INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Sistema de Recomendação em Real-Time para Reserva de transfers

Pedro André Freitas Camacho

Mestrado em Engenharia Informática

Orientadora:
Doutora Ana Maria de Almeida, Professora Associada,
Iscte - Instituto Universitário de Lisboa

Coorientador:
Doutor Nuno António, Professor Auxiliar Convidado,
NOVA IMS - Universidade Nova de Lisboa

Novembro, 2020

TECNOLOGIAS
E ARQUITETURA

Sistema de Recomendação em Real-Time para Reserva de transfers

Pedro André Freitas Camacho

Mestrado em Engenharia Informática

Orientadora:
Doutora Ana Maria de Almeida, Professora Associada,
Iscte - Instituto Universitário de Lisboa

Coorientador:
Doutor Nuno António, Professor Auxiliar Convidado,
NOVA IMS - Universidade Nova de Lisboa

Novembro, 2020

"Work hard in silence, let your success be your noise."
- Frank Ocean

Agradecimentos

Os méritos que esta dissertação possa ter, devem-se também aos contributos das pessoas que durante a sua elaboração, me proporcionaram testemunhos de vários géneros. Agradeço assim a todos os que me apoiaram neste projeto e tornaram possível a conclusão desta fase.

Gostaria de agradecer e destacar o papel fundamental neste trajeto pela minha orientadora pertencente ao Departamento de Ciências e Tecnologias da Informação (DCTI) da ISTA – Escola de Tecnologias e Arquitetura, Prof. Dra. Ana Maria Carvalho de Almeida, pelo profissionalismo, disponibilidade e apoio em todos os momentos no desenvolvimento desta dissertação. Fez-me usufruir das suas capacidades académicas ao longo deste trabalho em conjunto: pragmatismo, saber profundo e sentido pedagógico.

Agradeço também ao meu orientador, Prof. Dr. Nuno Miguel da Conceição António pelo apoio no desenvolvimento desta dissertação, pela sua generosidade, sensibilidade e profissionalismo, que foram fulcrais para atingir os nossos objetivos. Desta forma, agradeço a ambos os meus orientadores pela constante capacidade construtiva e objetiva que me ajudou sempre a melhorar o nosso trabalho, resultando nesta dissertação e num artigo aceite para a conferência internacional em turismo, tecnologia e sistemas, "ICOTTS'20 - International Conference on Tourism, Technology Systems"(Anexo A).

O meu agradecimento vai, também, para todos os docentes do ISCTE - Instituto Universitário de Lisboa, pela transmissão de conhecimentos e orientação no percurso académico e que foram fundamentais para a conclusão desta dissertação.

Aproximando-se este final de ciclo, gostaria de agradecer também à minha família, em especial à minha noiva, à minha madrinha, aos meus pais e aos meus sogros, por sempre apoiarem este meu objetivo de vida e todas as decisões inerentes ao mesmo.

Por fim quero agradecer a todos os meus colegas e amigos com os quais partilhei bons e maus momentos durante este ciclo do meu percurso académico, ajudando-me nas mais diferentes situações.

Resumo

O continuado crescimento do número de turistas dos últimos anos é proporcional à progressiva utilização de serviços de transfers, sendo também, a oferta deste tipo de serviços, cada vez mais uma tendência. Os clientes de hoje são mais exigentes e procuram uma experiência online mais simplificada e personalizada, que pode ser obtida através de técnicas de antecipação do comportamento do cliente. Na sociedade contemporânea, a procura por mecanismos que possam recomendar ou auxiliar na escolha de produtos ou serviços é cada vez mais uma tendência, fomentando os conceitos de *cross-selling* e *upselling* nas empresas. A aquisição de serviços privados de transfer através de reservas nos websites, geram uma grande quantidade de dados que podem ser utilizados para segmentar clientes e construir sistemas de recomendação que sugerem outros produtos ou serviços ao cliente. No decorrer desta dissertação, apresentamos e desenvolvemos um modelo de classificação híbrido tendo por base uma empresa de transfers, sediada no Algarve, que pretende aumentar as vendas dos seus serviços paralelos (experiências/*tours*). De forma a identificar-se o comportamento e padrões nos clientes da empresa, é efetuada uma análise exploratória, assim como, aplicadas técnicas de segmentação de clientes. O sistema de recomendação proposto, funciona com um modelo de classificação em que, identifica, numa primeira fase, possíveis compradores de experiências e, posteriormente, numa segunda fase, sugere qual das experiências disponíveis será mais adequada a cada cliente. Apenas uma baixa percentagem de clientes que compra serviços de transfers, também compra experiências e pretende-se aumentar esta percentagem.

Palavras-chave: Transfers, Experiências, *Clickstream*, Recomendação, Classificação, Segmentação de Clientes.

Abstract

The continued growth in the number of tourists in recent years is proportional to the incresead use of transfer services. The offer of this type of service is becoming a trend. Today's customers are more demanding and require a more streamlined and personalized online experience, which can be achieved through techniques to anticipate customer behaviour. In contemporary society, the search for mechanisms that can recommend or assist in choosing products or services is increasingly a trend, fostering the concepts of cross-selling and upselling in companies. The acquisition of private transfer services through reservations on the websites generate a large amount of data that can be used to segment customers and build a recommendation system that suggest other products or services to the customer. In the course of this dissertation, we present and develop a hybrid classification model based on a transfer company based in the Algarve, which intends to increase sales of its parallel services (experiences/ tours). An exploratory analysis was carried out to identify the company's customers' behaviour and patterns and apply customer segmentation techniques. The proposed recommendation system works with a classification model in which it determines, in the first stage, potential buyers of experiences. Later, in a second phase, it suggests which of the available experiences will be best suited to each client. Only a low percentage of customers who buy transfer services also buy experiences and are intended to increase this percentage.

Keywords: Transfers, *Tours*, Clickstream, Recommendation, Classification, Customer Segmentation.

Conteúdo

Lista de Figuras

CAPíTULO 1

Introdução

Neste capítulo introduzimos o âmbito desta dissertação, apresentando o enquadramento, motivação, objetivos que propomos atingir e por último, a estrutura do documento.

1.1. Enquadramento

A inteligência artificial (IA) e mais concretamente, as técnicas de aprendizagem automática (AA), têm evoluído ao longo dos anos, ao mesmo tempo que a aplicação destas técnicas nos modelos de negócio presencia-se cada vez mais em vários setores, como é o caso do turismo [**1**]. A grande quantidade e diversidade de dados gerados no turismo, proporcionam novas perspetivas e possibilidades para melhorar a experiência de jornada do cliente [**2**]. A aprendizagem automática oferece a possibilidade de obter padrões e previsões a partir de grandes e diversificadas quantidades de dados, para que sejam respondidas questões intrínsecas às empresas, melhorando o seu alinhamento com a estratégia de marketing.

O investimento das empresas em formas de marketing digital e disponibilização dos seus serviços online, a fim de melhorar as vendas, proporcionam ao cliente uma experiência online melhorada e personalizada. Estes serviços disponibilizados online, fornecem informações enriquecedoras e valiosas sobre a jornada e interação dos clientes, que podem ser utilizadas para descobrir padrões e antecipar as suas necessidades. O Google Analytics, por exemplo, é uma plataforma da Google que acompanha a interação dos clientes com os websites das empresas, recolhendo e agregando informações sobre a jornada dos clientes que os visitam. A utilização de técnicas para segmentação de clientes sob estes dados recolhidos, permitem melhorar a oferta de produtos e serviços, com base nas características e preferências dos cliente.

Os sistemas de recomendação são uma das aplicações melhor sucedidas e difundidas através dos algoritmos de aprendizagem automática, recomendando produtos/serviços e personalizando a jornada de cada cliente. A sua utilização é transversal aos vários domínios de negócio online, sendo aplicado ao comércio eletrónico, publicidade online, aplicações em dispositivos móveis, redes sociais e outras grandes áreas que envolvem transações pessoais e comunicações [**3**]. O objetivo crítico da aprendizagem automática, neste contexto, é aprender uma função que consiga prever a oportunidade de sugerir um novo produto ou serviço baseado no perfil de utilizador.

Os serviços de transfer privados, são uma subárea do turismo que é responsável pelo transporte rodoviário de clientes entre locais, utilizando veículos privados da empresa [**4**]. Com o aumento destes serviços, as empresas tendem a encontrar formas de melhorar a experiência dos clientes, antecipando as suas necessidades e diversificando a sua panóplia

de serviços oferecidos. A venda de serviços paralelos aos serviços de transfers privados, como a venda de experiências *tours*, são um dos exemplos de serviços disponibilizados para cativar os clientes e de estratégia *cross-selling*.

A YellowFish é uma empresa de transfers privados, sediada no Algarve, responsável principalmente por serviços de transfer do aeroporto de Faro para todo o Algarve e vice-versa. A empresa iniciou sua atividade em Janeiro de 2010 tendo mais de 5000 clientes e várias dezenas de milhares de reservas por ano, desde a sua criação. Para além de operar no Algarve, a Yellowfish fornece serviços de transfers no Alentejo, Lisboa e sul de Espanha, utilizando a sua própria frota de automóveis e mini autocarros com capacidade para 4 a 8 passageiros. A empresa vende ainda experiências lúdicas, que podem ser desfrutadas no Algarve e no decorrer da estadia dos clientes.

1.2. Motivação

O turismo tem denotado um enorme crescimento ao longo das últimas décadas. Em termos mundiais, em 2018, verificou-se uma tendência crescente nas chegadas de turistas internacionais com 5,6%, sendo que a Europa continua a concentrar maior parte dos turistas internacionais, com 50,9% dos visitantes. Relativamente a Portugal, constatou-se um crescimento de 7,5% no número de chegadas de turistas, face a 2017. O Algarve mantém-se como principal destino turístico em Portugal, registando 30,2% das dormidas em 2018 [**5**].

A crescente utilização de serviços de transfers é proporcional ao aumento do turismo, sendo também, a oferta deste tipo de serviços, cada vez mais uma tendência [**6**], [**7**]. Dado o aumento exponencial das empresas de prestação de serviços de transfers privados, os clientes despendem mais tempo na escolha da empresa indicada e quais os serviços adequados escolher aquando da reserva do serviço. No entanto, existe alguma falta de complementaridade na investigação sobre a perceção do cliente de transfers privados.

Na sociedade contemporânea, a procura por mecanismos que possam recomendar ou auxiliar na escolha de um novo produto ou serviço é cada vez mais uma tendência. Mais ainda, a procura por preços mais acessíveis ou outros serviços em websites terceiros às empresas fornecedoras, designadas por afiliadas, é uma ameaça para a margem de lucro da empresa que aprovisiona os serviços. Existe a necessidade das empresas de transfers filtrarem, priorizarem e fornecerem informações relevantes de forma eficiente, com a finalidade de conseguir cativar o cliente em micro segundos, para que, não só a compra seja feita com a empresa diretamente, mas também seja possível sugerir adicionar novos serviços ou produtos de valor acrescentado. Os sistemas de recomendação conseguem fazer face a este problema através da análise de grandes quantidades de informação geradas, de forma a fornecer aos utilizadores conteúdos e serviços personalizados.

A venda de serviços ou produtos secundários, nem sempre têm o mesmo volume de vendas que os serviços/produtos principais têm numa empresa. Isto implica um desequilíbrio no número de exemplos entre as classes existentes no conjunto de dados. Nos

problemas de aprendizagem automática, são comuns os problemas com dados desequilibrados, por exemplo, na deteção de fraude em que a classes minoritária contem apenas 2-5% dos exemplos [**8**].

Apenas uma percentagem reduzida dos clientes de transfers (0,20%) adquire experiências. Ao longo deste documento, propomos e desenvolvemos um sistema de recomendação que identifica possíveis compradores de experiências, e sugeridas a estes uma experiência em específico. Sob forma de identificarmos padrões de comportamento nos clientes de transfers, são utilizadas técnicas de segmentação de clientes que alimentam o sistema de recomendação. O sistema, baseado num modelo de classificação híbrido, utiliza um conjunto integrado de algoritmos de aprendizagem automática com técnicas de *oversampling* e *undersampling* nos dados, que permitem identificar e antecipar padrões no comportamento do cliente, num conjunto de dados desequilibrado.

1.3. Objetivos

O objetivo desta investigação, passa por desenvolver um estudo empírico recaindo sobre o negócio de transfers privados. Como objetivos centrais desta investigação, pretende-se descobrir, interpretar e comunicar padrões significativos nos dados da YellowFish, que ajudem a caracterizar os padrões de comportamento dos clientes, e ainda desenvolver um sistema de recomendação, baseado num modelo de classificação híbrido, que consiga aumentar a venda de serviços paralelos. O sistema proposto, pretende sugerir aos clientes de serviços de transfer, experiências lúdicas, vendidas também pela empresa de transfers. São utilizados dados históricos das reservas e o modelo de classificação divide-se em duas fases, numa primeira fase as vendas de transfers são classificadas entre possível comprador de experiência ou não (classificação binária), e numa segunda fase, o modelo identifica qual das experiências será mais adequada aos registos que foram classificados como possíveis compras de experiências (classificação multi-classe).

Esta investigação pode ser dividida em dois sub problemas, a análise comportamental dos clientes de transfers e experiências através da análise exploratória e segmentação de clientes, e o desenvolvimento de um modelo de classificação que consiga identificar quando sugerir uma experiência específica a um cliente de transfer.

No final desta investigação pretendemos responder à seguinte questão:

- Dadas as características de cada reserva de transfer e as características de cada cliente, como desenvolver um sistema capaz de identificar possíveis clientes de serviços adicionais e como sugerir um destes serviços em especifico?

As questões de investigação que pretendem ser endereçadas no desenvolvimento desta investigação recaem sobre:

- Que fatores caracterizam o comportamento dos clientes na navegação e reserva de serviços no website da YellowFish.
- Como identificar o comportamento de clientes que realmente pretendem reservar serviços, dos restantes, que apenas visitam o site.

- Distinguir como é que os clientes de transfers se correlacionam de forma específica relevante para campanhas de marketing.
- Quais as experiências a oferecer online a cada cliente em complemento ao transfer de modo a fomentar uma estratégia de cross-selling?

1.4. Estrutura do Documento

O presente documento encontra-se organizado em cinco capítulos:

- No **Capítulo 2** é apresentada a revisão da literatura, que consiste na análise e revisão de estudos científicos elaborados na mesma área de investigação. Ciência dos dados, sistemas de recomendação e técnicas para equilíbrio dos dados são as principais áreas de análise.
- O **Capítulo 3** apresenta o caso de estudo e as características associadas à empresa sob a qual a investigação incidiu.
- No **Capítulo 4** é descrito todo o processo de limpeza, integração, análise descritiva, seleção e transformação dos dados.
- No **Capítulo 5** é desenvolvido o sistema de recomendação e detalhado o modelo de classificação híbrido. São ainda descritos os métodos, técnicas, bibliotecas e ferramentas que foram utilizados, assim como, resultados para cada abordagem.
- O **Capítulo 6** apresenta uma comparação empírica entre as várias abordagens, demonstrando que características optimizam o desempenho do modelo de classificação.
- Por último, o **Capítulo 7** apresenta as conclusões desta investigação empírica, assim como, indicações para investigação futura a fim de melhorar o desempenho do sistema de recomendação e modelo de classificação.

CAPíTULO 2

Revisão da Literatura

Neste capítulo apresenta-se a análise dos estudos científicos, teóricos e empíricos, que têm sido desenvolvidos nas áreas de estudos inerentes a esta investigação. Desta forma, são apresentados e descritos os técnicas, algoritmos e modelos que têm sido utilizados para extrair conhecimento sobre os clientes e desenvolver sistemas de recomendação. A revisão literária é dividida em três tópicos principais: ciência dos dados, sistemas de recomendação, e técnicas para equilíbrio dos dados.

2.1. Ciência dos Dados e *Analytics* no Turismo

A ciência dos dados, sendo a área cientifica que agrega e analisa grandes quantidades de dados, tem o objetivo, dado um problema, de extrair, descobrir, interpretar e comunicar padrões com significado e prever eventos que possam surgir destas análises. No contexto de negócios, a ciência dos dados ajuda a perceber as oportunidades de negócio ocultas nos dados e informações que são todos os dias geradas, sob forma de se obter o máximo valor dos dados e extrair conhecimento sobre os clientes, utilizando dados explícitos e implícitos. A capacidade de recolher, armazenar e analisar enormes quantidades de dados, implica uma nova era de personalização relativamente à experiência e jornada dos clientes [**9**], [**10**].

Os padrões encontrados através da análise dos dados, permitem predições para o futuro baseadas nas tendências existentes. A indústria do turismo tem evoluído de forma a prever o comportamento do cliente e encontrar padrões que permitem antecipar expetativas e ações. A análise de sentimento e perfis dos clientes, permitem a análise de milhares de *reviews*, fotos e comentários, sob forma de ser tomadas decisões rápidas e estratégicas, baseadas no que os clientes pensam sobre as empresas turísticas [**11**]. O investimento em marketing online objetiva maximizar a venda de produtos ou serviços sem obrigar a expor-se em demasia. A maximização da conversão de visitantes em clientes, com o menor investimento é o principal alvo do marketing. Os algoritmos preditivos estimam a taxa de conversão, otimizando o marketing online e entendendo as necessidades dos clientes com base nos vários fatores que parecem afetar as suas preferências [**12**].

Os sistemas de recomendação têm sido largamente difundidos, valendo-se da grande quantidade de informação disponível, implícita e explicita, e que ajudam os utilizadores a tomar as melhores decisões e a melhorar as estratégias de negócio das empresas. A utilização de sistemas de recomendação no turismo é cada vez mais uma tendência, em que os clientes são abordados para comprar produtos ou serviços (hotéis, voos, transfers, *tours*, entre outros), com base no seu perfil ou informação contextual. A recomendação

de hotéis com base no perfil e *reviews* de outros utilizadores é uma das aplicações mais disseminadas [**13**], [**14**], [**15**]. Não obstante, os sistemas de recomendação são transversais às várias áreas de negócio e continuam em constante crescimento [**16**].

Especificamente em relação ao transporte de turistas, a investigação tem sido alargada à analise estatística do comportamento dos turistas, relativamente aos padrões de deslocações e métodos de transporte [**17**], bem como a influência dos transportes na satisfação para prever a intenção de revisitar um destino [**18**].

2.1.1. Experiência e Jornada do Cliente

A jornada do cliente caracteriza-se pelo conjunto de eventos nos quais o cliente interage durante a sua navegação nos websites das organizações [**19**]. Um dos objetivos da jornada dos clientes é mapear diagramas que ilustram os principais passos que os clientes tomam ao se conectarem com uma determinada empresa, seja uma experiência online, compras, serviços ou outra combinação [**20**]. Através da concepção e análise da jornada do cliente, pretende-se maximizar o valor do cliente e da organização, sob forma de personalizar ao máximo a experiência do cliente com a organização [**21**]. A análise da jornada do cliente, fornece às organizações conhecimento sobre a experiência dos atuais e potenciais clientes, percecionando os pontos críticos e enfatizando como a atual experiência se difere comparada com as suas expetativas. O conhecimento ganho através desta análise pode ser utilizado para desenhar e implementar soluções ligadas à experiência dos clientes, que vá de encontro às suas necessidades e aos objetivos da organização, de forma a ser atingida vantagem competitiva [**22**]. O registo do comportamento e jornada dos clientes na navegação de um website, designa-se por *clickstream*.

Várias ferramentas estão disponíveis para recolha, agregação e análise dos dados gerados online, através da interação com os serviços disponibilizados pelas empresas online, como o Google Analytics, que se destaca pela sua popularidade e versatilidade. A grande quantidade de informações recolhidas pelo Google Analytics representa uma fonte valiosa de informações implícitas, tais como as páginas com mais cliques, em que tipo de informações os clientes estão mais interessados, quais os caminhos críticos que os utilizadores preferem, entre outras. Estas informações podem ser submetidas a análise temporal, a fim de gerar previsões e conhecimentos úteis sobre os clientes que as visitam [**23**].

A personalização de experiências é uma área de investimento crescente que requer atenção das organizações, dado que os clientes esperaram presenciar mais este tipo de interação. Numa recente análise efetuada a clientes, comprova-se que estes têm altas expetativas para encontrar experiências personalizadas nos websites que visitam, assim como expressam o seu desapontamento quando existe a falta de personalização nas suas experiências de compras online [**24**]. Verifica-se, assim, a relevância e importância da personalização de experiências a disponibilizar ao cliente aquando da sua navegação ou compras nos websites, impulsionando as compras, lucros, fidelidade do cliente e melhorando, no geral, a satisfação do cliente.

Antes de se conseguir recomendar ou sugerir algum tipo de personalização ao serviço ou produto, é necessário identificar e classificar a intenção de compra por parte de um determinado cliente. A clara identificação da intenção de compra pelos clientes é importante, na medida em que os sistemas apenas devem sugerir/recomendar algum produto/serviço, quando for o momento ideal e propício. A negligência na recomendação precoce de informação desnecessária ao utilizador pode implicar a perda do mesmo. Os algoritmos de aprendizagem automática conseguem ajudar na procura pela melhor precisão na classificação de intenção de compra dos utilizadores [**25**].

Através da interação com websites é gerada uma enorme quantidade de informação dinâmica, quer seja ela implícita ou explícita. Informação implícita é extraída através do comportamento do utilizador sem haver uma direta intencionalidade de a fornecer, como é o caso dos dados sobre o caminho de navegação no website, os vídeos que clicou, o número de músicas ouvidas, entre outros. Na informação explícita, existe intencionalidade por parte do utilizador em fornecer a informação ao sistema, como é o caso, por exemplo, dos *ratings*, em que o utilizador classifica através de texto ou escala um determinado produto ou serviço. Toda esta informação pode ser utilizada para caracterizar o perfil do utilizador, assim como comportamentos e semelhanças entre utilizadores, quer seja no âmbito do feedback implícito ou explícito, de forma a serem construídas experiências personalizadas (Figura 2.1 [**26**]) [**27**],[**28**].

Feedback Explícito. O feedback explícito caracteriza-se pelo fornecimento, por parte dos utilizadores, de um produto ou serviço, uma classificação ou comentário, que acarreta grande valor, visto que caracteriza o pensamento do utilizador sobre o produto/serviço em questão e que o sistema toma em consideração para efetuar futuras recomendações. O feedback explícito acarreta uma preocupação acrescida para o cliente, a qual, muita das vezes não é considerada pelo mesmo, não respondendo a qualquer tipo de feedback. Para além da possibilidade de os clientes não fornecerem qualquer tipo de feedback ao sistema sobre a sua experiência, existe a possibilidade de o feedback fornecido ser de baixa qualidade ou não ter qualquer tipo de veracidade, por exemplo, no caso de o cliente, por alguma razão desconhecida, fornecer feedback aleatório ou totalmente contrário àquilo que pensa. Isto implica que o treino e teste dos algoritmos resultaram em fracos resultados, no caso de serem fornecidos poucos feedbacks ou de má qualidade.

Feedback Implícito. A recolha de feedback explícito, torna-se, por vezes, uma tarefa difícil dado que os utilizadores têm que expressar intencionalmente os seus interesses, classificações ou observações sobre um determinado produto ou serviço. A dificuldade passa pela complexidade que estes feedbacks exigem aos utilizadores, em termos de perceção da informação e gasto de tempo. Na tarefa de personalização de experiências, surge a necessidade de trabalhar com outro tipo de dados que caracterizem as preferências dos utilizadores, de forma a serem feitas recomendações personalizadas com sucesso. Caracteriza-se por feedback implícito, o comportamento de navegação dos utilizadores num determinado sistema de informação. No que diz respeito à utilização de feedback

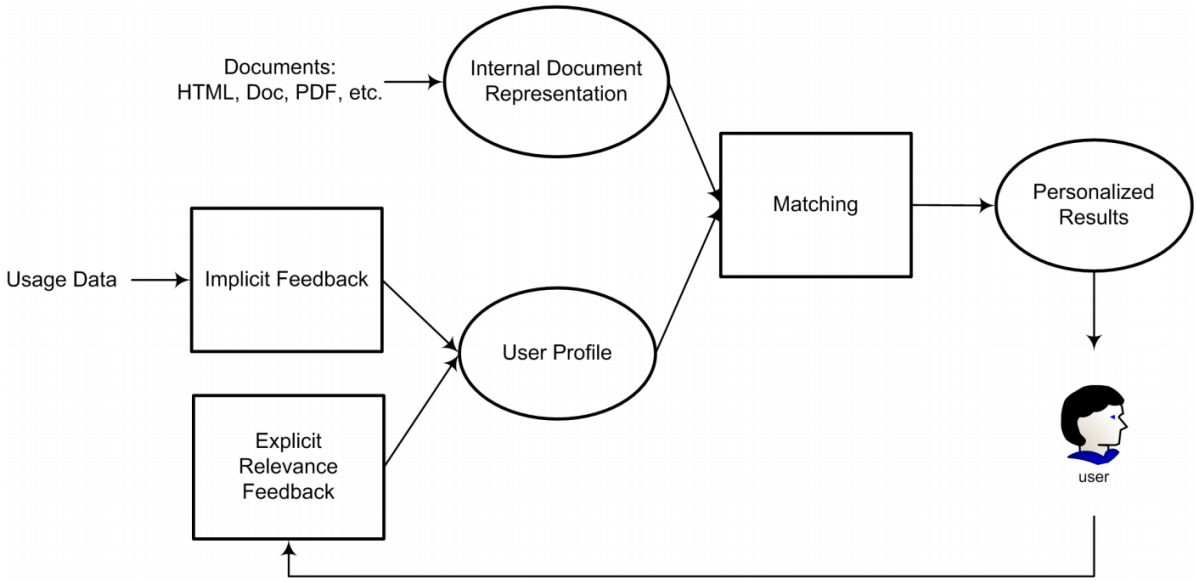


FIGURA 2.1. Feedback implícito e explicito utilizado para aprender e manter atualizado o perfil dos utilizadores utilizado durante a personalização.

implícito na criação de experiências personalizadas, estes dados têm sido utilizados, com sucesso, em sistemas de recomendação baseados em filtragem colaborativa [**29**].

A figura 2.2 ilustra um exemplo de um fluxo de comportamento que faz parte do Google Analytics. Apesar da plataforma permitir a agregação dos fluxos principais que os utilizadores percorrem, é necessário obter, a um nível mais granular, o comportamento dos utilizadores, de forma a ser possível a alimentação dos modelos com estes dados. É importante perceber se os clientes que navegam no website têm intenções de concluir uma compra, a fim de lhes serem sugeridas algumas recomendações a cerca dos produtos, serviços interessantes para o mesmo ou, até mesmo, pacotes de serviços.

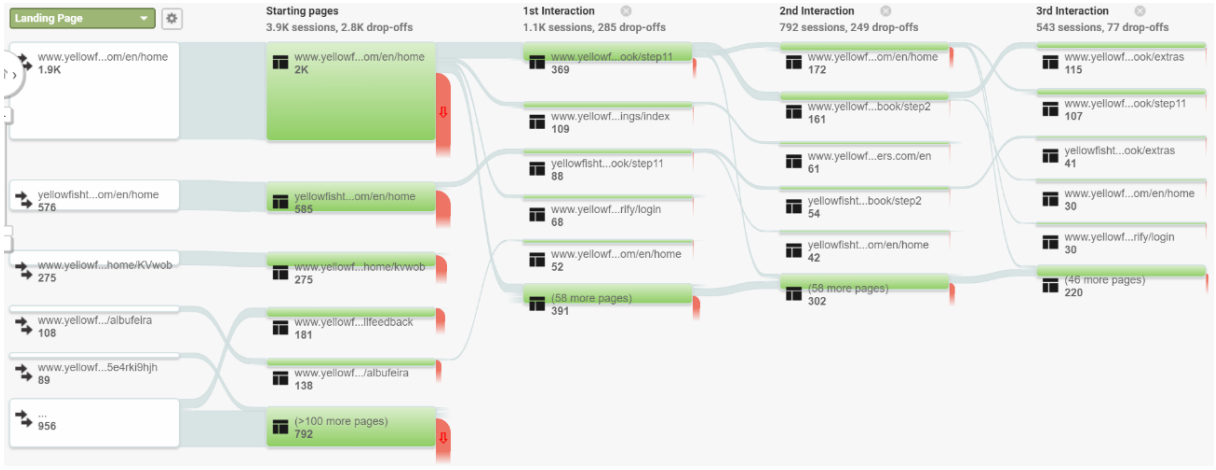


FIGURA 2.2. Caminho de navegação dos clientes no website da YellowFish recolhida através da plataforma Google Analytics.

Sheil *etal.* descreve uma rede neuronal que prevê a intenção de compra num ambiente de comércio electrónico, abordando a significância sobre o investimento em *feature*

engineering para melhorar os resultados dos modelos [**25**]. Os resultados demonstram que o estudo alcançou 98,4% de desempenho, em comparação com o estado da arte sobre predição de compra em comercio eletrónico, sem utilizar *features* explícitas.

Os ratings implícitos em relação ao explícitos, diferem e ganham vantagem na medida em que os utilizadores não têm de responder a perguntas ou dar feedback intencional, para que o treino dos algoritmos funcione. Contudo, ambas as abordagens podem ser combinadas para melhorar o conhecimento sobre as preferências e pensamentos dos utilizadores. Nesta investigação será considerada a combinação de ambos os feedbacks.

2.2. Sistemas de Recomendação

Os sistemas de recomendação são uma forma de ajudar os utilizadores a tomar decisões em ambientes de informação complexa [**30**]. De uma perspetiva de comércio electrónico, os sistemas de recomendação são definidos como uma ferramenta que ajuda os utilizadores a encontrar itens semelhantes, tendo em conta o interesse e as preferências de outros utilizadores [**31**]. A evolução dos sistemas de recomendação tem registado várias *nuances*, desde a utilização e criação de várias metodologias até à utilização de diferentes tipos de algoritmos de aprendizagem automática, de forma a melhorar personalização de ofertas a clientes [**32**].

A figura 2.3 ilustra as várias técnicas utilizadas para a construção de sistemas de recomendação. As aplicações de sistemas de recomendação aplicam-se a domínios, tais como, bibliotecas online, medicina, comercio electrónico, música e outros segmentos de mercado. A Amazon, por exemplo, uma das maiores empresas de comércio electrónico, utiliza algoritmos de diversificação de tópicos para produzir as suas recomendações aos utilizadores. A técnica de filtragem colaborativa é utilizada para gerar uma matriz de itens semelhantes com base na interligação entre itens na matriz. A técnica também recomenda outros itens aos utilizadores, baseado no seu comportamento e perfil de compras.

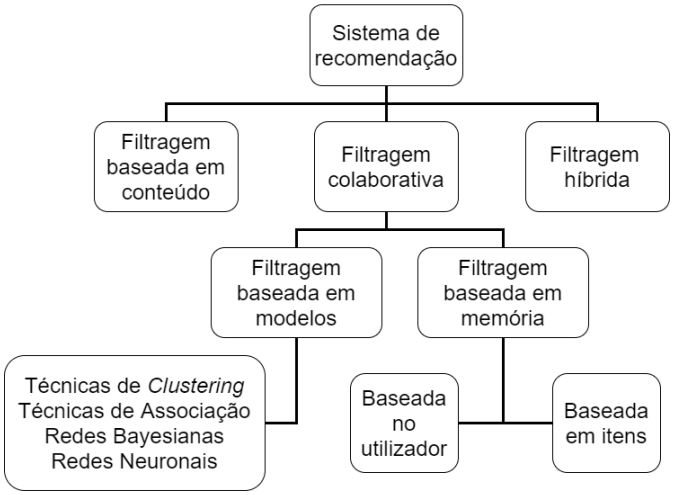


FIGURA 2.3. Técnicas utilizadas na construção de sistemas de recomendação.

Filtragem colaborativa é uma das abordagens mais utilizadas para desenvolver sistemas de recomendação. Os métodos de filtragem colaborativa são estabelecidos na recolha

e análise de grandes quantidades de informação baseadas em comportamentos, atividades ou preferências e antecipando o gosto do utilizador em causa utilizando para isso a sua semelhança com outros utilizadores [**33**]. Os objetos recomendados por filtragem colaborativa são selecionados baseando-se em avaliações passadas de grandes grupos de utilizadores.

Estes sistemas podem ser vistos como um método para apoiar e melhorar as decisões dos utilizadores, baseados em conhecimento de decisões de outros utilizadores com características pessoais similares, quando não existe informação suficiente acerca de determinados utilizadores [**34**]. Os sistemas de recomendação oferecem também a possibilidade de apresentar recomendações personalizadas, de serviços e conteúdos exclusivos, quando os utilizadores se deparam com um excesso de informação ao navegar em websites. Para desenvolver estas soluções com sucesso, os sistemas de recomendação baseiam-se em três técnicas principais: filtragem colaborativa [**35**], filtragem baseada em conteúdo [**36**], e filtragem híbrida [**37**].

Filtragem baseada em conteúdo tenta recomendar itens aos utilizadores com base na "contagem" de similaridade que é avaliada por esse utilizador positivamente no passado [**33**]. Nos métodos de filtragem baseada em conteúdo, o problema de recomendação é fundido num problema de classificação que prevê se os utilizadores gostaram ou não de um determinado item, ou num problema de regressão (prevê a classificação a ser dada por um utilizador a um item).

Técnicas de filtragem baseadas em conteúdo e colaboração têm sido amplamente utilizadas no desenvolvimento de sistemas de recomendação. Contudo, existem alguns problemas e limitações, tais como, conteúdo limitado para análise, dispersão de dados, sobre-especialização, arranque a frio e escalabilidade [**38**]. As técnicas de filtragem híbridas resolvem algumas das limitações que as técnicas de filtragem baseados em conteúdo e colaboração possuem, através da combinação de dois ou mais métodos de filtragem, de forma a melhorar o desempenho e precisão dos sistemas de recomendação [**39**].

A utilização de dados de *clickstream* dos utilizadores, para a criação de sistemas de recomendação tem sido frequentemente utilizados, de forma a serem identificados padrões de comportamentos na navegação destes utilizadores. Contudo, a combinação de dados históricos de compras e *clickstream* fornece previsões e padrões sobre os utilizadores mais precisas, melhorando significativamente a finalidade do sistema de recomendação [**40**].

No domínio das notícias online, a personalização e previsão para os utilizadores com base em dados de *clickstream*, melhora a qualidade da notícia recomendada e atrai visitantes mais frequente ao site da Google News, através de uma abordagem híbrida, combinando uma framework Bayesiana e um modelo existente de filtragem colaborativa [**41**]. O desenvolvimento de sistemas de recomendação utilizando dados de *clickstream* tem vindo a aumentar. No artigo [**42**], os autores propõem um sistema que analisa os dados de *clickstream* obtidos de um website de comércio eletrónico e prevêem os valores de preferência dos clientes para os produtos vistos mas não comprados, utilizando classificadores

mais eficientes como Random Forests e gradiente boosting, e filtragem colaborativa para recomendar produtos. Constatou-se, também, que a utilização de melhores medidas de similaridade, i.é., com um eficiente algoritmo de *clustering* em filtragem colaborativa ajuda a produzir melhores recomendações. A Netflix é também um exemplo de utilização de um sistema de recomendação híbrido [**43**]. O website faz recomendações comparando os hábitos de visualizações e pesquisa dos utilizadores semelhantes (filtragem colaborativa), assim como, oferecendo filmes que partilham características com filmes que o utilizador tem avaliado com melhores pontuações (filtragem baseada em conteúdo).

Em [**44**], foi desenvolvido um novo sistema de recomendação com base na análise estendida do comportamento dos utilizadores. O sistema proposto extrai informações de *recall*, *precision*, número de itens clicados, sequência de itens clicados, duração do acompanhamento, número de acompanhamentos para o mesmo item, gostou/não gostou, regras de associação dos itens clicados e considerações para os itens, resultando numa solução melhor sucedida em relação ao estado do arte para os resultados dos métodos de filtragem colaborativos. Os sistemas de recomendação baseados em filtragem de conteúdo utilizam um paradigma diferente, onde o feedback não é relevante, mas sim a similaridade no conteúdo dos itens. No caso, se um utilizador demonstrou interesse, por exemplo, em uma música, o sistema de recomendação analisa o conteúdo dessa música e tenta encontrar e recomendar outras músicas semelhantes. Os algoritmos que endereçam estas capacidades de encontrar conteúdo similar, geralmente tiram partido das *features* de contagem e estabelecem frequências, sob forma de se obter grandes palavras-chave discriminantes sobre o conteúdo em análise.

2.3. Técnicas para Equilibrar Conjuntos de Dados

A venda de experiências pela YellowFish ainda representa uma pequena parte do negócio da empresa, sendo registadas poucas vendas de serviços de transfer com um serviço de experiência associado, resultando em apenas cerca de 0,20% dos clientes de transfer que adquire experiências. Com isto, são registados no conjunto de dados muitas mais observações de serviços de transfer sem experiência do que com experiência associada, gerando-se assim, dados desequilibrados do ponto de vista do número total de exemplos para cada tipo de serviço vendido.

A utilização de técnicas de equilíbrio dos dados é cada vez mais uma tendência, tanto em aspetos teóricos como práticos, dado o aumento de problemas com esta particularidade. Dados com distribuição de classes extremamente irregulares causam problemas em 3 dimensões:

- **Problema com a máquina** - Os algoritmos de aprendizagem automática são criados para minimizar os erros. Dado que a classe maioritária têm um maior número de exemplos disponíveis, a probabilidade de um determinado exemplo pertencer a essa classe é elevada. Tendo isto em consideração, os algoritmos de aprendizagem automática tendem a classificar estes exemplos como da classe maioritária, incorrendo em falsos positivos (FP) ou falsos negativos (FN) . Por

exemplo, no caso de deteção de fraude em operações bancárias, grande parte das operações são designadas como normais em cerca de 99% dos casos e os algoritmos tendem a classificar todas as operações como normais. Por outro lado, os 1% que correspondem a fraudes, podem implicar graves riscos se não forem corretamente detetados [**45**].

- **Problema intrínseco** - Dependo do problema em questão e daquilo que se pretende minimizar, a classificação de falsos positivos ou falsos negativos podem acarretar graves consequências. Contudo, na vida real, a deteção de falsos negativos implica um custo maior que o erro na deteção de casos positivos. Para a maior parte dos algoritmos de aprendizagem automática, a deteção de falsos positivos e falsos negativos é penalizado com a mesma importância. Tendo em conta um problema de deteção de cancro, se os modelos preverem que um determinado paciente não tem este problema, mas depois verificar-se que o tem, estamos perante um falso negativo em que o custo de tal classificação, no limite, pode custar a vida do paciente. Por outro lado, um falso positivo neste caso, originaria consequências menores, como a preocupação e medidas preventivas. Nos problemas mais relativos, como é o caso de recomendação de produtos ou serviços aos utilizadores, o impacto do erro em classificações de FP e FN é menor, dado que a pior das hipóteses com um FP é a perda de um possível cliente e de FN com a perda de vendas para a empresa [**46**].
- **Problema humano** - A utilização de teorias transversais a vários problemas pode não ser adequada, dado que os padrões comportamentais das amostras podem ser diferentes e não se obter resultados favoráveis. O que pode ser ótimo para um certo tipo de recomendação, pode não ser ótimo para outro caso. No caso de crédito de risco, práticas comuns são estabelecidas por peritos, em vez de estudos empíricos [**47**].

Algumas soluções têm sido desenvolvidas para endereçar estes problemas, tanto com abordagens ao nível dos algoritmos como a nível dos dados.

Solução Algorítmica: Dado que os algoritmos de aprendizagem automática penalizam os FP e FN de forma igualitária, uma forma encontrada para mitigar este problema é através de algoritmos que aumentem o desempenho preditivo nas classes minoritárias, através de aprendizagem baseada em reconhecimento ou aprendizagem sensibilidade-custo [**48**], [**49**], [**50**].

Solução de Reamostragem: Uma das abordagens mais comuns neste tipo de problema, é o equilíbrio dos dados, através do aumento do número de exemplos das classes minoritárias, ou decréscimo das classes maioritárias. Uma das vantagens da utilização de técnicas de equilíbrio de dados, é a flexibilidade e utilização de algoritmos de aprendizagem automática gerais e atualizados [**8**]. As duas técnicas mais comuns são *oversampling* e *undersampling*. *Oversampling* aumenta o número de registos das classes minoritárias no conjunto de treino, tendo como vantagem a integridade da informação, com todas as

observações das classes maioritárias e minoritárias a serem mantidas, contudo, é propicio a *overfiting*. Pelo lado contrário, a técnicas de *undersampling*, reduzem o número de observações das classes maioritárias e pode ser perdida alguma informação crítica. Alguns estudos demonstram que a combinação das duas abordagens gera bons resultados no desempenho dos modelos.

2.3.1. Técnicas de *oversampling*

O problema com dados desequilibrados é um desafio inerente aos problemas de aprendizagem automática, com cada vez mais investigações cientificas transversais a vários domínios. As técnicas de oversampling aplicadas em dados desequilibrados, têm como objetivo aumentar o número de observações presentes nas classes minoritárias. Synthetic Minority Oversampling (SMOTE) [**51**] e ADAptive SYNthetic (ADASYN) [**52**] são os dois algoritmos principais associados a estas técnicas.

Uma das técnicas mais comuns é a geração de observações aleatórias a partir dos dados reais, contudo, uma grande desvantagem é a geração não descriminada que pode gerar *overfitting*. Em vez de simplesmente aumentar o número de observações através da replicação de observações das classes minoritárias, o que originaria *overfitting*, o algoritmo **SMOTE**, cria observações sintéticas baseadas nas observações das classes minoritárias existentes. Estas observações são criadas com base nas k amostras mais próximas à classe que se quer aumentar. Dependendo do número de amostras que se pretende aumentar, um ou mais vizinhos são selecionados para criar as observações sintéticas. A utilização da técnica SMOTE em áreas como a previsão de rotação dos clientes em comércio eletrónico, indica que o equilíbrio dos dados é uma boa abordagem para melhorar os desempenhos dos modelos preditivos [**53**].

A abordagem **ADASYN** baseia-se na utilização de distribuição de pesos para as diferentes amostras das classes minoritárias, de acordo com o seu nível de dificuldade em aprender, onde mais dados sintéticos são gerados para os exemplos das classes minoritárias que são mais difíceis de aprender, comparadas com os exemplos minoritários que são mais fáceis dos modelos aprenderem. Esta baseia-se no conceito de geração de observações das classes minoritárias, através da adaptação às suas distribuições utilizando os k vizinhos mais próximos.

As diferenças entre o algoritmo SMOTE e ADASYN prendem-se no facto de o algoritmo SMOTE gerar o mesmo número de amostras sintéticas para cada amostra minoritária original, enquanto que o ADASYN utiliza uma distribuição de densidade como critério para decidir automaticamente o número de amostras sintéticas que devem ser geradas para cada amostra minoritária, alterando adaptativamente os pesos das diferentes amostras minoritárias para compensar a distribuições distorcidas.

2.3.2. Técnicas de *undersampling*

As técnicas de *undersampling* objetivam reduzir o número de amostras das classes maioritárias, com o mínimo de perda de informação crítica. A aplicação de técnicas de *undersampling* é transversal a problemas de aprendizagem automática [**54**].

As técnicas de *undersampling* são baseadas em técnicas como *RandomUnderSampling*, *NearMiss*, *CondensedNearestNeighbour*, *TomekLinks*, *EditedNearestNeighbours*, *Neigh- bourhoodCleaningRule*, *ClusterCentroids*.

2.3.3. Combinação de abordagens de reamostragem

A utilização de combinação de abordagens *oversampling* e *undersampling* num mesmo problema tem sido uma vertente investigada com vista a melhorar o desempenho dos modelos preditivos [**55**].

A técnica SMOTE em combinação com a técnica ENN faz a fusão destes dois conceitos. Através do algoritmo SMOTE é feito *oversampling*, complementando com a técnica de limpeza e *undersampling* ENN. A vantagem deste algoritmo combinado passa pela limpeza de todas as classes, removendo exemplos também da classe maioritária, em vez de apenas da classe minoritária. A técnica ENN tende a remover mais observações, fazendo uma limpeza mais profunda, comparativamente à técnica *TomekLinks*.

Outra técnica utilizada é a fusão do algoritmo SMOTE com a técnica de limpeza e *undersampling* *TomekLink*, que tem um comportamento semelhante à SMOTE + ENN, mas, como já referido, com a particularidade de remover menos exemplos.

CAPíTULO 3

Caso de Estudo - YellowFish

Neste capítulo é descrito o caso de estudo inerente a esta investigação, apresentando a empresa e o seu negócio, assim como, o problema inerente que se pretende resolver com este trabalho e o conjunto de dados associado, composta por uma grande quantidade de dados sobre as operações e clientes da empresa.

3.1. Descrição da Empresa

A Yellowfish iniciou a sua atividade em 2010 como agência de viagens e tem demonstrado um crescimento exponencial nos últimos 9 anos. A empresa é composta, principalmente, pela venda de serviços de transfers privados entre localizações (Yellowfish Transfers). Contudo, o grupo expandiu-se, sendo agora composto por agências especializadas que cobrem toda a experiência turística no Algarve, desde reserva de hotéis e lugares para ficar (Yellowfish Beds) à venda de experiências lúdicas que podem ser desfrutadas na zona do Algarve com passeios de quadriciclo e *buggy* (Yellowfish Adventures). Para além destes projetos da empresa, brevemente serão lançados mais dois projetos com vista a melhorar a orientação turística dos seus clientes (YellowFish Guide) e ainda aluguer de viaturas elétricas (Kwako) no Algarve [**56**].

A Yellowfish Transfers é a principal filial da empresa Yellowfish, sendo a mais antiga e apresentando o volume de negócios mais elevado. Para além da Yellowfish Transfers, a Yellowfish Adventures tem denotado um crescimento desde a sua criação, contudo a um muito menor ritmo que a Yellowfish Transfers.

3.1.1. Yellowfish Transfers

A Yellowfish Transfers tem como objetivo de negócio a venda de transfers privados entre diferentes localizações, transportando individualmente o(s) cliente(s) de cada reserva feita. Executa maioritariamente serviços de transfers na zona Algarvia, desde o aeroporto para hotéis e vice-versa. Contudo, possibilita ainda a concretização de serviços em Lisboa, Alentejo e sul de Espanha. Para além de serviços de transfers privados, a Yellowfish Transfers, especializa-se em transfers golfe, que se caracterizam pelo transporte de clientes para clubes de golfe, com possibilidade de só ida ou ida e volta.

As reservas de transfers são efetuadas maioritariamente através do website da Yellowfish Transfers ou por parceiros afiliados. Relativamente ao canal online da venda de serviços de transfer no website da empresa, o processo começa pela escolha dos pontos de partida e chegada, assim como as datas do serviço, sentido (ida ou ida e volta), horários e número/tipo de passageiros. Seguidamente, são identificados o número e tipo de bagagens, finalizando com o respetivo pagamento. Após o pagamento com sucesso é inserido

na base de dados da empresa o serviço comprado. Na Fig. 3.1 [**57**] encontra-se ilustrado todo este processo de compra de um serviço de transfer.

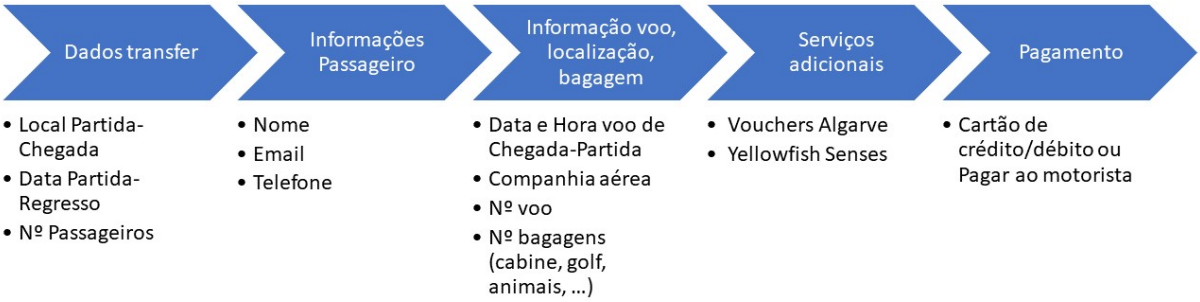


FIGURA 3.1. Processo de compra de serviços de transfer online no website da Yellowfish Transfers.

Os serviços de transfers são efetuados por motoristas contratados pela empresa, que utilizam a frota disponível (carros, mini-autocarros e mini-vans), com uma capacidade máxima de até 8 pessoas, para concretizar os serviços nos dias e horários marcados. Em casos esporádicos, a empresa contrata empresas terceiras para realizar um serviço associado a uma reserva efetuada, o que acontece apenas em situações em que existe um maior fluxo de reservas e a frota não é capaz de dar resposta a todas.

Apesar do objetivo principal de negócio da Yellowfish ser a venda de transfers, a empresa também vende experiências lúdicas na região do Algarve, que se distinguem por passeios de *buggy* ou moto quatro e possibilitam o conhecimento dos diversos pontos turísticos da região Algarvia.

3.1.2. YellowFish Adventures

A Yellowfish Adventures foi criada em 2017 através da paixão dos seus fundadores pela natureza e atividades de todo-o-terreno. Com sede e base de operações em Albufeira, a empresa proporciona experiências únicas de contato com a natureza do interior do Algarve, partilhando a beleza natural das paisagens, vilas típicas e locais de interesse histórico e cultural. Todas as atividades de passeio são realizadas a baixa velocidade por forma a que os clientes desfrutem em segurança dos locais e paisagens do percurso e se evite ao máximo perturbar a vida selvagem e residentes locais [**58**].

As experiências vendidas distinguem-se em: *Experience Tour*, *Feel Tour* e *Tour personalizada*. O tipo de experiência, *Experience Tour*, caracteriza-se por uma experiência no interior do Algarve ao volante de um *buggy* ou moto quatro, com uma duração de 90 minutos e passagem por vilas típicas, locais com significado histórico ou cultural, vinhas, pomares, sempre acompanhados por instrutor/guia. À semelhança da *Experience Tour*, a experiência *Feel Tour* representa o mesmo conceito, diferenciando-se na duração da experiência e paragens efetuadas. A sua duração aumenta para três horas e o número de paragens é diversificado, permitindo aos turistas fotografias panorâmicas e banhos em lagoas/praias deslumbrantes. Relativamente à experiência personalizada, permitem ao cliente criar o seu próprio roteiro pelo interior do Algarve, com paragem para almoço,

visitando locais distintos como barragens ou piscinas naturais. Apenas podem usufruir destas experiências clientes em que os condutores disponham de carta de condução com mais de 3 anos e mais de 21 anos de idade.

A venda destas experiências está disponível num website paralelo, sob alçada da YellowFish, com o nome de YellowFish Adventures. Para reservar uma experiência através do website, o cliente primeiramente escolhe a experiência pretendida, seguidamente seleciona a data de inicio e de fim, referente a um intervalo de tempo em que o passeio pode ser realizada, caso exista experiências disponíveis nesse intervalo, o cliente seleciona a data que mais lhe agrada. No próximo passo, o cliente seleciona o tipo de transporte a utilizar no passeio, que pode ser do tipo quad e/ou *buggy*, selecionando também o número de pessoas por veículo. O processo de compra fica completo com a informação pessoal do cliente e respetivo pagamento da reserva. O processo de compra de experiências encontra-se representado na Fig. 3.2. A venda de experiências, de forma semelhante aos transfers, pode também ser vendida por parceiros terceiros e o serviço efetuado apenas pela YellowFish. Em todas as experiências é ainda fornecido, caso necessário, transporte gratuito entre o local de recolha e entrega nas áreas de Albufeira e entre Falésia e Salgados.

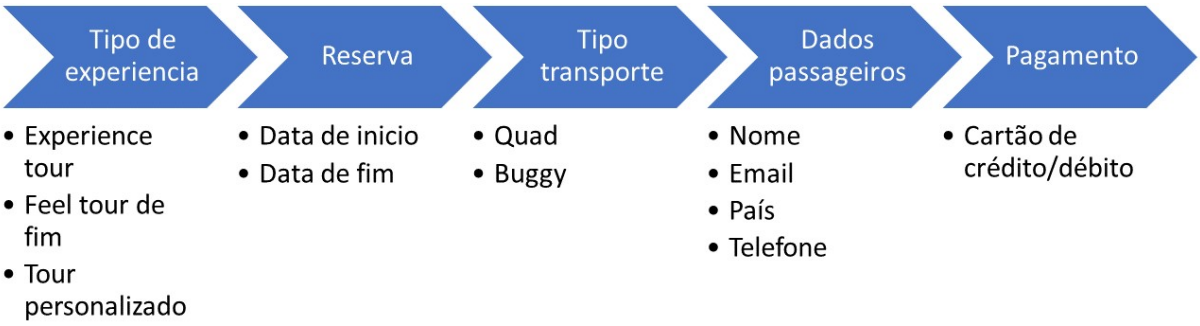


FIGURA 3.2. Processo de compra de transfers online.

3.2. Problema Empresarial

A venda de transfers representa uma grande parte do volume de negócios da Yellowfish. Já a venda de experiências e passeios constitui ainda uma reduzida parcela no seu espetro empresarial. Posto isto, a empresa pretende aumentar a venda de experiências da sua filiada Yellowfish Adventures, através da sugestão de aventuras e experiências aos seus clientes que reservem online um transfer.

Apesar de um serviço de experiência poder ser vendido de uma forma independente aos serviços de transfer, grande parte das vendas de experiências está associada a uma reserva de transfer. Pretende-se desenvolver um modelo de recomendação que consiga identificar padrões e características nos clientes que compram transfers e posteriormente experiências, de forma a generalizar para clientes que apenas compram transfers, sugerindo-lhes experiências adequadas e fomentando, assim, o conceito de *cross-selling* na empresa.

3.3. Conjunto de Dados

Para aplicação de técnicas de ciência dos dados e aprendizagem automática, foram utilizados dados históricos constantes nas bases de dados da YellowFish, utilizadas para registo de reservas de transfers, venda de experiências e demais informação inerente. Os dados de *clickstream*, que se caracterizam pelo registo do comportamento e histórico do cliente, aquando da navegação no website da YellowFish, foram obtidos através da plataforma Google Analytics.

3.3.1. Manifesto

Perante os dados fornecidos pela empresa, assume-se que o modelo de dados que a YellowFish utiliza para o registo de reservas de transfers e experiências, assim como respetivas informações inerentes às mesmas, se caracteriza por 15 diferentes tipos de conjuntos de dados: manifesto, clientes, motoristas, afiliados, reservas de afiliados, visitas de afiliados, reservas canceladas, *feedback*, locais, países, transportes, veículos, voos, reservas de experiências e bagagem. O ficheiro "manifesto" representa o elemento principal que alberga e sumariza as reservas de transfers da empresa, assim como as suas principais características. No entanto, os restantes conjuntos de dados fomentam a informação complementar às reservas registadas no manifesto.

Dos conjuntos de dados mencionados, apenas foram utilizados os dados de manifesto, clientes, países, afiliados, reservas de afiliados, reservas canceladas, voos, reservas de experiências e bagagem. O critério para seleção deste conjunto de dados prendeu-se com o facto de serem os mais úteis e não redundantes para o objetivo do projeto: o desenvolvimento de um sistema de recomendação de experiências e segmentação de clientes. As restantes informações foram desconsideradas para o contexto do problema, dado que não acrescentariam informação viável, nem a nível exploratório, nem para os modelos de *machine learning*. Na figura 3.3 encontra-se representado o modelo de dados que dá suporte aos diversos conjuntos de dados fornecidos.

O conjunto de dados, manifesto, é alimentado pelos diversos conjuntos de dados (tabelas), como ilustrado na figura 3.3 e é constituído, inicialmente, por 1 067 336 registos correspondendo cada entrada a um serviço de ida ou volta, sendo que as viagens de ida e volta representam dois registos na base de dados. Cada registo de reserva no manifesto tem associado: um número de adultos, crianças e bebés; a data-hora em que a reserva foi adquirida no website; o estado da reserva (bookstatus) que indica se a reserva foi ou não realizada; o ID do cliente associado à reserva; o tipo de reserva (código) - que pode ser do tipo OAL (Só ida do aeroporto para uma localização); OLL(Só ida de uma localização para outra localização); RLL(Ida e volta de uma localização para outra localização) ou RAL (Ida e volta do Aeroporto para uma localização) - dia (data de realização do serviço); dropoff (local de chegada); dropoff_gps (coordenadas GPS do local de chegada); dropoff_place_id (ID do local de chegada); flight_nr (número do voo de chegada/partida); fornecedores_id (ID do fornecedor caso tenha realizado o serviço); hora (hora de realização do serviço); motoristas_id (ID do motorista a realizar o serviço); operadores_id

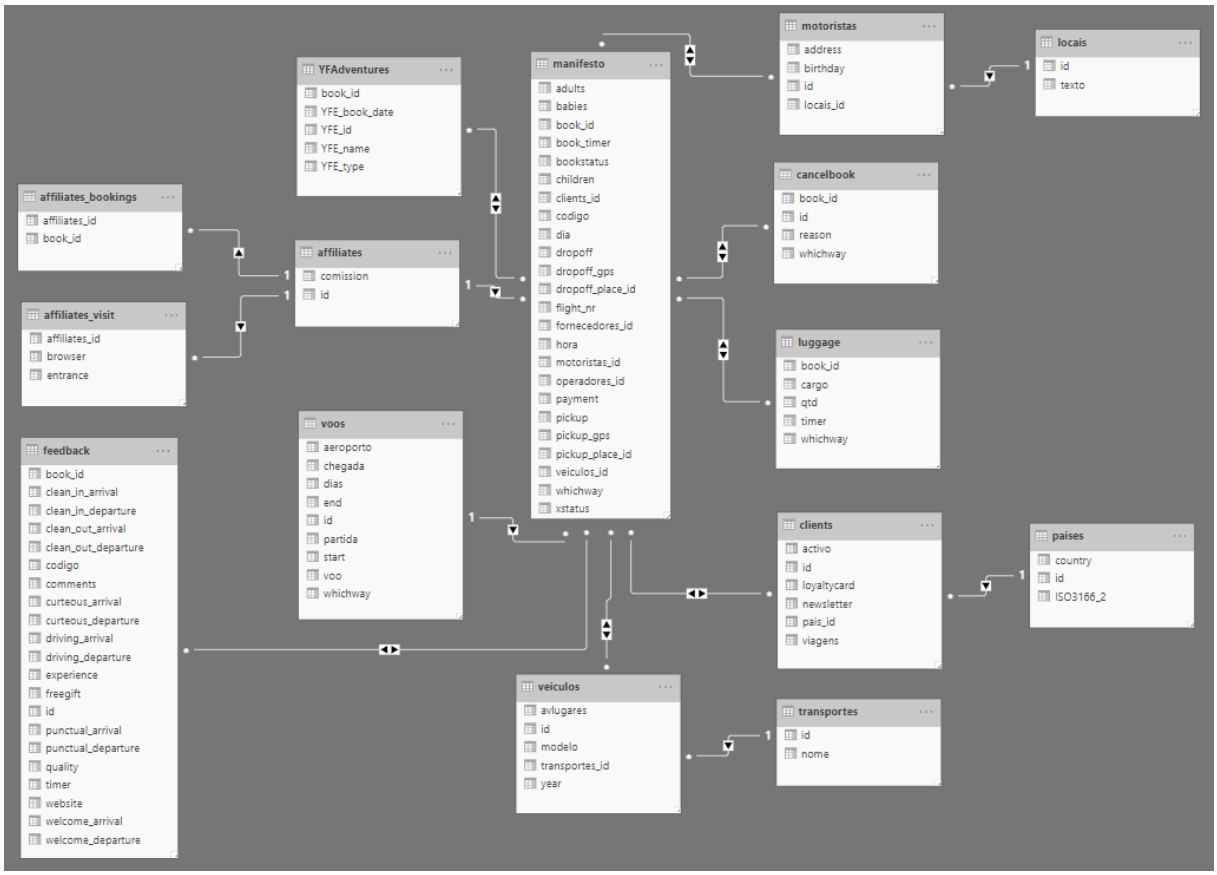


FIGURA 3.3. Diagrama de relação de entidades do conjunto de dados fornecidos pela YellowFish.

(ID do operador que realizou a reserva); payment (forma de pagamento); pickup (local de partida); pickup_gps (coordenadas gps do local de partida); pickup_place_id (ID do local de partida); vehicles_id (ID do veículo a ser utilizado no serviço); whichway (chegada ou partida); xstatus (estado de realização do serviço).

Relativamente ao número de registas presentes nas tabelas secundárias, o conjunto de voos é composto por 7 605 registos, feedback com 257 662 registos, visitas de afiliados com 455 633 registos, reservas de afiliados com 9 346 registos, afiliados com 73 registos, reservas de experiências com 441 registos, veículos com 148 registos, motoristas com 1 958 registos, reservas canceladas com 9 464 registos, bagagem com 629 162, clientes com 316 836 registos, transportes com 2 registos, locais com 224 registos, e países com 206 registos.

Os dados de *clickstream* são caracterizados pelo comportamento, registo de navegação e jornada do cliente aquando da exploração de um website. A YellowFish tem presente o serviço Google Analytics, que permite a visualização analítica e segmentação de clientes para os vários aspetos característicos dos seus clientes.

3.3.2. Google Analytics e REST API

Apesar do Google Analytics fornecer uma plataforma web integrada, que permite a exploração dos dados e padrões sobre os visitante do website, foi utilizado a REST API

fornecida pelo Google Analytics, de forma a se obter esses mesmos dados em larga escala e serem utilizados em Python. Desta forma, é possível haver uma maior manipulação e análise dos dados.

A aplicação de relatórios que a Google Analytics disponibiliza, permite obter dados dos websites associados à aplicação. É possível efetuar *queries* por datas, métricas, dimensões, filtrar, ordenar e segmentar. Dado que estamos a utilizar Python, foi utilizado a API desenvolvendo o código e *queries* personalizadas em ambiente Python. Para isso, foi necessário primeiramente configurar a *Reporting* API do Google Analytics [**59**], ativando a API, instalando a biblioteca cliente e a configuração inicial da amostra. Após a configuração bem sucedida foram levadas a cabo as diversas *queries* para obter os dados de *clickstream* dos clientes da YellowFish Transfers.

Um dos problemas inerentes à utilização gratuita da API de relatórios do Google Analytics, prende-se com a limitação de resultados que podem ser obtidos em cada pedido. Esta limitação existe quando o pedido excede o milhão de registos. Quando esta situação acontece, se houver mais registos para além dos permitidos, o Google Analytics apenas fornece uma amostra dos dados. De forma a ultrapassar este problema, foi desenvolvido código Python que permitisse fazer pedidos de relatórios parciais, que não excedessem o limite, partindo em espaços temporais mais pequenos.

Os pedidos de relatórios são compostos por uma vista (Yellowfish.com ou Yellowfish.co.uk), limites de datas, métricas e dimensões. Forçamos ainda a o nível de amostra para "grande"e o tamanho da página para 100000 de forma a minimizarmos a possível amostragem que a Google Analytics fornece-se nos relatórios.

CAPíTULO 4

Pré-processamento e Análise Exploratória

Neste capítulo são descritos os procedimentos para pré-processamento dos dados, tais como, limpeza, integração, seleção e transformação dos dados, a fim de serem submetidos aos modelos de aprendizagem automática. É ainda apresentada uma análise exploratória e descritiva dos dados, avaliando os principais padrões encontrados nos dados de reserva e dados de *clickstream*. Nesta análise descritiva são ainda desenvolvidas e descritas técnicas para segmentação de clientes nos dados de reservas da YellowFish de forma a avaliar as características dos clientes que compram transfers.

4.1. Limpeza dos Dados

A primeira operação de limpeza de dados passou pela substituição de todas as células que apenas continham espaços em branco, substituindo-o pelo valor NaN (*Not a Number*), em todos os conjuntos de dados. Com isto foi possível a unificação de algumas registos duplicados que apenas continham algumas células diferenciadas pelos espaços em branco. Pelo lado oposto, existiam registos duplicados apenas diferenciados em determinadas colunas descritivas dos dados. As colunas 'bookstatus' e 'xstatus' contêm a indicação do estado da reserva e da realização do serviço, respetivamente 0 - por efetuar, 1 - efetuada(o) e -1 - não efetuada(o). Assim, todos os registos duplicados erradamente, diferenciados apenas pelos valores NaN em determinadas colunas, foram removidos.

Ainda relativamente a registos duplicados, assinalou-se que haviam registos que apenas se diferenciavam em algumas colunas como o "*bookid*" e a data/hora da compra. Removeram-se estes registos, que apenas acrescentavam ruído aos dados, através da verificação de duplicados tendo em conta o agrupamento pelo identificador numérico do cliente, tipo de reserva (OAL, OLL, RLL, RAL), local de destino, dia da reserva e hora da reserva. Sob forma de mantermos apenas um registo para cada reserva de transfer no conjunto de dados, eliminaram-se todas as reservas duplicadas e mantendo apenas o registo com a book_id mais elevada, salvo quando a mesma reserva de transfer (book_id) reservou mais do que uma experiência.

Relativamente a registos com valores negativos em algumas colunas, como é o caso do número de adultos, crianças, bebés e book_id, foi considerado aplicar o respetivo valor absoluto para cada um dos registos, dado que foram apenas registados casos esporádicos de valores negativos que, ao modificar para o seu valor absoluto, faria sentido na reserva. Por exemplo: o valor de -12 adultos numa reserva foi convertido em 12 adultos. Foi ainda aplicada uma filtragem nas reservas, considerando apenas aquelas com menos de 40 adultos. Mesmo tendo em conta que a YellowFish trabalha com grandes grupos de turistas,

foi decidido que não se justificava manter o pequeno número de registos contendo reservas com mais de 40 adultos que foram, portanto, considerados *outliers*. Para uniformização do conjunto de dados manifesto, converteu-se o texto de todas as células com valores nesse formato para maiúsculas e foram removidos os registos duplicados, neste caso, que apenas diferiam em algumas células por letras maiúsculas/minúsculas.

4.2. Integração de Dados

A principal operação de integração de dados recaiu sobre a fusão de todos os conjuntos de dados num único ficheiro, tendo em conta o ficheiro manifesto e reunindo toda a informação sobre as reservas de transfers num único conjunto. Desta forma, foi possível integrar e obter mais informação proveniente dos outros conjuntos de dados, sobre as reservas e experiências, trabalhando apenas com um conjunto de dados que agrega todas as informações inerentes às reservas.

O ficheiro de dados clientes representa a informação para cada um dos clientes que efetuou uma reserva com a YellowFish, tal como, o ID do cliente, o ID de País, *newsletter* (1/0), número de loyalty card, número de viagens já efetuadas com a empresa e, por último, se se encontra ativo como cliente. Relativamente ao conjunto de dados de reservas de afiliados, este indica quais foram as reservas que foram feitas através de um dos afiliados associados à empresa. Quanto aos dados de reservas canceladas, indicam quais das reservas foram canceladas, assim como a sua razão. Por último, o conjunto de dados de países indica, para cada ID de país qual o seu nome e nomenclatura ISO. Estes conjuntos de dados foram adicionados ao manifesto através da sua conjugação, utilizando para isso campos comuns em ambos os conjuntos de dados, como é o caso de bookid.

Dado as características de um transfer de golfe, que apenas acontece esporadicamente e em situações específicas de lazer, decidiu-se classificar todos estes eventos como registos de experiências e não como registos de transfer, sob forma de conseguirmos identificar e recomendar experiências golfe aos clientes que compram transfers. Desta forma, não só temos mais uma classe para classificar no sistema de recomendação, como aumentamos o número de registos (observações) de experiências.

Após uma integração centralizada num único ficheiro com toda a informação sobre as reservas presentes no manifesto, verificou-se que algumas reservas continham o código de reserva diferente para a partida e/ou chegada. Partindo do pressuposto que uma reserva com o código 'RAL' - Return Airport to Location, implica um serviço de transfer de ida e volta, não podendo existir para a mesma reserva de transfer códigos diferentes para partidas e chegadas. Dado a pouca quantidade de registos nesta situação, foi decidido eliminar estas reservas do conjunto de dados, dado que apenas acrescentaria ruído.

Após as tarefas de limpeza e integração de dados, o conjunto de dados unificado ficou com 303 993 registos.

Relativamente aos dados de *clickstream*, foram obtidos 3 ficheiros através de relatórios presentes no Google Analytics, sendo estes: os dados de *clickstream* por ID de utilizador

visitante; clientes que compraram uma experiência golfe por ID de utilizador visitante; clientes que compraram transfers por ID de utilizador visitante.

Como primeira abordagem eliminaram-se todos os registos duplicados de ID's dos motores de busca nos ficheiros de clientes de transfers e experiências, para que apenas existam ID's únicos. Seguidamente fazemos uma fusão entre os dados de clickstream, clientes de transfers e clientes de experiências golfe. Desta forma, é possível identificar a jornada dos clientes através das várias páginas visitadas.

4.3. Análise Descritiva

4.3.1. Manifesto

O manifesto representa os dados associados às reservas de transfers e experiências, assim como, informação inerente aos clientes. De forma a sumarizar e analisar estes dados, foram utilizadas técnicas de ciência dos dados para providenciar características, padrões e análises exploratórias dos mesmos. Numa primeira análise relativamente ao número de serviços de transfers registados, verifica-se que existe uma tendência crescente na procura destes serviços desde 2012 (ver Fig. 4.1). No entanto, no ano de 2019 verificou-se uma diminuição nesta tendência crescente que poderá ser explicada pela indefinição do *Brexit*, dado que os clientes do Reino Unido representam quase metade das reservas.

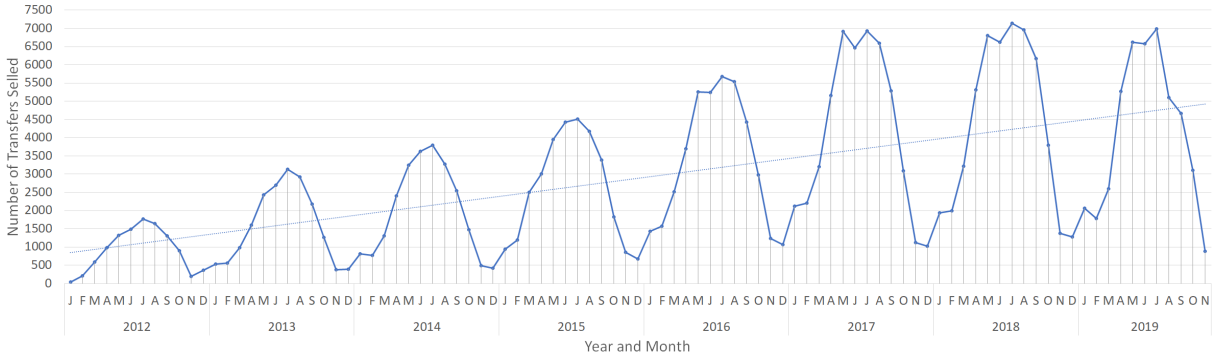


FIGURA 4.1. Número de reservas de transfers entre 2012 e 2019.

O Reino Unido representa grande parte das reservas de serviços de transfers, seguido da Irlanda (Fig. 4.2). Os turistas de ambos os países preferem viajar nas estações do ano Verão e Outono. O número total de reservas sem país associado representa o terceiro valor mais elevado entre o número total de serviços de transfers. As reservas de transfers sem país associado foram principalmente registadas nas estações do ano Verão e Primavera. Dado o número significativo, decidimos manter estes registos para análise. França, Portugal e Alemanha, apesar de representarem um volume de compra muito menor, são os três países, para além de Reino Unido e Irlanda, que mais reservam serviços de transfer.

Pretende-se tirar partido do elevado número de reservas efetuadas de serviços de transfers para vender experiências. Através dos dados fornecidos é notória a diferença de registos entre a venda de transfers e experiências. Foram registados desde Janeiro de 2012, cerca de 300 mil serviços de transfers, e em contrapartida, apenas vendidas 2504 experiências desde Fevereiro de 2012, sendo elas, 2077 do tipo golfe, 281 do tipo *Feel Tour*, 131

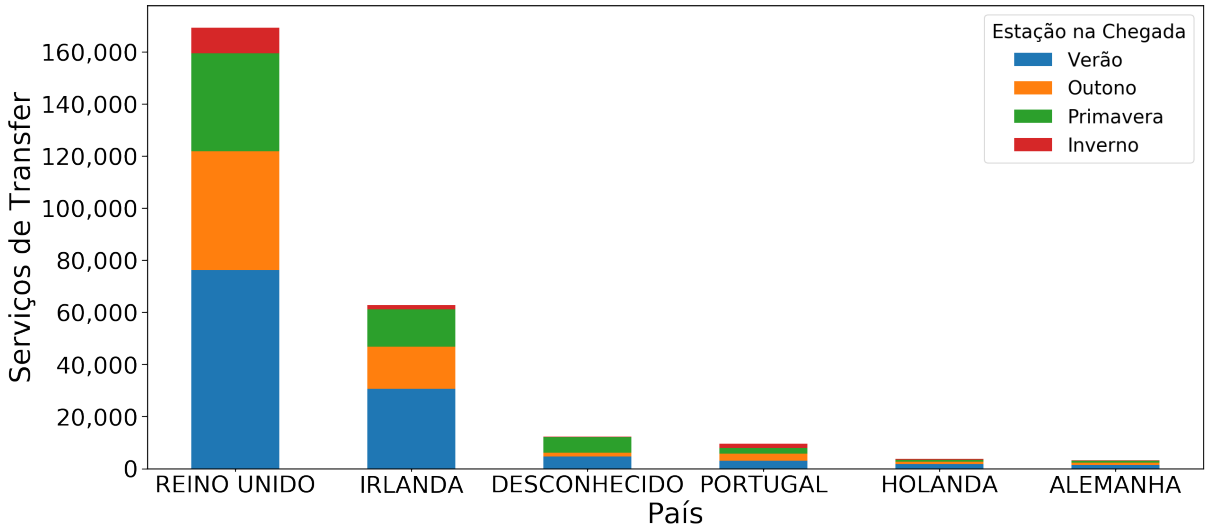


FIGURA 4.2. Número de serviços de transfer vendidos por país dos clientes e estação na chegada.

do tipo *Experience Tour*, e 15 do tipo *Feel Tour + Caves & Coastline*. Esta particularidade, origina observações extremamente desequilibradas, em que a venda de experiências, apenas representam cerca de 0,20% do número total de reservas. Com um rácio de 1 experiência vendida para cada 121 transfers vendidos. Como verificado na Fig. 4.3, os clientes compram mais experiências do tipo golfe e são maioritariamente provenientes do Reino Unido e Irlanda.

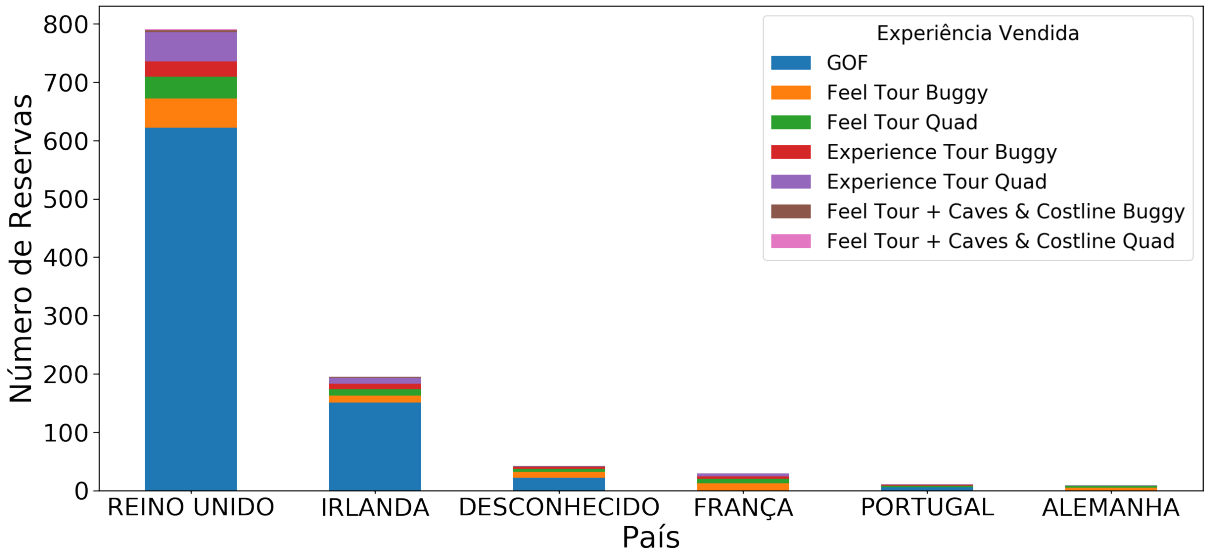
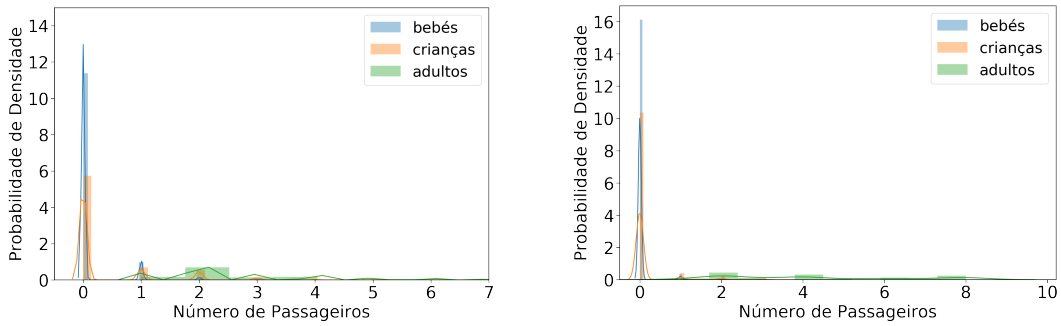


FIGURA 4.3. Número de serviços de experiências vendidos por país dos clientes e tipo de experiência.

Como constatado na Fig. 4.4, que representa a distribuição de passageiros nos serviços de chegadas e partidas, a maioria das reservas de transfer são reservadas para dois ou quatro adultos, com zero a duas crianças, e zero a um bebé. As reservas associadas a um número mais elevado de adultos estão ligadas, geralmente, a jogadores de golfe, dado que

muitas delas têm como destino um campo de golfe ou têm bagagem de golfe associada. Em relação às companhias aéreas, os clientes tendem a viajar utilizando empresas de baixo custo, que representam mais de 60 % das observações.



(A) (B)

FIGURA 4.4. Distribuição de passageiros nos serviços de transfer e experiências: a) transfers; b) experiências.

Tendo em conta a venda de experiências a clientes que compraram serviços de transfer à empresa YellowFish, analisámos os padrões destes clientes, de forma perceber se existem características padrão nos clientes de experiências. Na Fig. 4.5 encontra-se ilustrada a distribuição normalizada de adultos, crianças e bebés associadas à venda de experiências. Grande parte dos clientes de transfers que compram experiências estão associados a reservas com poucas ou nenhumas crianças/bebés. Verificou-se que grande parte das experiências vendidas são provenientes de serviços de transfers com um número superior a 3 adultos.

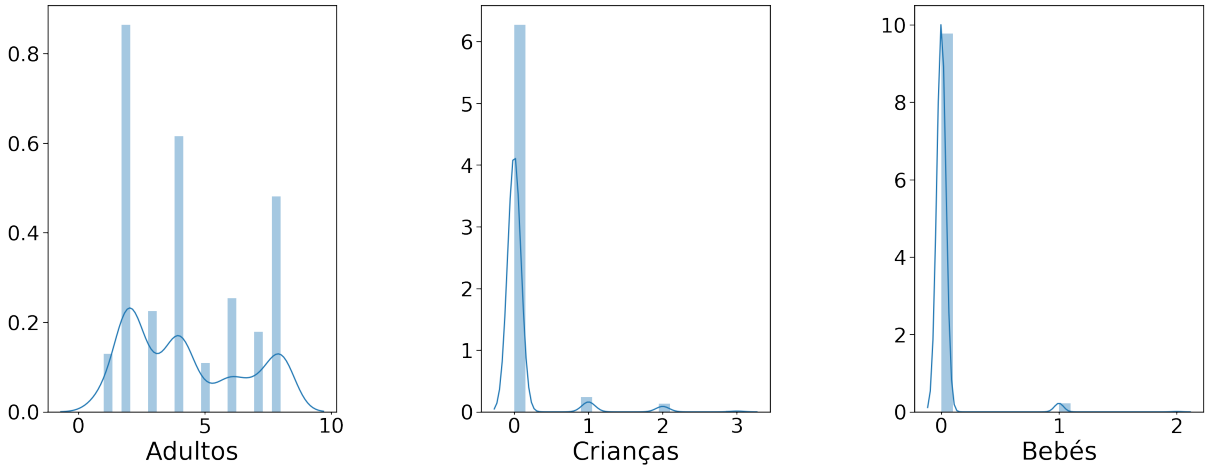


FIGURA 4.5. Distribuição de adultos, crianças e bebés na compra de experiências por país.

A compra de experiências pode ser efetuada antes ou depois da compra de um serviço de transfer. A compra de serviços de experiências antes do serviço de transfer pode implicar que o cliente é motivado principalmente pela experiência e não pelo serviço de transfer em si, sendo o transfer apenas o meio para chegar até ao local da experiência ou

ao hotel de destino. A Fig. 4.6 apresenta o número de reservas de experiências vendidas antes de ser comprados os serviços de transfer associados. Verifica-se, novamente, que os países Reino Unido e Irlanda são os que mais compram experiências antes de ser comprado o serviço de transfer, sendo os serviços Golfe e *Experience Tour Quad*, os mais vendidos, respetivamente.

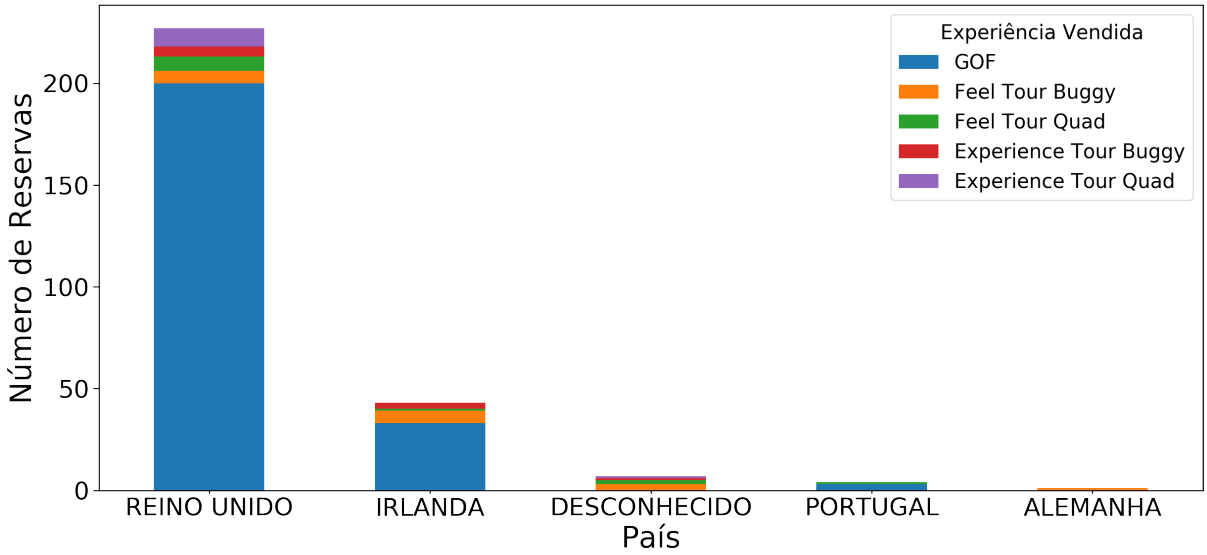


FIGURA 4.6. Número de experiências compradas antes da compra do serviço de transfer.

Analisando os serviços de experiência que foram comprados após a compra do serviço de transfer, verifica-se que existe um maior número total de experiências vendidas após a compra do serviço de transfer. A Fig. 4.7 mostra que grande parte das experiências são compradas depois da compra de um serviço de transfer. Reino Unido e Irlanda continuam a destacar-se como os países que mais compram serviços de experiências, sendo os serviços Golfe e*Experience Tour Quad Experience Tour Quad*, os mais vendidos, respetivamente.

O aeroporto de Faro é o principal ponto de partida dos serviços de transfer (Fig. 4.8). Os clientes de serviços de transfer tendem a utilizar maioritariamente companhia aéreas *low-cost*, sendo a Ryanair e a Easyjet as companhias mais utilizadas para se deslocarem até ao aeroporto de Faro. As localizações Albufeira, Vilamoura, Alvor, Lagos e Carvoeiro são outras das localizações mais requisitadas de carga para serviços de transfer, contudo, com frequência muito menor comparado às cargas no aeroporto de Faro.

A Fig. 4.9 ilustra os locais mais frequentes de destino dos turistas em função das companhias aéreas em que chegaram. Albufeira e Vilamoura são os locais mais frequentes de destino final para os turistas que procuram serviços de transfer, sendo maioritariamente turistas que viajam em companhias *low-cost*, como a Ryanair e Easyjet. O aeroporto de Faro, sendo o segundo principal local de descarga, indica os serviços de transfer de regresso dos turistas de volta ao aeroporto, no fim da sua estadia.

Dado que muitas das vezes os serviço hoteleiro e mais especificamente, os serviços de transfers estão mais ligados a famílias ou grupos organizados para fazerem determinadas

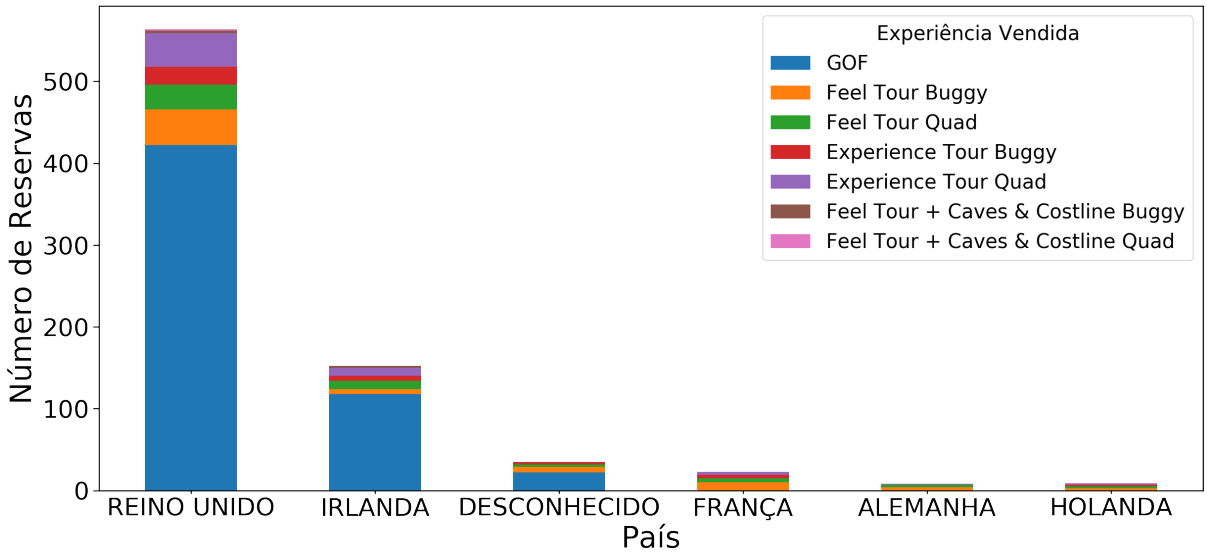


FIGURA 4.7. Número de experiências compradas depois da compra do serviço de transfer.

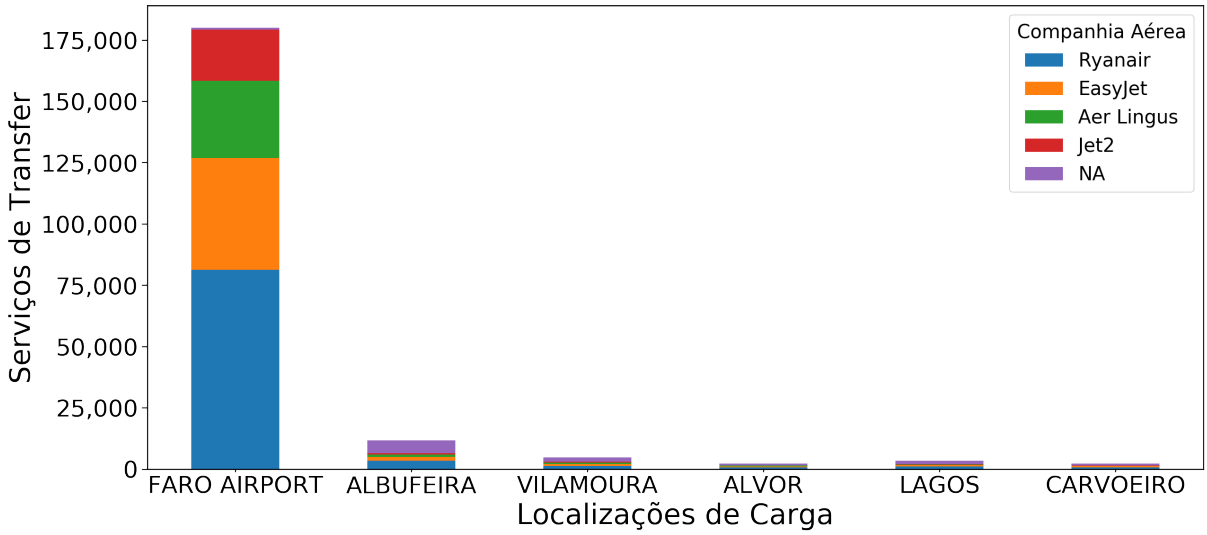


FIGURA 4.8. Locais mais frequentes de carregamento dos clientes de transfers.

viagens, analisamos os padrões destes grupos nos dados do manifesto. Para isso, nesta análise foram considerados apenas dois grupos: famílias ou grandes grupos. As reservas de famílias podem ser definidas contendo um adulto ou mais, um bebé ou mais ou uma criança ou mais. Exemplo: Grupo de dois adultos, um bebé e/ou uma criança. As reservas de grandes grupos incluem mais de cinco adultos, sem considerar crianças ou bebes, exemplo: Grupo de dez adultos sem crianças. Segundo a opinião da empresa, existem grandes grupos que se organizam para jogar golfe no Algarve e utilizam os serviços da YellowFish Transfers para se deslocarem, interessando à empresa compreender um pouco mais acerca destes clientes.

Como indicado anteriormente, a venda de serviços de transfers pela empresa YellowFish transfers tem vindo a aumentar de ano para ano, sendo que os clientes tendem a

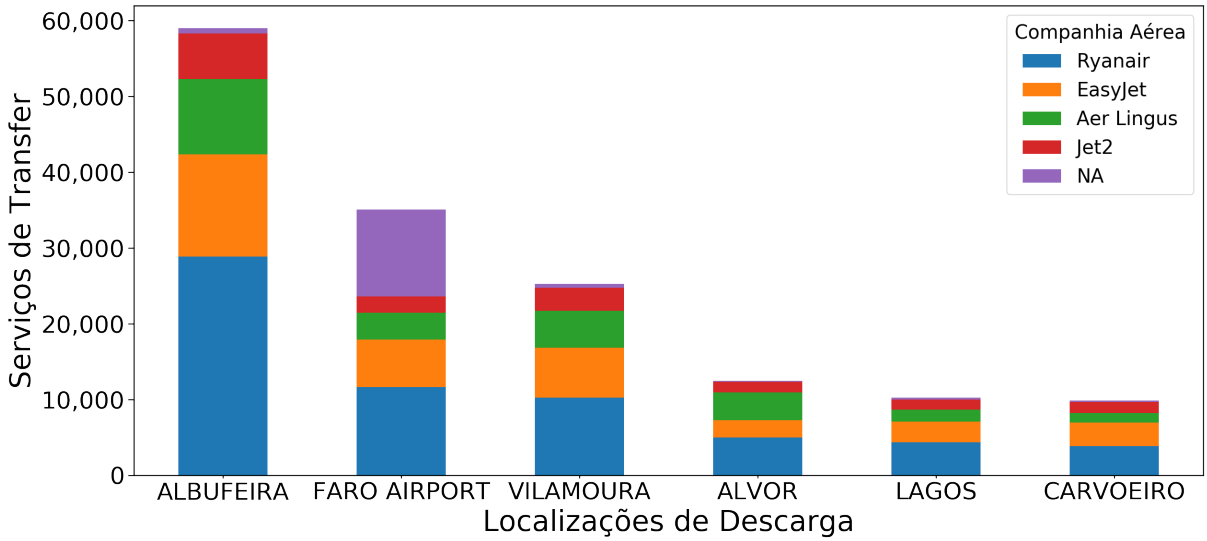


FIGURA 4.9. Locais mais frequentes de descarregamento dos clientes de transfers.

reservar nos primeiros dias de cada semana e mais frequentemente entre os meses de Março e Setembro. A Fig. 4.10 explica visualmente esta variação das reservas ao longo dos anos, meses e dias da semana.

4.3.2. Google Analytics

Os clientes que visitam o website da YellowFish para efetuar uma reserva de transfer ou golfe, têm a sua jornada registada na plataforma Google Analytics. As visitas ao website têm aumentado exponencialmente, permitindo, de uma forma clara, percecionar e quantificar a eficácia do website para os demais objetivos da empresa.

A Fig. 4.11 demonstra o número de visitas por dia recebidas no website da Yellowfish Transfers entre fevereiro e março de 2019. Os picos, que de uma forma consistente aparecem ilustrados na Fig. 4.11, foram analisados sob forma de se perceber se existe um padrão inerente. Estes picos semanais ocorrem às segundas-feiras, onde são registados o maior número de visitas, e os valores mais baixos semanais são registados aos sábados.

Como expectável, o Reino Unido mantém-se como sendo o país de onde são originários a maior fatia de visitantes das páginas da Yellowfish, com cerca de 248 mil visitas desde Fevereiro até Junho de 2019. Logo de seguida, a Irlanda, com cerca de 118 mil visitas e depois, por ordem, Portugal com cerca de 44 mil visitas, Estados Unidos com cerca de 9800 visitas, Alemanha com 7160 visitas, Holanda com 6313 visitas e França com 5900 visitas, são os países com mais visitas ao website. No entanto, a página é visitada por pessoas de quase todos os países do mundo. A Fig. 4.12 apresenta os dez países que mais visitam o website da YellowFish Transfers, ordenados por ordem descendente.

Cerca de um quarto do número total de visitantes na página da YellowFish Transfers converte através da compra de um serviço de transfer. Na Fig. 4.13 verifica-se o número de visitantes a nível mensal e as consequentes receitas geradas com base no número de visitas que converteram, comprando um serviço de transfer. Nos meses de março, abril,

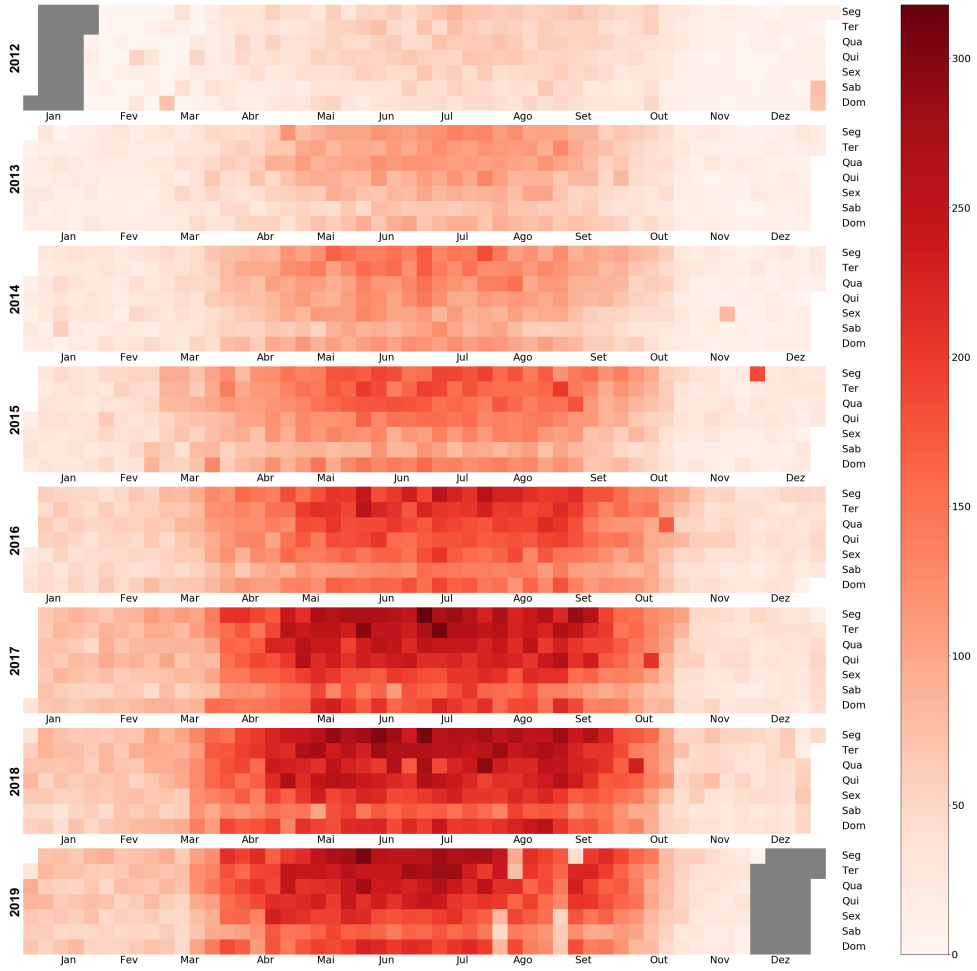


FIGURA 4.10. Mapa de calor tendo em conta a marcação de reservas anuais, por mês e dia da semana.

maio, junho e julho, foram registadas taxas de conversão de 13%, 15%, 17%, 16% e 22%, respetivamente.

Os visitantes do website da Yellowfish que utilizam sistemas operativos Windows ou iOS são os que mais reservas fazem no site website. Na Fig. 4.14 encontra-se ilustrado a receita gerada por sistema operativo do cliente e parte do dia em que foi comprada a reserva de serviço de transfer. Os horários para as partes do dia são definidos da seguinte forma: manhã (entre as 5 e as 11h), almoço (entre as 12 e as 14h), tarde (entre as 15 e as 17h), anoitecer (entre as 18 e as 22h) e noite (entre as 23 e as 4h). A parte do dia em que é gerada mais receita é durante o período da manhã.Contudo, de um ponto de vista geral, as melhores alturas do dia a nível de vendas são durante a manhã e durante o anoitecer em que são geradas mais receitas por diferentes tipos de sistemas operativos.

A Yellowfish faz publicidade em vários canais online, originando tráfego de várias fontes para o seu website. Grande parte dos utilizadores que visita o website é proveniente do motor de busca da Google, através da procura voluntária e sem pagamento pela empresa para tal (Fig. 4.15). A segunda origem da pesquisa aparece no canal (direct), que indica tráfego onde a referência ou fonte é desconhecida.

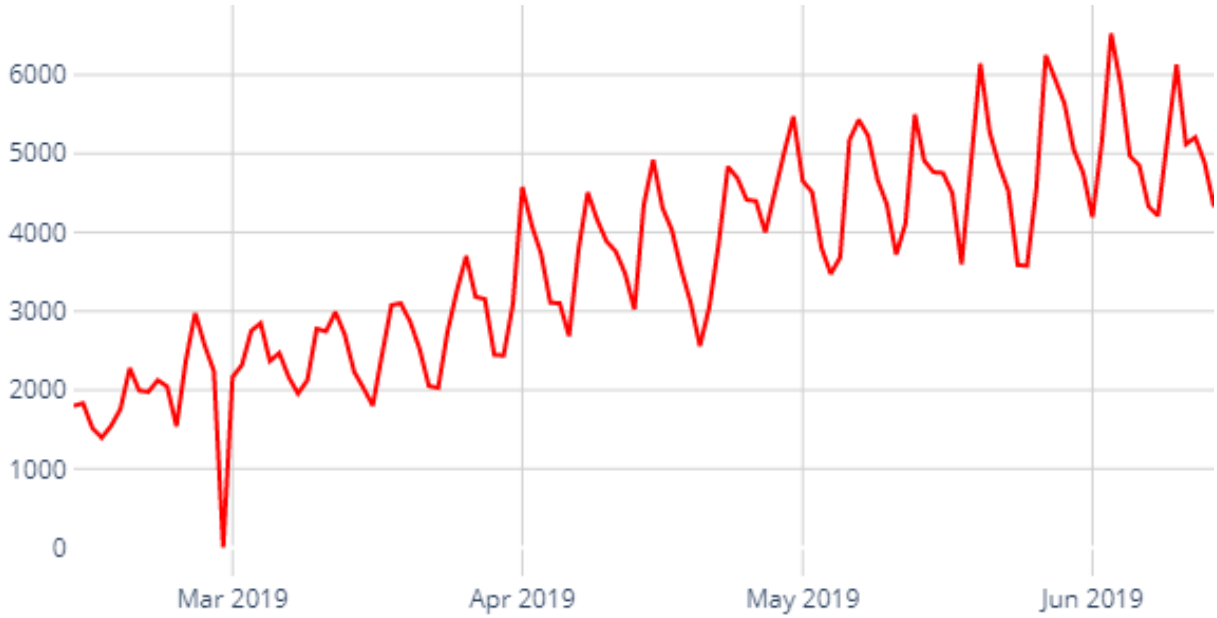


FIGURA 4.11. Visitas por dia no website da YellowFish de fevereiro a março de 2019.

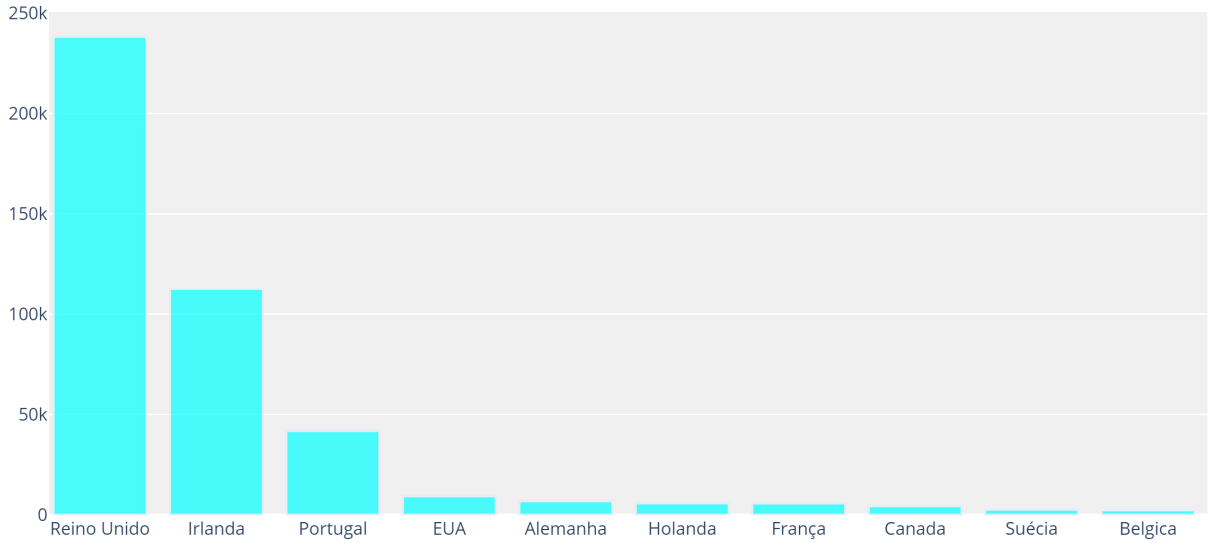
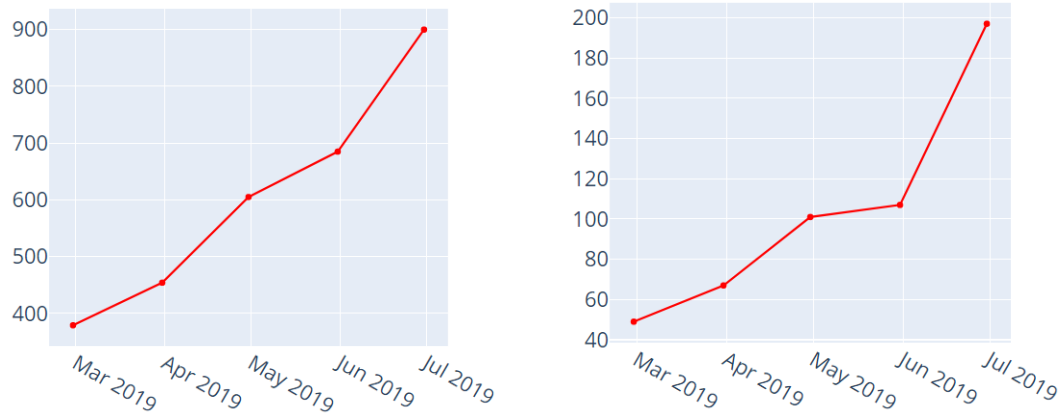


FIGURA 4.12. Gráfico de barras com o número total de visitas online por país de Fevereiro a Março de 2019.

Para além da fonte orgânica (Google) ser a principal fonte de visitas no website da Yellowfish, verifica-se, também, que grande parte das visitas ocorrem no inicio da semana, ou seja, durante domingo, segunda e terça.

4.4. Seleção de Dados

A seleção de dados tem como objetivo identificar os dados que possam explicar a variável alvo, neste caso, a aquisição de experiências. Com o conhecimento obtido das análises anteriormente realizadas, é possível selecionar os dados que devem ser considerados com



(A) (B)

FIGURA 4.13. Número de visitas por mês no website da *yellowfishtransfers.com* que geraram receita no intervalo de Fevereiro a Março de 2019. (A) Número de visitantes únicos. (B) Número de visitantes únicos que converteram.

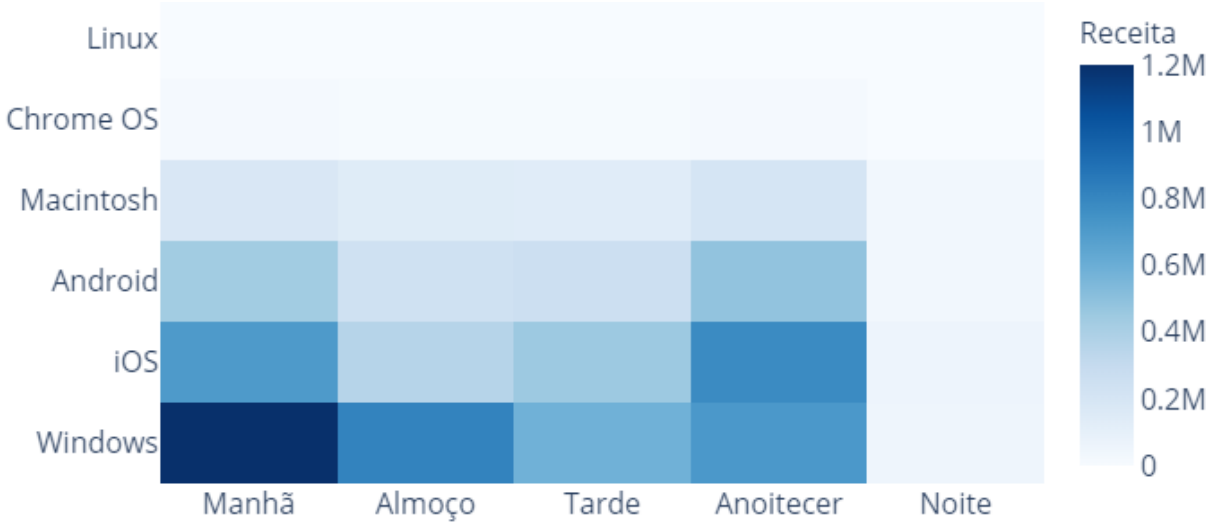


FIGURA 4.14. Geração de receita por dispositivo e parte do dia de Fevereiro a Março de 2019.

maior poder explicativo e assim evitar dados que podem produzir ruído para as futuras análises.

Dado que o objetivo da investigação é desenvolver um sistema de recomendação que consiga prever se um cliente de transfer vai ou não comprar um serviço de experiência e, ainda, sugerir uma experiência específica, vão ser utilizados os dados do manifesto contendo apenas informação sobre a reserva do serviço de transfer com a associação do *cluster* identificado anteriormente. A compra ou não de uma experiência, tendo em conta cada observação de reserva de transfer, será a variável alvo.

Não foi possível utilizar os dados do Google Analytics (*clickstream*) dado que não existe uma associação entre uma reserva registada no manifesto e os registos de *clickstream*. Isto deve-se às políticas restritivas do Google Analytics para utilizadores que não

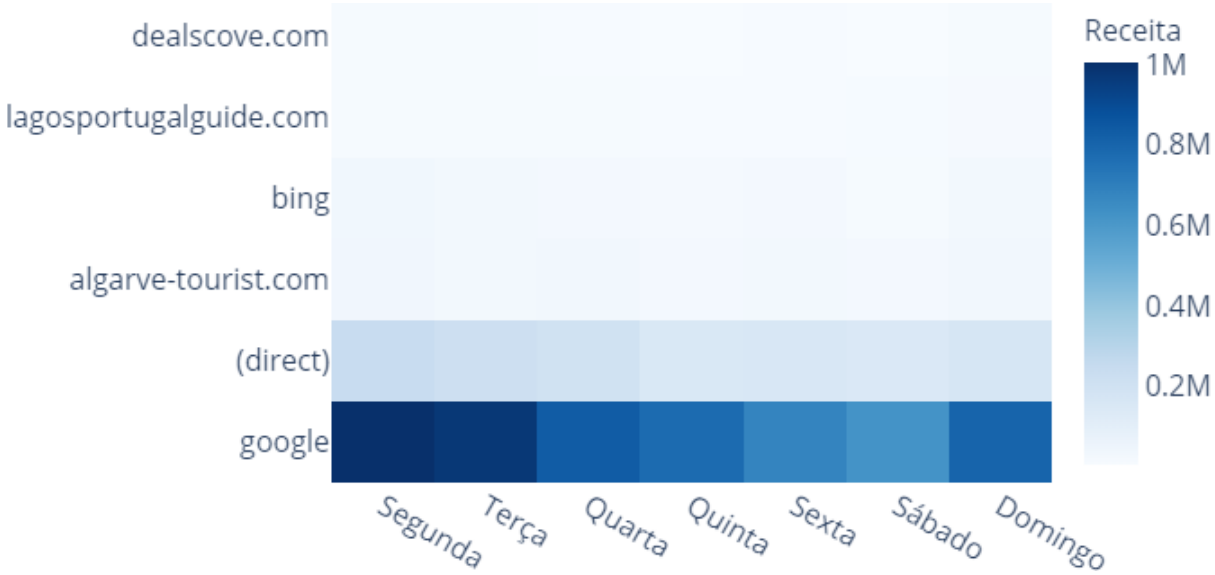


FIGURA 4.15. Receita total por fonte de tráfico e dia da semana de Fevereiro a Março de 2019.

sejam *premium* Google Analytics 360, uma vez que restringem o número de amostras que podem ser exportadas das suas bases de dados. No entanto, o número elevado e qualidade das *features* encontradas no manifesto e que caracterizam cada uma das reservas de transfers e experiências, permitem uma solução viável para o desenvolvimento do sistema de recomendação. A utilização de técnicas de *feature engineering* permitiu que se aumentasse a quantidade de *features* disponíveis de forma a serem utilizadas no sistema de recomendação. Na subsecção seguinte é explicado o processo de criação das novas *features*, a designação de cada delas e a importância das mesmas.

4.4.1. *Feature Engineering*

A construção de novas *features* a partir de informação inerente ajudam a que os modelos de aprendizagem automática consigam melhorar o seu desempenho preditivo [**60**].

Todas as colunas do tipo "data"foram convertidas em novas colunas de dados, passando a conter informação do dia da semana, dia do mês, mês, ano, quarto do ano e estação do ano. Da mesma forma, as colunas com horas foram convertidas em informação mais generalizada, correspondendo à parte do dia a que a hora diz respeito, isto é, entre as 5 e as 11 horas da manhã, inclusive, considera-se "manhã", entre as 12 e as 14h, considera-se "almoço", entre as 15 e as 17h considera-se "tarde"e entre as 18 e as 22h considera-se "anoitecer". Todos os restantes casos são considerados "noite". Desta forma, foi possível criar, para as datas e horas das reservas e serviços (chegada ou partida) de transfers ou experiências as *features* mencionadas.

Através de uma abordagem cumulativa, quantificaram-se todas as viagens que cada cliente efetuou na YellowFish Transfers, fornecendo, desta forma, o grau de frequência com que cada cliente utiliza os serviços de transfer. Para se compreender se os clientes reservam com antecedência os seus serviços de transfer, criou-se uma nova *feature* que

calcula quantos dias antecedem a realização do serviço da reserva, possibilitando analisar que características diferenciam os clientes que compram com antecedência e ainda correlacionando com a compra de experiências. Partindo da hipótese que pode haver correlação entre um feriado local com a chegada a um destino turístico, decidiu-se criar uma *feature* que conseguisse encontrar o feriado mais próximo para o país de cada reserva, tendo em conta a primeira data de realização do serviço. Para isso, foi necessário utilizar a biblioteca *holidays* [**61**], que armazena todos os feriados em diversos anos, para um conjunto alargado de países. Apesar de alguns países constantes nas reservas não estarem presentes na biblioteca *holidays*, apenas uma pequena percentagem da marcação de reservas é que se encontrava nesta situação, não havendo lugar a prejuízo para o objetivo do problema.

Esta preparação dos dados resultou na criação das seguintes *features*:

- **country_code**: abreviação do nome do país de cada reserva.
- **nearest_arrival_holiday**: dias para o feriado mais próximo, tendo em conta o dia do serviço de transfer de ida e o país do cliente.
- **nearest_departure_holiday**: dias para o feriado mais próximo, tendo em conta o dia do serviço de transfer de regresso e o país do cliente.
- **book_part_of_day**: parte do dia em que foi feita a reserva do transfer [noite; manhã; almoço; tarde; madrugada].
- **book_weekday**: dia da semana em que foi feita a reserva do transfer [0-6].
- **book_day**: dia do mês em que foi feita a reserva do transfer [1-31].
- **book_month**: mês em que foi feita a reserva do transfer [1-12].
- **book_year**: ano em que foi feita a reserva do transfer [2012-2020].
- **book_quarter**: quarto do ano em que foi feita a reserva do transfer. [1-4].
- **book_season**: estação do ano em foi feita a reserva do transfer [1-4]
- **arrival_quarter**: quarto do ano aquando da realização do serviço de ida do transfer [1-4].
- **arrival_season**: estação do ano aquando da realização do serviço de ida do transfer [1-4].
- **departure_quarter**: quarto do ano aquando da realização do serviço de regresso do transfer [1-4].
- **departure_season**: estação do ano aquando da realização do serviço de regresso do transfer [1-4].
- **YFE_book_day**: dia do mês em que foi feita a reserva da experiência [1-31].
- **YFE_book_weekday**: dia da semana em que foi feita a reserva da experiência [0-6].
- **YFE_book_month**: mês em que foi feita a reserva da experiência [1-12].
- **YFE_book_quarter**: quarto do ano em que foi feita a reserva da experiência [1-4].
- **YFE_book_part_of_day**: parte do dia em que foi feita a reserva da experiência [noite; manhã; almoço; tarde; madrugada].

- **viagens**: número total de serviços de transfer que o cliente já efetuou numa determinada data.
- **lead_time**: número total de dias entre a reserva e o serviço de transfer de ida.
- **dias_de_estadia**: número total de dias entre a chegada e a ida do cliente.
- **last_service**: número total de dias entre a reserva mais recente e a reserva anterior.
- **antiguidade**: número total de dias de existência do cliente na empresa.

4.5. Segmentação de Clientes no Manifesto

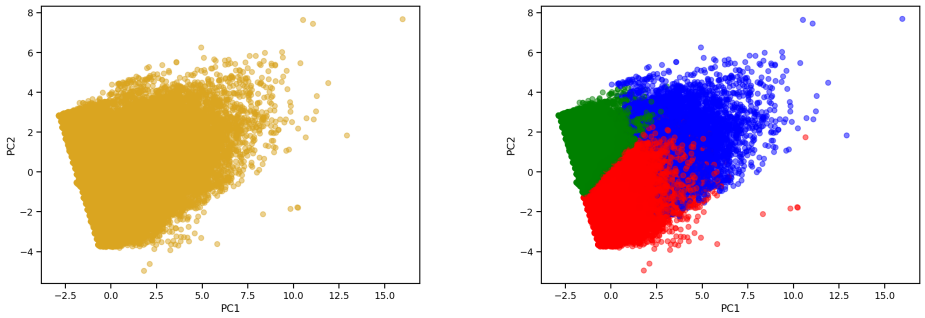
A segmentação de clientes consiste em agrupar clientes com comportamentos, dados demográficos e interesses semelhantes. A técnica permite que se personalize a experiência para cada um dos perfis de agrupamento, explorando o conteúdo relevante e os aspectos que estes partilham [**62**]. Grandes quantidades de informações sobre os clientes permitem oferecer uma personalização da jornada dos clientes única. Desta forma, é possível alcançar maior número de utilizadores, aprimorando a experiência e a fidelidade do cliente e otimizando as vendas [**63**].

Tendo em conta os dados de vendas de transfers, pretende-se distinguir e caracterizar os diferentes tipos de clientes, assim como, descobrir novas *features* com o processo. A segmentação de clientes vai permitir à empresa de transfers adaptar os seus produtos e serviços, a fim de melhor satisfazer os clientes, bem como melhorar a oferta de produtos para cada grupo de clientes. A segmentação de clientes torna o processo de compra mais rápido, ajudando a construir lealdade para com os clientes quando baseada na interação relevante dos clientes durante a sua jornada [**64**].

Para a realização de segmentação de clientes foi utilizada a técnica de Análise de Componentes Principais, PCA (do inglês, Principal Components Analysis) com dados uniformizados, seguido da aplicação do método K-means [**65**] para realizar o agrupamento. A partir da análise PCA, foram selecionados os primeiros 6 componentes principais, uma vez que capturavam 80% da variância dos dados. Numa primeira análise, e visualizando os dois primeiros componentes da análise PCA (Fig. 4.16 (a)), não existe nenhuma separação significativa entre os pontos dos dados. De forma a inferir, aproximadamente, quantos *clusters* poderiam existir no conjunto de dados, utilizou-se a heurística "Método do Cotovelo" [**66**]. O método apontou para $K = 3$ *clusters* como a melhor escolha, estando os resultados subsequentes da aplicação do K-means nos dados do PCA ilustrados na Figura 4.16 (b).

A Figura 4.17 ilustra os *clusters* encontrados, mostrando diferentes características quantitativas que caracterizam o comportamento e o contexto demográfico dos clientes de que reservam serviços de transfer.

Analisando as características dos clientes nos três grupos indicados pelas técnicas PCA e K-means, identificou-se que: o primeiro *cluster* (a vermelho) representa os clientes que costumam reservar o serviço de transfer em maio, chegam em junho e têm uma estadia média de seis dias. Estes clientes geralmente reservam com um mês de antecedência e



(A) (B)

FIGURA 4.16. (a) Análise PCA dos dois primeiros componentes; (b) Predição utilizando o algoritmo K-means (K= 3).

tendem a fazer novas reservas de dois meses em dois meses. As reservas de transfer deste grupo de clientes, têm em média três adultos, sem bebés ou crianças, e trazem três itens de bagagem. O segundo *cluster* (verde) representa os clientes que reservam mais tarde (agosto) e com um tempo de antecedência mínimo. Geralmente reservam e chegam em agosto com um tempo de antecedência de 11 dias. Estes clientes ficam por pouco tempo, em média três dias, mas compram serviços de transfer com mais frequência. Viajam com menos bagagem, com uma média de dois adultos e sem bebés ou crianças nas suas reservas. O último *cluster* (azul), representa os clientes que viajam com crianças ou bebés e trazem bagagem mais variada. Estes clientes geralmente chegam em julho, ficam por uma semana, reservam em média com 23 dias de antecedência e reservam em média em dois meses (57 dias).

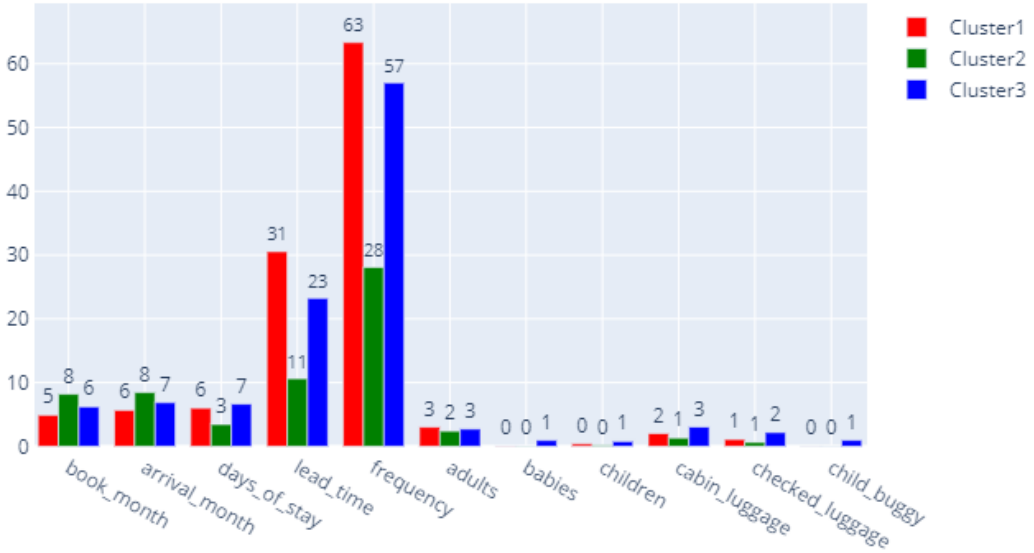


FIGURA 4.17. Segmentação de clientes que compram serviços de transfer, divididos em três *clusters*.

CAPíTULO 5

Desenvolvimento do Sistema de Recomendação

Neste capítulo é apresentado o desenvolvimento do sistema recomendação, que tem como objetivo recomendar as experiências vendidas pela YellowFish Adventures aos seus clientes de transfers. Através da modelo sugerido para o sistema, pretende-se utilizar dados de *clickstream* e dados das reservas para obter uma abordagem holística sobre a experiência e jornada do cliente de modo a recomendar experiências *tours* com base no histórico comportamental e demográfico dos clientes. Apresentamos ainda a metodologia utilizada para desenvolvimento dos modelos base e híbridos.

5.1. Algoritmos Aprendizagem Automática

De forma a construir os modelos foi utilizada a biblioteca de código aberto para Python, *scikit-learn* [**67**], e os seguintes algoritmos pertencentes à mesma: *K-nearest Neighbors Algorithm*, *Naive Bayes*, *Random Forests*, *Support Vector Machines*, *Gradient Boosting*, *XGBoost*, *Linear Discriminant Analysis*, *DecisionTrees*, *AdaBoost* e *Neural Networks*. Utilizou-se um *Emsemble* com classificação por votos que engloba todos os algoritmos designados anteriormente, sob forma destes deliberarem em conjunto sobre a classe de uma determinada observação a teste. No caso da maioria dos modelos classificadores, englobados no classificador por votos, estarem de acordo acerca da classe alvo de uma determinada observação a teste, esta mesma observação é classificada segundo classe alvo que a maioria decidiu.

O algoritmo *K-nearest Neighbors Algorithm* (KNN), ou k-vizinhos mais próximos, é um método não paramétrico utilizado para classificação e regressão. Em ambos os casos, a entrada consiste na função de semelhança e no parâmetro k. O algoritmo utiliza a função de semelhança e escolhe os k exemplos de treino mais próximos, no espaço das *features*, para votar na classe maioritária. Quando o algoritmo é utilizado para classificação, a saída é uma classe. Um exemplo é classificado através da classe mais comum entre os k-vizinhos mais próximos. No que diz respeito à utilização do algoritmo para regressão, a saída é o valor da propriedade para o exemplo. Este valor é a média dos valores para os k vizinhos mais próximos.

O classificador *Naive Bayes* (NB) é um modelo de *machine learning* probabilístico que é utilizado para tarefas de classificação. O fulcro/coração do algoritmo é baseado no teorema de Bayes e na assunção de independência de todos os acontecimentos (atributos), ou seja, a presença de uma *feature* particular, não afeta a outra. Através deste teorema, é possível encontrar a probabilidade de A acontecer, dado que B ocorreu. Neste caso, B é a evidência e A a hipótese. O algoritmo assume que as *features* são independentes.

Decision Trees (DT), ou árvores de decisão, são modelos que tiram partido de exemplos previamente treinados, para decidir sobre novos exemplos apresentados ao modelo. O modelo constrói uma árvore de decisão (ou de regras), sendo esta composta por nós e folhas. As folhas representam o uma classe (*target*) e os nós representam a conjugação das *features* que levam ao nome de cada classe. O modelo decide com base numa série de caminhos (sequência de regras sobre os valores dos atributos em cada nó) orientados da raiz para uma folha, a que corresponde sempre uma classificação. A relação entre os elementos da árvore (nós e folhas) e os atributos, valores e classificações pode ser entendida da seguinte forma: cada nó interno testa um atributo, cada ramo corresponde a um valor do atributo (que o nó testa) e cada folha atribui uma classificação.

Random Forests (RF), ou florestas aleatórias, é um algoritmo de aprendizagem supervisionada que combina várias árvores de decisão, treinadas, na maior parte dos casos, com o método de *bagging* de modo a melhorar o resultado geral dos modelos. Uma grande vantagem do algoritmo de florestas aleatórias é que pode ser utilizado tanto para tarefas de classificação quanto regressão, que representam a maior parte dos problemas de aprendizagem automática atuais.

Support Vector Machines (SVC), ou máquina de vetores de suporte, são um conjunto de métodos de aprendizagem supervisionados utilizados para classificação, regressão e detecção de *outliers*. O objetivo do algoritmo passa por encontrar um hiperplano num espaço N-dimensional (N = número de características) que classifica distintamente os diversos pontos dos dados das várias classes. O algoritmo é eficaz em espaços de alta dimensão e nos casos em que o número de dimensões é maior do que o número de amostras. É ainda um algoritmo eficiente em utilização de memória, dado que utiliza um subconjunto de treino na função de decisão (vetores de suporte). A função de decisão pode utilizar diferentes funções do *kernel* e ainda podem ser personalizados os núcleos, tornando-o assim um algoritmo versátil.

AdaBoost (AB) é um meta-estimador que começa por treinar um classificador no conjunto de dados original e, em seguida, encaixa cópias adicionais do classificador no mesmo conjunto de dados, mas onde os pesos das observações classificadas incorretamente são ajustadas de modo que os classificadores subsequentes se concentram nas observações mais dificeis de classificar. Por exemplo, o algoritmo começa por treinar uma árvore de decisão na qual a cada observação é atribuída um peso igual. Depois de avaliar a primeira árvore é aumentado os pesos dessas observações que são difíceis de classificar e reduzidos os pesos para aquelas que são mais fáceis de classificar. A segunda árvore é, portanto, treinada com esses dados ponderado, com o objetivo de melhorar as previsões da primeira árvore. O novo modelo é, portanto, a conjugação das duas árvores. É então calculado o erro de classificação deste novo modelo com a conjugação das duas árvores e treinada uma terceira árvore para prever as observações. Este processo é repetido segundo um número de iterações especificadas. As árvores subsequentes ajudam a classificar observações que

não tenham sido bem classificadas pelas árvores anteriores. As previsões do modelo final são a soma ponderada das previsões feitas pelos modelos das árvores anteriores.

Gradient Boosting (GB) é uma técnica de aprendizagem automática adaptada tanto para problemas de regressão como de classificação, que produz um modelo de previsão na forma de um *ensemble* de modelos de predição fracas, tipicamente árvores de decisão. O algoritmo treina vários modelos de forma gradual, aditiva e sequencial e tem um comportamento semelhante ao algoritmo AB, diferenciando-se na forma como identificam os modelos fracos. Enquanto o modelo AB identifica as fraquezas dos modelos utilizando os pontos de dados com um peso alto associado, o algoritmo GB executa o mesmo utilizando gradientes na função de perda. A função de perda é uma medida que indica o quão bons são os coeficientes do modelo para se encaixar nos dados subjacentes. Uma compreensão lógica da função de perda dependo do que está sendo otimizado pelo algoritmo. Por exemplo, se o objetivo é classificar possíveis cancros, a função de perda seria uma medida do quão bom o modelo preditivo é na classificação de se nosso objetivo é classificar a má classificação de tipos de cancro. O algoritmo permite otimizar uma função de custo especificada pelo utilizador, em vez de uma função de perda que geralmente oferece menos controle e por vezes não corresponde às aplicações de aprendizagem automática do mundo real.

XGBoost - Extreme Gradient Boosting (XGB) é uma técnica de aprendizagem automática *ensemble* que utiliza o algoritmo GB para fazer previsões. A sua rápida execução e escalabilidade são duas das vantagens deste algoritmo. No algoritmo XGB é modificado o algoritmo GB para que possa treinar com qualquer função de perda diferente.

Linear Discriminant Analysis (LDA) ou Análise Discriminante Linear, é uma técnica de redução de dimensionalidade que é comummente utilizada nos problemas de classificação supervisionados. É utilizada para modelar observações em grupos, ou seja, separar duas ou mais classes. É utilizada para projetar as características de um espaço de dimensão superior num espaço de dimensão inferior. Uma regra discriminante tenta dividir o espaço de dados em K regiões desarticuladas que representam todas as classes. Com essas regiões, a classificação por análise discriminante significa simplesmente que alocamos x para a classe j se x estiver na região j. O x pode ser alocado numa determinada região seguindo uma das seguintes regras: regra de probabilidade máxima ou regra de *bayes*.

Neural Networks (NN), ou redes neuronais, é uma família de algoritmos que se esforça para reconhecer as relações subjacentes num conjunto de dados através de um processo que tem por base o modo como os neurónios humanos comunicam. Nesse sentido, as redes neuronais referem-se a sistemas de "neurónios", orgânicos ou artificiais.

Pode ainda ser utilizada uma abordagem englobando o parecer de vários modelos para decidir se um exemplo deve ser classificado de certa forma, denominada de *Emsemble*. No caso de um classificador, são utilizados vários algoritmos e a classificação é decidida por votação. No nosso caso, usámos os algoritmos anteriormente descritos, para que, de uma forma democrática, decidam por maioria de voto, a classe a que deve pertencer um

exemplo do conjunto de teste. O sistema de votos é composto por um método *hard* e *soft*. O método *hard* utiliza as predições das classes para a regra de votação maioritária. No que diz respeito à votação *soft*, a saída para a predição da classe é baseada no argumento máximo da soma das probabilidades preditivas, sendo este método recomendado para combinação de classificadores bem calibrados. É ainda possível a atribuição de pesos para equilibrar a ocorrência de classes preditivas (*hard*) ou probabilidades das classes antes da média (*soft*).

5.2. Preparação, Transformação e Amostragem dos Dados

O desafio neste projeto passa por construir um sistema de recomendação com dados extremamente desequilibrados e onde as classes dos produtos a recomendar apenas representam cerca de 0,20% do número total de observações, perfazendo um rácio de 1:121 em relação aos serviços de transfer e experiências vendidas. Como verificado na Tabela 1 a classe 0, que representa as observações que apenas compraram um serviço de transfer, é significativamente maior em número de observações que as restantes.

TABELA 1. Número de observações no conjunto de dados por tipo de serviço comprado.

Serviço	Classe	Total
Só transfer	0	273 341
Golfe	1	172
Feel Tour	2	162
Experience Tour	3	93
Feel Tour + Caves & Coastline	4	15

Como espectável, ao utilizar os dados com esta discrepância no número de observações para cada classe, foram obtidos resultados fracos que demonstram que os modelos aprendem muito bem a classe zero (apenas compra de transfer) e não são bons classificadores de observações com experiências vendidas.

Dado que a experiência *Feel Tour + Calves & Coastline* apenas contém 15 registos, foi decidido eliminar estes registos do conjunto de dados e não os considerar.

De forma a resolver este problema de equilíbrio entre classes e os modelos treinados não recaírem no problema de *overfitting* para a classe 0 e *underfitting* para as restantes classes, foram utilizadas diferentes técnicas de amostragem, nomeadamente quando da limpeza dos dados e para controlo de quantidade de observações. Foi utilizada a biblioteca para Python *imbalanced-learn* **[68]** **[69]** que oferece um vasto número de técnicas de *undersampling* e *oversampling*, das quais usamos algumas para serem aplicadas nos dados e que serão descritas no que se segue.

No que diz respeito aos algoritmos para efetuar *undersampling*, que objetivam reduzir o número de observações da classe maioritária, o número de observações pretendido por classe tem que ser previamente especificado. Para isso, foram utilizados os seguintes

algorítmos da biblioteca *imbalanced-learn*: *Random Undersampling*, *Near Miss*, *Cluster Centroids*, *TomekLinks*, *NeighbourhoodCleaningRule* e *InstanceHardnessThreshold*.

O algoritmo *Random Undersampling* consiste na escolha aleatória e uniforme de observações para a classe maioritária. Este método pode potencializar perda de informação, contudo, se as observações da classe maioritária estiverem cingidas ao mesmo espaço, podem ser obtidos bons resultados com este método. De forma a mitigar o problema da perda potencial de informação inerente à técnica *Random Undersampling*, também foram aplicados métodos de *undersampling* utilizando a vizinhança de observações, havendo diversas variações nos métodos.

O método *Near Miss* baseia-se nas regras heurísticas de algoritmos de vizinhos mais próximos para seleção de observações, calculando a distância para os vizinhos [**70**]. Em primeiro lugar o método calcula as distâncias entre todas as observações da classe maioritária e as observações da classe minoritária. Em seguida, seleciona as K observações da classe maioritária que têm as menores distâncias para as observações da classe minoritária. Se houver N instâncias na classe minoritária, o método resultará em K*N observações da classe maioritária. A variação *NearMiss-1* seleciona as observações da classe maioritária onde as distâncias médias a três instâncias mais próximas da classe minoritária são as menores. *NearMiss-2* utiliza as três amostras mais distantes da classe minoritária. *NearMiss-3* utiliza dois passos para seleccionar as observações. Em primeiro lugar, são selecionados os vizinhos mais próximos de cada observação da classe minoritária. No segundo passo, são selecionadas as observações da classe maioritária onde a distância média para os N vizinhos mais próximos é maior.

Cluster Centroids realiza *undersampling*, ou seja redução das observações da classe maioritária através da geração de centróides com base em métodos de agrupamento [**71**]. A classe maioritária é reduzida através da substituição das amostras maioritárias pelo centróide do *cluster* do algoritmo *K-means*. Este algoritmo mantém N amostras maioritárias, utilizando o algoritmo K-means com N *clusters* para a classe maioritária e aplicando as coordenadas dos centróides dos N *clusters* como as novas amostras maioritárias. Exemplo: se a classe maioritária tem 1000 amostras e a classe minoritária 100, neste caso podem ser formados 100 *clusters* da classe maioritária e substituir 1000 pontos por 100 pontos de dados, ou seja, pelo centróide de cada *cluster*.

A *Edited Nearest Neighbor Rule* (ENN) pretende remover qualquer observação cuja classe é diferente da classe de pelo menos dois dos seus três vizinhos mais próximos [**72**]. A técnica remove todas as instâncias que estejam próximas ou ao redor das diferentes classes, com base no conceito de vizinho mais próximo. O método devolve probabilidades e nem sempre é possível obter um número específico de amostras.

TomekLinks é uma técnica de limpeza dos dados que não permite a especificação do número de observações que se pretende em cada classe [**73**]. Sejam dadas duas observações, A e B, pertencentes a classes diferentes e que estão separados por uma distância, D(A,B). O par (A, B) é denominado de *TomekLink* se não houver uma observação C tal que

D(A,C) < D(A,B) ou D(B,C) < D(A,B). As observações no espaço entre os *TomekLinks* são consideradas ruído e removidas. O método devolve probabilidades, não sendo sempre possível obter um número específico de amostras.

NeighbourhoodCleaningRule baseia-se na regra de limpeza da vizinhança [**74**]. O método utiliza *Edited Nearest Neighbor Rule* (ENN) e K-means para remover amostras que apenas trazem ruído ao conjunto de dados. Para cada instância no conjunto de dados, o algoritmo encontra os três vizinhos mais próximos. Se a instância pertence à classe maioritária e a classificação dada pelos seus três vizinhos mais próximos é o oposto da classe da instância escolhida, então a instância em questão é removida. Se a instância escolhida pertence à classe minoritária e é mal classificada pelos seus três vizinhos mais próximos, então os vizinhos mais próximos que pertencem à classe maioritários são removidos. De forma semelhante ao método anterior, também este método devolve probabilidades e nem sempre é possível obter um número específico de amostras.

InstanceHardnessThreshold é um algoritmo específico no qual um classificador com indicação de probabilidade é treinado sobre os dados e as amostras com probabilidades mais baixas são removidas [**75**]. Por defeito é utilizado o classificador *Random Forests*. Este método poderia ser considerado um método de *undersampling* controlado. No entanto, dado que o método devolve probabilidades, nem sempre é possível obter um número específico de amostras.

As técnicas de *oversampling* consistem na geração de novas observações de dados semelhantes à classe minoritária que queremos obter mais registos. Foram utilizadas as seguintes técnicas: *RandomOverSampler*, SMOTE (*Synthetic Minority Over-sampling Technique*), *BorderlineSMOTE*,*SMOTE-NC*, *SVMSMOTE*, *SMOTEENN*, *SMOTETomek*, ADASYN (*Adaptive Synthetic*).

RandomOverSampler simplesmente replica aleatoriamente as observações da classe minoritária. As principais desvantagens desta técnica prende-se com o facto de não haver controlo da replicação dos dados e com isto haver dados sem significado a serem replicados, aumentando assim a probabilidade dos modelos caírem no problema de *overfitting* e causar fraca generalização no conjunto de teste.

De forma a evitar o problema de *overfitting* da técnica *RandomOverSampler* foi proposto a técnica *Synthetic Minority Over-sampling Technique* (SMOTE), sendo esta considerada estado-da-arte e funcionando em diversos cenários de aumento de observações. O método gera dados sintéticos com base nas semelhanças do espaço de *features* entre as observações das classes minoritárias existentes. A fim de criar uma observação sintética, o método encontra os K vizinhos mais próximos de cada observação minoritária, seleciona aleatoriamente uma deles e, em seguida, calcula interpolações lineares (podendo ser apenas utilizada com dados numéricos) para produzir uma nova observação minoritária.

BorderlineSMOTE é uma variante do algoritmo SMOTE original em que as amostras no limite de decisão são detectadas e utilizadas para gerar novas amostras sintéticas.

Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC) é uma variação do algoritmo SMOTE que pode ser utilizada em conjunto de dados contendo variáveis contínuas e categóricas.

SVMSMOTE é uma variante do algoritmo SMOTE que utiliza o algoritmo SVM para gerar novas amostras sintéticas. O algoritmo SVM é utilizado para localizar o limite de decisão definido pelos vetores de suporte e as observações das classes minoritárias que se aproximam dos vetores de suporte tornam-se o foco para gerar exemplos sintéticos. Se o número total de observações da classe maioritária for maior do que os seus vizinhos mais próximos da classe minoritária, novas observações são criadas com extrapolação para expandir a área da classe minoritária em direção à classe maioritária.

A técnica *SMOTEENN* utiliza o algoritmo base SMOTE para realizar *oversampling* e *undersampling,* ao mesmo tempo que faz limpeza nos dados através da técnica ENN.

A técnica *SMOTETomek* utilizado o algoritmo base SMOTE para realizar *oversampling* e *undersampling,* ao mesmo tempo que faz limpeza nos dados através da técnica *TomekLinks*.

A técnica ADASYN (*Adaptive Synthetic*) gera amostras da classe minoritária de acordo com suas distribuições de densidade. São gerados mais dados sintéticos para as amostras das classes minoritárias que são mais difíceis de aprender, em comparação com aquelas amostras minoritárias que são mais fáceis de aprender. O algoritmo calcula os K vizinhos mais próximos de cada observação minoritária, em seguida, obtém o rácio das observações das classes maioritárias e minoritárias para gerar novas amostras. Ao repetir este processo, muda adaptativamente o limite de decisão para focar nas amostras difíceis de aprender.

5.3. Modelo de Classificação Base

De forma a se poder antecipar as expectativas dos clientes e identificar o tipo de experiência mais adequada a cada perfil de cliente, foi utilizado aprendizagem automática para construir modelos que sejam capazes de prever se um determinado cliente, a comprar um serviço de transfer, vai-se interessar por comprar experiências e de que tipo.

No decorrer desta investigação empírica para classificação da intenção de compra de serviços de experiência e respetivo tipo, dado uma determinada compra de transfer, foram utilizadas várias abordagens. Em primeiro lugar foram desenvolvidos modelos utilizando todas as classes disponíveis, observações de compra apenas de transfer (classe 0) e classes de experiências (classes 1,2,3) de uma forma geral, sem distinção de veículos a utilizar, Visto que o número de exemplos para cada classe utilizando a distinção por veículos seria muito baixa, foi decidido não considerar esta abordagem nos modelos.

Foram treinados modelos de classificação incluindo todas as classes ao mesmo tempo, com o intuito de prever se um cliente que comprou um transfer também compraria uma experiência.

5.4. Modelo de Classificação Híbrida

De forma a otimizar os resultados obtidos com os modelos base, foi construído um modelo híbrido de classificação baseado em dois submodelos que visam melhorar o desempenho das classificações de intenção de compra de experiências.

Nesta abordagem híbrida, inicialmente um modelo binário classifica se uma reserva de transfer irá ou não comprar uma experiência. Todas as observações que o modelo binário corretamente identifica como potenciais compradores de experiências são transmitidos a um segundo modelo, multi-classe, responsável por classificar apenas entre as diferentes classes de experiências (classe 1,2,3) qual delas será a mais indicada.

5.4.1. Modelo de Classificação Binária

A classificação binária denomina-se pela classificação de duas classes, normalmente entre a classe zero e um, que representam neste caso a compra de apenas um serviço de transfer (classe 0), e a compra de um serviço de transfer e um serviço de experiência (classe 1).

Para o desenvolvimento deste modelo binário, todos as observações com classes diferentes de zero foram convertidas para classe 1, indicando que são uma venda de transfer que comprou experiência. As restantes observações de apenas compra de serviço de transfer (classe 0) mantêm-se sem alterações.

Dado a discrepância entre o número total de observações de cada classe (0, 1), foi necessário equilibrar os dados utilizando técnicas de *undersampling*, selecionando um número de amostras de classe 0 semelhante ao número de amostras da classe 1. Contudo, foi tida em consideração que existem muitas mais compras de transfers do que experiências, sendo esta informação foi passada implicitamente ao modelo através de uma seleção de amostras de classe 0 ligeiramente superiores aos da classe 1.

A utilização de um modelo binário é essencial para identificar quais os padrões que podem ser utilizados para prever quais os clientes que podem potencialmente comprar serviços adicionais. O objetivo principal é construir um modelo que possa prever se um cliente comprará uma experiência. Os modelos foram ajustados para maximizar as taxas de verdadeiros positivos e a taxa de verdadeiros negativos.

Os verdadeiros positivos que o modelo binário corretamente identificou são enviados para um segundo modelo, multi-classe, responsável por classificar entre as várias experiências qual delas se assemelha ao perfil do cliente em questão e recomendá-la.

5.4.2. Modelo de Classificação Multi-Classe

O modelo de classificação multi-classe é responsável pela classificação da experiência adequada ao perfil de um cliente. Com a classificação multi-classe pretende-se treinar modelos que consigam identificar e classificar apenas os tipos de experiências disponibilizadas para venda, sendo elas: golfe, *Feel tours* e *Feel Tour and Caves Costline*. Ao receber um conjunto de amostras que o modelo binário corretamente identificou como compradores de experiências, o modelo multi-classe classifica cada uma destas amostras consoante o tipo

de experiências. Desta forma é possível ter apenas um classificador focado na aprendizagem dos padrões dos clientes que compram cada tipo de experiência.

Em contexto real, fora do ambiente de treino, o modelo binário enviará todos os casos em que identifica como positivo e o modelo multi-classe encarregar-se-á de classificar o tipo de experiência mais adequada ao perfil do cliente.

5.4.3. Modelo do Sistema de Recomendação

O modelo do sistema foi desenhado para que seja possível integrar várias fontes de dados que caracterizam o comportamento do cliente e consiga perceber se um cliente que esteja a navegar na página web da YellowFish Transfers e comprador de um ou mais serviços de transfer, pretende adquirir uma experiência também. Para isso, o modelo do sistema contempla dados sobre a jornada do cliente sobre a navegação no website, provenientes do Google Analytics, onde estão registadas todos os eventos de interação, assim como, as características intrínsecas de cada cliente online. Para além da utilização de fontes de dados online, o modelo do sistema considera os dados das reservas dos transfers, registados no manifesto.

De um ponto de vista teórico, a conjugação dos dois conjuntos de dados, *clickstream* e manifesto, que caracterizam a jornada completa do cliente, seria utilizada para treinar um modelo híbrido baseado em duas principais etapas. De um ponto de vista empírico, nesta investigação, apenas foram utilizados dados do manifesto para recomendação de experiências.Dado que não é possível fazer a correspondência de dados encontradas entre o manifesto e os dados de *clickstream* provenientes do Google Analytics, nem os dados têm qualidade para se poder treinar os modelos utilizando as informações de *clickstream*, foram apenas utilizados dados das bases de dados históricas nos modelos.

Analisando a Fig. 5.1 que ilustra o modelo proposto para o sistema de recomendação como um todo, onde o principal objetivo é sugerir uma experiência aos clientes que visitam o site da YellowFish, o modelo está dividido da seguinte forma: o visitante da página web da YellowFish Transfers, ao adquirir um transfer passa a ser um cliente, tendo a sua jornada registada no Google Analytics; após adquirir um transfer essa informação é inserida na base de dados de transfers; sob forma dos modelos serem treinados, a informação do Google Analytics é extraída através da REST API e a informação sobre os transfers é extraída para o ficheiro manifesto; a combinação destas duas fontes de dados representam a jornada do cliente para a compra de um transfer e experiências. Seguidamente o modelo binário e modelo multi-classe são treinados e disponibilizados como *web service* a ser aplicado para recomendar experiências aos próximos clientes de transfers online.

Na primeira etapa, um modelo binário é treinado para prever se um cliente que comprou serviço(s) de transfer irá comprar ou não uma experiência. Na segunda etapa, treinamos um modelo multi-classe com apenas observações de serviços de transfers com compra de experiências associadas, ou seja, sem considerar a classe zero. Com este modelo, pretendemos classificar qual das experiências disponíveis um cliente de transfer identificado como comprador de experiências vai comprar. Neste segundo modelo, as observações que

foram classificadas como verdadeiros positivos (modelo previu que cliente de transfer vai comprar uma experiência e realmente comprou), são enviadas para este segundo modelo para classificar o tipo de experiência que o cliente vai comprar.

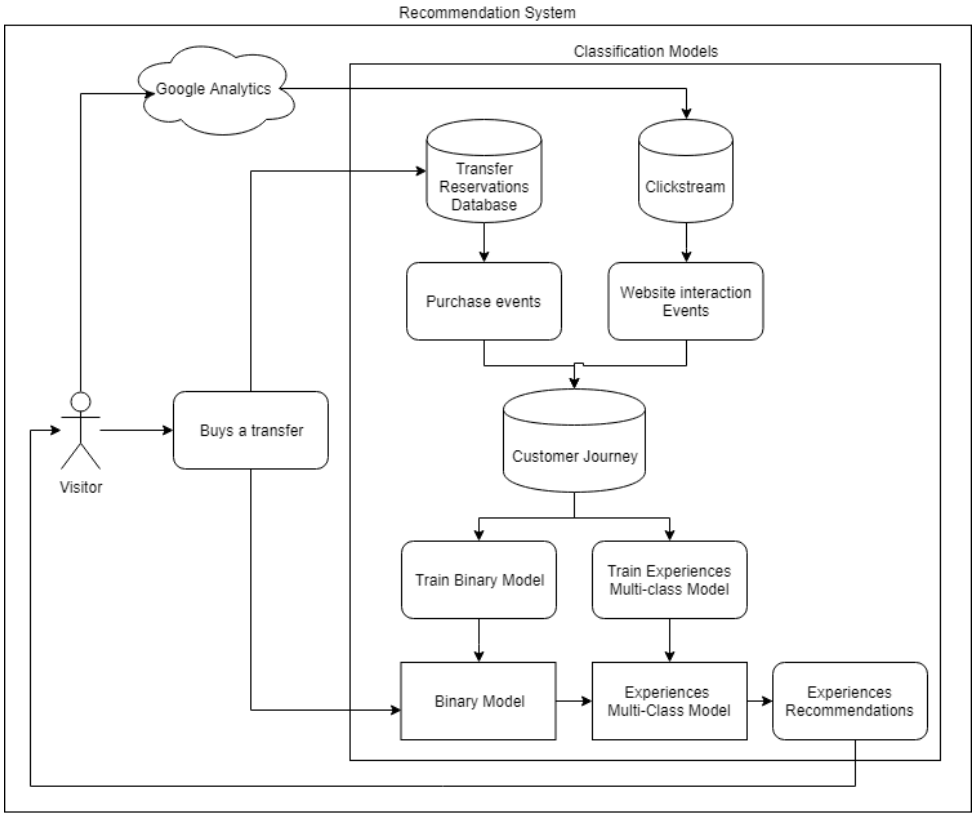


FIGURA 5.1. Modelo do sistema de recomendação híbrido.

5.5. Treino e Teste dos Modelos

O conjunto de dados foi dividido em 70% para treino e 30% para teste como verificado na Tabela 2. Dado que a divisão em conjunto de treino e conjunto de teste originou uma maior discrepância entre o número de observações para cada classe, foram aplicadas as técnicas de *undersampling* e *oversampling* para reduzir e/ou aumentar as observações das classes. As técnicas de *undersampling* e *oversampling* apenas devem ser aplicadas ao conjunto de treino para que não exista influência nas observações a teste, caso contrário, os modelos podem memorizar as observações sintéticas criadas e os resultados não serem fiáveis em dados desconhecidos e em contexto real. Desta forma, o aumento e redução das observações das classes minoritárias e maioritárias, através das técnicas de *undersampling* e *oversampling*, apenas foram aplicadas ao conjunto de treino.

Sob forma de ser obtido um maior equilíbrio dos dados para cada uma das classes, os modelos foram treinados com diferentes tamanhos de exemplos de treino para a classe 0, assim como, para as classes de experiências (classes 1,2,3). Utilizando as diferentes abordagens de *undersampling* e *oversampling* foi possível diminuir e/ou aumentar as observações consoante o tipo de teste para cada modelo. Dado que temos disponível uma

Tabela 2. Número de observações no conjunto de dados por tipo de serviço comprado.

Serviço	Classe	Total	Treino	Teste
Só transfer	0	273341	191340	82003
Golfe	1	172	120	52
Feel Tour	2	162	114	48
Experience Tour	3	93	65	28
Feel Tour + Caves & Coastline	4	15	10	5

grande quantidade de observações para a classe 0, foram utilizados diferentes tamanhos alvo para esta classe, aplicando apenas técnicas de *undersampling* para as reduzir, sob forma de analisarmos o impacto deste diferente número de observações no desempenho dos modelos. Para as classes de experiências (1,2,3) apenas foram utilizadas técnicas de *oversampling* dado que estas são as classes minoritárias, tendo um número de observações semelhantes entre cada uma destas. Todas as técnicas de *undersampling* e *oversampling* forma utilizadas nos vários testes levados a cabo, sob forma de ser analisado o impacto das mesmas no desempenho dos modelos.

Para cada iteração de treino/teste foi possível aplicar diferentes tamanhos de amostras da classe 0, diferentes técnicas de *undersampling* ou *oversampling*, e diferentes algoritmos, identificando no final qual a melhor combinação que se adequava ao problema. Em cada ciclo de treino/teste é definido um número de amostras da classe 0 a serem utilizadas no conjunto de treino e teste dos modelos, seguidamente, num ciclo seguinte é definida a técnica de *undersampling* ou *oversampling* a ser aplicada nas classes do conjunto de treino. Posteriormente é dividido em conjunto de treino e teste, onde o conjunto de dados divide-se em 70% para treino e 30% para teste, e é aplicada a técnica de *undersampling* ou *oversampling* no conjunto de treino. Neste processo é também criado um conjunto de treino/teste binário a ser utilizado nos modelos binários pertencente aos modelos híbridos, em que as classes 1,2,3 passam a ser apenas a classe 1 e a classe 0 mantém-se sem alterações. De forma a treinar os modelos, são inicializados vários modelos utilizando os algoritmos apresentados anteriormente, inicializando-os com os parâmetros por defeito. Logo de seguida, os modelos são treinados utilizando o procedimento heurístico de hiperparâmetros *GridSearchCV* da biblioteca *scikit-learn* [**67**], para identificar os melhores parâmetros para os diferentes algoritmos de forma a melhorar o desempenho de cada um deles. O ajuste foi realizado utilizando cinco validações cruzadas para otimização dos hiperparâmetros de forma a encontrar os melhores parâmetros e estimar o desempenho dos modelos em dados desconhecidos. Ao determinar os melhores parâmetros encontrados para cada algoritmo, é construído e inicializado para cada algoritmo, um modelo com esses melhores parâmetros, sendo este submetido a teste, classificando cada exemplo no conjunto de teste. É então construída uma tabela com os resultados obtidos para cada modelo, tendo em conta as várias métricas utilizadas. Para os modelos baseados em algoritmos de árvores são construídos gráficos com a importância de cada característica

no conjunto de dados que ajudam os modelos a tomarem corretas decisões. São ainda construídos os gráficos com a curva de ROC, curva de precisão e *recall*, estatística de *Kolmogorov–Smirnov*, curva de ganho cumulativo, curva de elevação e matriz de confusão. São ainda gravados os melhores modelos treinados, de forma a poderem ser utilizados posteriormente e efetuar previsões em dados novos.

No próximo capítulo são apresentados os resultados obtidos com o modelo de classificação híbrido e comparados os resultados das diferentes abordagens utilizadas.

CAPíTULO 6

Resultados e Discussão

Neste capítulo são apresentados os resultados obtidos com o modelo híbrido e efetuada a comparação com os modelos base. São ainda demonstrados os impactos das abordagens de *undersampling* e *oversampling* no desempenho dos modelos.

6.1. Métricas de Avaliação

A utilização de diferentes métricas para avaliação dos modelos permitem estimar os resultados atuais corretamente e melhorá-los. Contudo, uma compreensão errada das métricas leva à estimativa errada da capacidade dos modelos e uma visão diferente do estado do problema. Foram utilizadas diferentes métricas para avaliar tanto os modelos base como os modelos híbridos, tendo em conta a desproporção de exemplos nas diferentes classes a teste. Foi utilizada a biblioteca *scikit-learn* para implementação das seguintes métricas: taxa de verdadeiros positivos, taxa de verdadeiros negativos, precisão, *F1-score*, *Matthews correlation coefficient*, *area under receiver operating characteristic curve* e *geometric mean*.

A avaliação principal dos modelos recai sobre um problema binário, sendo este a classificação de uma observação de venda só transfer ou de uma observação de venda de transfer com experiência associada. Dado que as observações em contexto de teste também estão desequilibradas, não tendo sido aplicada qualquer técnica de *undersampling* e *oversampling*, foram aplicadas as métricas que melhor avaliam o desempenho em contexto de modelos testados com dados desequilibrados.

A métrica mais simples para medir a classificação é a *accuracy* que avalia a razão de previsões corretas para o número total de amostras no conjunto de dados. No entanto, no caso de classes desequilibradas, a métrica pode não refletir corretamente os resultados, podendo fornecer valores elevados que não demonstram a verdadeira capacidade de previsão para a classe minoritária. A métrica pode apresentar 99% de *accuracy*, mas ainda assim ser ter uma péssima capacidade de previsão nas classes minoritárias, como é o caso, por exemplo, da deteção de anomalias.

De forma a ser obtida melhor compreensão sobre a capacidade preditiva dos modelos, são identificados os quatro casos possíveis para as observações aquando da sua classificação:

- Verdadeiro positivo (VP)— o rótulo da amostra é positivo e é classificado como positivo.
- Verdadeiro negativo (VN) — o rótulo da amostra é negativo e é classificado como negativo.

- Falso positivo (FP)— o rótulo da amostra é negativo, mas é classificado como positivo.
- Falso negativo (FN)— o rótulo da amostra é positivo, mas é classificado como negativo.

Com base nestes quatro valores, podem ser derivadas as seguintes métricas paras as classes desequilibradas:

Taxa de Verdadeiros Positivos (TVP)/Sensibilidade/*Recall* identifica quantas observações relevantes são classificadas corretamente através da quantificação do número de verdadeiros positivos feitas a partir de todas as previsões positivas que poderiam ter sido feitas. A métrica fornece uma indicação da quantidade de observações verdadeiras positivas que falharam em ser classificadas, fornecendo assim uma noção da cobertura da classe positiva pelos modelos. É calculada através do número total de verdadeiros positivos, divididas pelo somatório do número total de verdadeiros positivos e falsos negativos. O resultado da métrica varia entre 0 e 1, em que o valor 0 indica um mau resultado, e 1, um resultado perfeito para TVP/*recall*. A métrica é apropriada num cenário em que se pretende minimizar os falsos negativos.

$$TVP = \frac{VP}{VP + FN}$$

Taxa de Verdadeiros Negativos (TVN)/Especificidade é calculada através do número de previsões negativas corretamente identificadas divididas pelo número total de verdadeiros negativos e falsos positivos. A métrica varia entre 0 e 1, sendo o valor 0 o pior resultado e 1 o melhor. Por exemplo, um teste que identifica todas as pessoas saudáveis como sendo negativas para uma determinada doença é muito específico. Outro teste que identifica incorretamente 30 % das pessoas saudáveis como tendo a doença seria considerado menos específico, tendo uma taxa de falsos positivos (TFP) mais elevada.

$$TVN = \frac{VN}{VN + FP}$$

Precisão indica quantas observações classificadas são relevantes. A métrica foca-se mais na classe positiva do que na classe negativa, medindo a probabilidade de detecção correta das classes positivas. A métrica varia entre 0 e 1, sendo o valor 0 o pior resultado e 1 o melhor e é apropriada no cenário em que se presente minimizar os falsos positivos.

$$Precisao = \frac{VP}{VP + FP}$$

F1-Score, ou métrica F1, é a média harmónica entre as métricas precisão e TVP/*recall*, sendo uma escolha adequada para o cenário de classificação com um número desequilibrado de observações das classes. Por vezes os modelos podem obter uma excelente precisão, mas uma TVP/*recall* muito baixa, ou alternadamente, a uma precisão muito baixa, mas um excelente TVP/*recall*. A métrica F1 expressa ambas as preocupações numa única métrica. A métrica F1 varia entre 0 e 1, onde 1 é classificação perfeita e 0 é falha total.

$$F1 = 2 * \frac{VP}{VP + FP + FN}$$

Matthews correlation coefficient **(MCC)**, ou coeficiente de correlação de *Matthews*, é uma métrica que considera os verdadeiros positivos, falsos positivos e verdadeiros negativos, sendo uma medida equilibrada que é adequada para medir o desempenho dos testes quando existe classes de tamanhos muito diferentes. A métrica varia entre -1 e 1, onde 1 é pontuação para uma previsão perfeita, 0 é igual a previsões aleatórias pelo classificador e -1 indica discordância total entre a previsão e o verdadeiro rotulo da classe.

$$MCC = \frac{VP * VN - FP * FN}{\sqrt{(VP + FP) * (VP + FN) * (VN + FP) * (VN + FN)}}$$

Receiver operating characteristic curve - area under the curve **(ROC-AUC)**, ou área sob a curva de ROC, mede a capacidade de um classificador distinguir entre classes. Quanto maior o valor de AUC, melhor o desempenho do modelo na distinção entre as classes positivas e negativas. Quando o AUC é igual a 1, o classificador é capaz de distinguir perfeitamente entre todos os pontos positivos e negativos corretamente. Se, no entanto, o AUC for de 0, então o classificador estará a prever todas as observações negativas como positivas, e todos as positivas como negativas. Quando o valor de AUC encontra-se entre 0.5 e 1, há uma grande probabilidade do classificador ser capaz de distinguir entre os valores positivos e os valores negativos das classes. Isso ocorre porque o classificador é capaz de detectar um maior número de verdadeiros positivos e verdadeiros negativos em vez de falsos negativos e falsos positivos. Quando o valor de AUC = 0,5, o classificador não é capaz de distinguir entre as observações das classes positivas e negativas, ou seja, ou o classificador está a prever as classes aleatoriamente ou indica a mesma classe constantemente para todos as observações das classes. Assim, quanto maior o valor de AUC para um classificador, melhor a sua capacidade de distinguir entre as classes positivas e negativas.

$$AUC = \frac{TVP + TVN}{2}$$

Geometric Mean **(G-mean)**, ou média geométrica, é a raiz do produto da TVP/ sensibilidade com a TVN/especificidade da classe. Esta medida tenta maximizar a precisão em cada uma das classes, mantendo-as equilibradas.

$$G - mean = \sqrt{(TVP * TVN)}$$

Dado que o principal objetivo dos modelos desenvolvidos será conseguir identificar quando é que um comprador de serviço de transfer poderá comprar também uma experiência, a avaliação dos modelos foi realizada de uma perspetiva binária. No caso, os verdadeiros negativos são as observações em que o modelo previu que seria apenas comprado um serviço de transfer e na realidade foi, os verdadeiros positivos são as observação em que o modelo previu que seria também comprado um serviço de experiência e foi, os

falsos positivos são as observações em que o modelo previu que seria comprado um serviço de experiência mas na realidade não foi, os falsos negativos são as observações em que o modelo previu que não seria comprado um serviço de experiência mas na realidade foi.

Mesmo nos modelos multi-classe, agrupou-se os verdadeiros positivos como as observações corretamente identificadas para as classes de experiências (1,2,3), os falsos positivos são as observações em que o modelo previu que era uma das classes de experiências mas na realidade era da classe 0, e os falsos negativos são as observações em que o modelo previu que seria da classe 0, mas na realidade era um exemplo das classes de experiências. Isto deve-se ao facto de ser mais importante saber a intenção do cliente em comprar uma experiência, do que classificar exatamente qual das três experiências cliente irá comprar.

Do ponto de vista do utilizador, o número de falsos positivos é mais importante que os falsos negativos devido à quantidade de recomendações que podem ser feitas ao cliente aquando da sua navegação no website, ao finalizar a compra de um transfer. Se os modelos resultarem num grande número de falsos positivos estaremos a aumentar o ruído visual dos clientes com *cross-selling*, o que pode ser contraproducente e levar os clientes a desistir da compra do serviço de transfer e sair do website. Por outro lado, da perspetiva da empresa, a importância dos falsos negativos reflete-se no facto de haver oportunidades que foram perdidas pelo modelo de sugerir experiências em que havia muita probabilidade do cliente comprar, sendo que a empresa pretende maximizar a probabilidade de *cross-selling* e vender serviços de experiência.

Dada a importância de ambos os casos, do ponto de vista do cliente e da empresa, tendo em conta os falsos positivos e os falsos negativos, os resultados são ordenados pela métrica, média geométrica, que representa o melhor indicador para o objetivo do nosso problema: reduzir os falsos positivos e falsos negativos, aumentando por consequência os verdadeiros positivos e negativos.

6.2. Resultados dos Modelos Base

Os resultados dos modelos base apresentam o desempenho dos modelos que utilizam apenas uma fase para classificação de observações de transfers e experiências. Nestes modelos são treinadas e testadas todas as observações das classes ao mesmo tempo.

Verifica-se que o equilíbrio do número de observações da classe 0 para valores semelhantes ou ligeiramente superiores ao número de observações das classe 1,2,3, resultam em melhores resultados no desempenho dos modelos como podemos verificar na Fig. 6.1. Contudo, os modelos utilizando DT e LDA tendem a melhorar com 1000 e 3000 exemplos de treino para a classe 0, respetivamente.

Os algoritmos XGB e GB destacaram-se como os que melhor se adaptam ao problema, resultando em melhor desempenho na classificação de verdadeiros positivos (compra de transfer e experiência) e verdadeiros negativos (apenas compra de transfer). Utilizando diferentes tamanhos de *undersampling* para classe 0 foi possível analisar e alcançar melhor desempenho para os diferentes algoritmos aplicados. A Tabela 1 apresenta os resultados

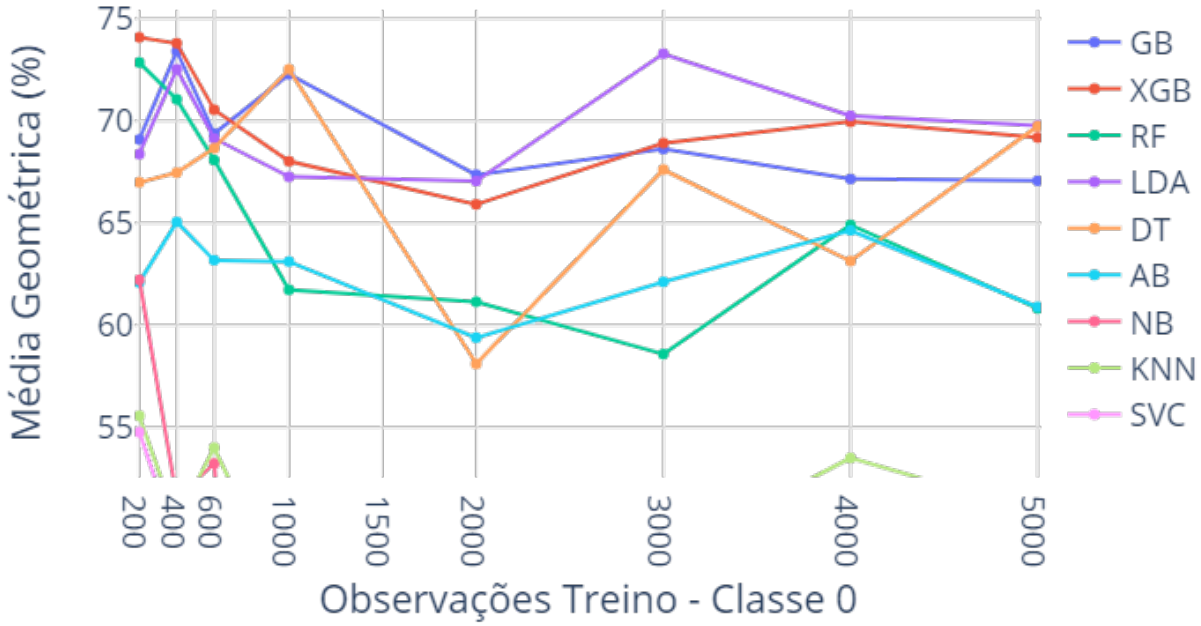


FIGURA 6.1. Comparação entre modelos com diferentes tamanhos para o treino da classe 0 utilizando *randomundersampling*.

dos dez melhores modelos base construídos. O algoritmo XGB obteve o melhor desempenho utilizando 200 e 400 observações no conjunto de treino para a classe 0. Com 200 observações para a classe 0, o modelo identificou corretamente 80 observações de compra de experiências num total de 128 observações das classes de experiência a teste. A segunda melhor abordagem foi obtida novamente com o algoritmo XGB, utilizando 400 observações para a classe 0, melhorando a taxa de verdadeiros negativos (corretamente identificadas apenas observações de transfers), mas foi reduzida a taxa de verdadeiros positivos (classe 1, 2 e 3) com 71 observações corretamente identificadas. As métricas de F1 e MCC não são apresentados nas tabelas porque os seus valores são demasiado baixos, o que poderia significar que os modelos distinguem muito mal entre amostras positivas e negativas, no entanto estes resultados baixos devem-se ao facto de existir um grande número de observações a teste da classe 0, o que desproporciona os resultados destas métricas.

O modelo com melhor desempenho tendo em conta a métrica GM, apesar de ter uma TVP mais alta, acertando num maior número de verdadeiros positivos (80), obteve uma TVN mais baixa quando comparada com os modelos da Tabela 1, sendo um modelo mais adequado do ponto de vista em que pretende maximizar a probabilidade da venda de experiências, apesar do risco ser maior com a apresentação/sugestão desnecessária de outros serviços aos clientes que não estão interessados (maior número de FP). Contudo, o segundo melhor modelo do ponto de vista da métrica GM, obteve uma TVN de 95,87% o que minimiza o risco de falsos positivos e desta forma não se torna cansativo para os clientes com abordagens de *cross-selling* desnecessárias, apesar da TVP ser menor com 58,60% e 71 verdadeiros positivos. Mais ainda, os modelos LDA treinados com 3000 e 10000 observações da classe 0 maximizam ainda mais TVN, obtendo um total de 98,71%

Tabela 1. Melhores resultados para os modelos base utilizando apenas *randomundersampling*.

C0	Modelo	VP	VN	FP	FN	TVP	TVN	GM
200	XGB	**80**	70319	11684	**45**	**64,00%**	85,75%	**74,08%**
400	XGB	71	78615	3388	54	56,80%	95,87%	73,79%
400	GB	76	70936	11067	46	62,30%	86,50%	73,41%
3000	LDA	68	80947	1056	57	54,40%	98,71%	73,28%
200	RF	72	76764	5239	55	56,69%	93,61%	72,85%
1000	DT	74	70514	11489	47	61,16%	85,99%	72,52%
400	LDA	71	73492	8511	50	58,68%	89,62%	72,52%
1000	GB	67	76070	5933	52	56,30%	92,76%	72,27%
10000	LDA	66	**80965**	**1038**	61	51,97%	**98,73%**	71,63%
400	RF	64	80845	1158	61	51,20%	98,59%	71,05%

e 98,73% com 1056 FP e 1038 FP, respetivamente. Apesar do número de verdadeiros positivos destes modelos LDA ser menor, com 68 e 66 observações de experiências corretamente identificadas, são adequados para o cenário em que não se pretende cansar os clientes com informações de outros serviços e que poderá levar à desistência da compra do serviço de transfer e saída do website.

A utilização de diferentes abordagens de *undersampling* e *oversampling* permitiu que os modelos conseguissem generalizar melhor, aprendendo as observações significativas para que conseguissem classificar com sucesso futuras observações desconhecidas. A Fig. 6.2 apresenta os melhores resultados para cada algoritmo e para diferentes tamanhos de observações a treino da classe 0, utilizando diferentes técnicas de *undersampling* e *oversampling*.

Na Tabela 2 encontram-se ilustrados os resultados para as diferentes abordagens, utilizando diferentes modelos base e técnicas de *undersampling* e *oversampling* para equilíbrio das observações das classes. Nas colunas 'OS' e 'US' são designadas as técnicas de *oversampling* e *undersampling*, respetivamente, para aumentar ou diminuir as observações das classes maioritárias e/ou minoritárias. No que diz respeito às técnicas de *oversampling* são denominadas pelos seguintes números: 0 - sem técnica aplicada, 1- *SMOTE*, 2- *ADASYN*, 3- *BorderlineSMOTE*, 4- *RandomOverSampler*, 5- *SMOTENC*, 6- *SVMSMOTE*,7- *SMOTEENN*, 8- *SMOTETomek*. As técnicas de *undersampling* são resumidas em: 1- *RandomUnderSampler*, 2- *ClusterCentroids*, 3- *NearMiss*, 4- *TomekLinks*, 5- *NeighbourhoodCleaningRule* , 6- *NearMiss*. Foram utilizadas diferentes combinações de técnicas de *undersampling* e *oversampling* para que fosse possível fornecer aos modelos o melhor equilíbrio de dados para treino, a fim de generalizar melhor no conjunto de teste e obter melhores desempenhos.

Os modelos base utilizando técnicas de *undersampling* e *oversampling* conseguiram generalizar o treino para as observações no conjunto de teste, sendo o melhor modelo, tendo em conta a métrica GM, capaz de acertar em 82 verdadeiros positivos e 75326

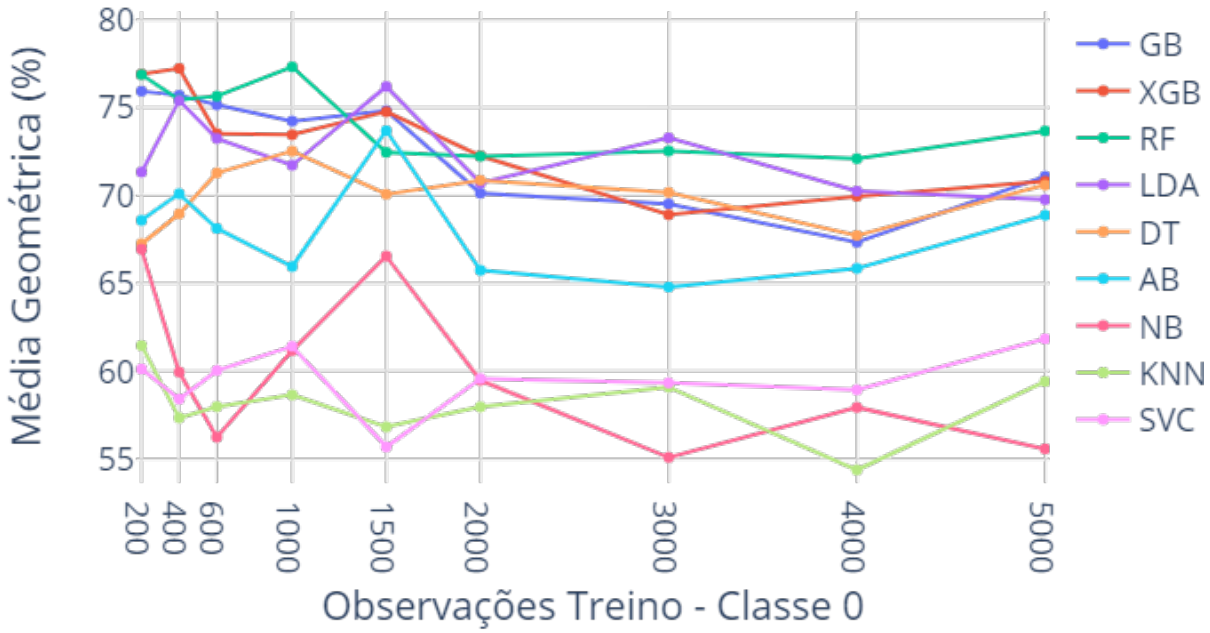


FIGURA 6.2. Comparação entre modelos com diferentes tamanhos para o treino da classe 0 utilizando diferentes técnicas de *undersampling* e *oversampling*.

TABELA 2. Melhores resultados para os modelos base utilizando técnicas de *undersampling* e *oversampling*

.

C0	C1	C2	C3	OS	US	Modelo	VP	VN	FP	FN	TVP	TVN	GM
1000	986	991	977	2	1	RF	82	75326	6677	44	65,08%	91,86%	**77,32%**
400	65	114	120	0	5	XGB	80	77036	4967	46	63,49%	93,94%	77,23%
200	65	114	120	0	4	XGB	85	71308	10695	40	68,00%	86,96%	76,90%
200	65	114	120	0	4	RF	80	74518	7485	43	65,04%	90,87%	76,88%
1500	500	500	500	3	1	LDA	77	75438	6565	45	63,11%	91,99%	76,20%
200	65	114	120	0	5	GB	**95**	60714	21289	**27**	**77,87%**	74,04%	75,93%
400	65	114	120	0	6	GB	82	69916	12087	40	67,21%	85,26%	75,70%
600	608	573	591	2	1	RF	82	71539	10464	43	65,60%	87,24%	75,65%
400	65	114	120	0	6	RF	73	**79934**	**2069**	52	58,40%	**97,48%**	75,45%
400	65	114	120	0	6	LDA	81	71937	10066	44	64,80%	87,72%	75,40%

verdadeiros negativos. Assim, foi possível aumentar o desempenho dos modelos tendo em conta a métrica GM para 77,32% utilizando o algoritmo Random Forest.

Comparando os modelos base em que apenas foi aplicado *randomundersampling* com os modelos base utilizando diferentes técnicas de *undersampling* e *oversampling*, o melhor modelo desta última abordagem aumentou tanto o número de VP de 80 para 82 como o número de VN de 70319 para 75326, como diminui tanto o número de FP de 11684 para 6677 e FN de 45 para 44.

Do ponto de vista da métrica TVP/ *recall*, foi atingido o melhor resultado com um modelo utilizando o algoritmo GB, acertando em 95 verdadeiros positivos, num total de 128, o que significa que o modelo consegue selecionar relativamente bem a classe positiva. Contudo, a taxa de verdadeiros negativos para o modelo GB em questão foi das mais baixas neste conjunto de testes, o que indica uma fraca capacidade do modelo para classificar verdadeiros negativos, originando uma grande quantidade de falsos positivos (21289). Se a empresa pretender arriscar um pouco mais para que seja maximizada a probabilidade de serem compradas experiências, podem ser utilizados os modelos GB e XGB que tiveram uma TVP de 77,87% e 68,00%, resultando em 95 e 85 VP, respetivamente, sob pena de haver um maior número de falsos positivos e por consequência um maior risco dos clientes saírem do website.

De uma perspetiva de maximização da métrica TVN, os modelos utilizando XGB e RF, treinados com 400 observações para a classe 0 obtiveram uma TVN de 93,94% e 97,48%, respetivamente, o que minimiza o risco de falsos positivos, e desta forma não se torna cansativo para os clientes com *cross-selling* desnecessário. Ainda assim, quando comparada a métrica TVN com os resultados apresentados na Tabela 1, esta não foi otimizada, mantendo-se o melhor resultado com o modelo LDA treinado com 3000 observações utilizando apenas *randomundersampling*, onde foi obtido um resultado de 98,71% enquanto que os resultados da Tabela 2 não foram além dos 97,48%. Utilizando diferentes técnicas de amostragem, o modelo XGB registou a melhor TVN, sendo este treinado com 400 observações da classe 0, obtidas através da técnica de *undersampling NeighbourhoodCleaningRule* e cerca de 1000 observações para cada uma das classes das experiências, geradas através da técnica de *oversampling ADASYN*. O modelo RF foi treinado com 400 observações da classe 0 obtidas através da técnica de *undersampling NearMiss* e não foram aplicadas técnicas de *oversampling* no conjunto de treino. O modelo XGB apesar de ter uma TVN menor quando comparada com o modelo RF, este obteve uma maior TVP o que indica um maior equilibrio entre o número de acertos tanto na classe positiva como na negativa. Contudo, se o objetivo passa pela minimização do fator incomodativo dos clientes com sucessivas recomendações, minimizando os falsos positivos, o modelo RF é uma boa escolha que acerta em quase todos os casos de apenas compra de transfer com um número muito baixo de falsos positivos (2069). Comparando com os resultados da Tabela 1, tendo em conta uma perspetiva de minimização do número de FP, os modelos utilizando diferentes técnicas de *undersampling* e *oversampling* tiveram piores resultados, tendo estes modelos como melhor resultado de TVN 97,48% e 2069 FP, enquanto que o melhor modelo em que apenas foi aplicada a técnica de *undersampling randomundersampling*, da mesma perspetiva, obteve uma TVN de 98,73% e 1038 FP. Porém, de ressalvar que do ponto de vista da métrica GM, que considera tanto a métrica TVP e TVN, os modelos utilizando diferentes técnicas de *undersampling* e *oversampling* obtiveram melhores resultados onde, no geral, foi melhorado tanto o número de acertos em VP como em VN, passando de 80 VP para 82 e de 70319 VN para 75326.

Em seguida são apresentados os resultados para os modelos com a nova abordagem híbrida desenvolvida, sob forma de ser melhorado os resultados apresentados anteriormente.

6.3. Resultados dos Modelos Híbridos

O desenvolvimento de um modelo híbrido objetiva melhorar o desempenho dos modelos anteriormente testados, de forma a ser possível melhorar o desempenho das classificações, minimizando os FP e FN, aumentando por consequência o número de VP e VN.

De forma semelhante aos testes realizados com os modelos base, numa primeira abordagem foi utilizado apenas *randomundersampling* para equilibrar as observações. Os resultados demonstram uma aumento das diferentes métricas nos vários testes realizados, melhorando o número de verdadeiros negativos e verdadeiros positivos nas classificações de cada modelo. A Tabela 3 apresenta os resultados para os diferentes testes utilizando diferentes modelos para uma abordagem híbrida com apenas *randomundersampling*, ordenados pela métrica GM.

Na Fig. 6.3 é ilustrada o desempenho dos diferentes modelos híbridos, representado a tracejado a primeira fase do modelo híbrido, o modelo binário, e representado a linha continua o desempenho dos modelos multi-classe na segunda fase. Verifica-se que os melhores resultados tendo em conta a métrica média geométrica são obtidos com modelos utilizando 600 exemplos de treino para a classe 0 e algoritmos GB e XGB. Apesar dos resultados da segunda fase do modelo híbrido (multi-classe) para a mesma quantidade de exemplos de treino (600) ser mais baixa em relação a outros testes representados na imagem, os modelos binários representados a tracejado, no contexto do problema, são os mais importantes dado que identificam claramente se um cliente vai ou não comprar uma experiência.

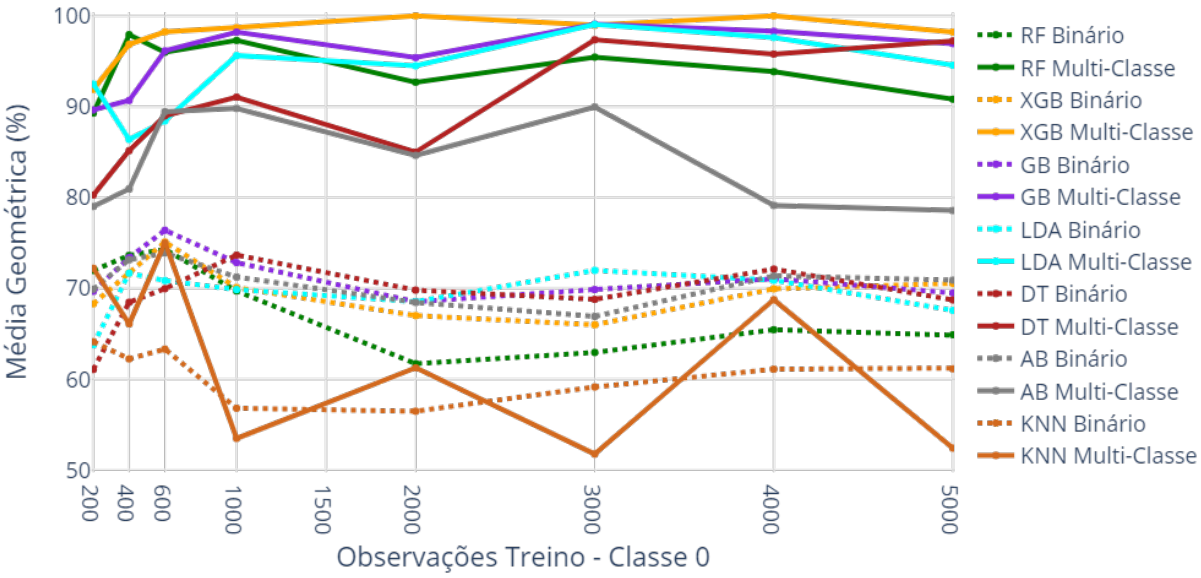


FIGURA 6.3. Melhores resultados para os modelos híbridos apenas com *randomundersampling*.

Tabela 3. Melhores resultados utilizando modelos híbridos com *randomundersampling*.

C0	Modelo	VP	VPm	VN	FP	FN	TVP	TVN	GM
600	GB	86	**81**	71234	10769	42	67,19%	86,87%	**76,40%**
600	XGB	74	72	80004	1999	54	57,81%	97,56%	75,10%
600	RF	72	68	**80471**	**1532**	56	56,25%	**98,13%**	74,30%
600	AB	76	65	75430	6573	52	59,38%	91,98%	73,90%
1000	DT	82	72	69467	12536	46	64,06%	84,71%	73,67%
400	RF	72	70	79039	2964	56	56,25%	96,39%	73,63%
400	GB	**88**	77	64224	17779	**40**	**68,75%**	78,32%	73,38%
400	AB	87	66	64561	17442	41	67,97%	78,73%	73,15%
1000	GB	74	72	75248	6755	54	57,81%	91,76%	72,84%
4000	DT	70	67	78002	4001	58	54,69%	95,12%	72,12%

Os melhores modelos foram obtidos com os algoritmos GB e XGB, apontando 86 e 74 verdadeiros positivos, respetivamente. Apesar do modelo XGB apenas ter acertado em 74 verdadeiros positivos, o número de verdadeiros negativos foi superior, apresentando uma das melhores taxas de verdadeiros negativos (97.56%). Como o primeiro objetivo do problema é maximizar o número de verdadeiros positivos, não descorando dos acertos nos verdadeiros negativos, o modelo treinado com o algoritmo GB com 600 exemplos da classe 0 é o mais adequado com 76.40% de GM.

Comparando com os resultados obtidos nos modelos base, tendo em conta a mesma abordagem com a técnica *randomundersampling*, foi possível aumentar o desempenho do melhor modelo em 2,32 pontos percentuais, sendo que o melhor modelo base do ponto de vista da métrica GM obteve 74,08% de GM e utilizando o modelo híbrido foi possível aumentar para 76,40%, aumentando o número de verdadeiros positivos de 80 para 86 e o número de verdadeiros negativos de 70319 para 71236, utilizando um modelo GB em vez de XGB.

Do ponto de vista da métrica TVP, tendo em conta os modelos base e híbridos utilizando *randomundersampling*, foi possível aumentar a métrica de 64% para 68,75%, o que indica que os VP aumentaram de 80 para 86 observações corretamente identificadas e os FN diminuíram de 45 para 42 observações mal classificadas. No entanto, quando comparada a abordagem híbrida utilizando *randomundersampling* com os modelos base utilizando diferentes técnicas *undersampling* e *oversampling*, ou seja, comparando as Tabelas 3 e 2, e do ponto de vista da métrica TVP, os resultados dos modelos base são superiores onde foi obtida uma TVP máxima de 77,87%, enquanto que os modelos híbridos não foram além dos 68,75%, o que indica que o modelo base GB com a técnica de *undersampling NearMiss* é um melhor classificador de verdadeiros positivos que minimiza os falsos negativos.

Em relação à metrica TVN, analisando os modelos base e híbridos utilizando *randomundersampling*, os modelos híbridos conseguiram um máximo para a métrica de 98,13% enquanto que os modelos base obtiveram um máximo de 98,73%, traduzindo-se em mais VN corretamente identificados e um menor número de FP. Posto isto, do ponto de vista da métrica TVN o modelo base consegue ter uma melhor especificidade com 80965 VN e 1038 FP, enquanto que o melhor modelo híbrido conseguiu 80471 VN e 1532 FP. Todavia, o modelo híbrido conseguiu uma melhor TVP com 74 VP em comparação aos 66 VP do modelo base em questão.

Ao utilizar técnicas de *undersampling* e *oversampling* nos modelos híbridos é possível verificar um comportamento semelhante ao fazer variar o número de observações das classes a treino (Fig. 6.3). Novamente, utilizando 600 exemplos de treino para a classe 0 foi possível obter o melhor desempenho, tendo em consideração a métrica média geométrica. Os modelos utilizando algoritmos AB e GB obtiveram os melhores resultados, contudo, o modelo utilizando AB apesar de ser um ótimo classificador binário, ao passar para a segunda fase do modelo híbrido, o modelo multi-classe, falha em conseguir identificar qual das experiências o cliente irá comprar. Em contrapartida, o modelo híbrido utilizando o algoritmo GB é ótimo a classificar tanto na 1ª fase (binário) como na 2ª fase (multi-classe) do classificador. Ao selecionar um número de observações maior que 600 para a classe 0, os modelos tendem a perder desempenho.

A Tabela 4 apresenta detalhadamente os resultados obtidos para os diferentes modelos apresentados na Fig. 6.4.

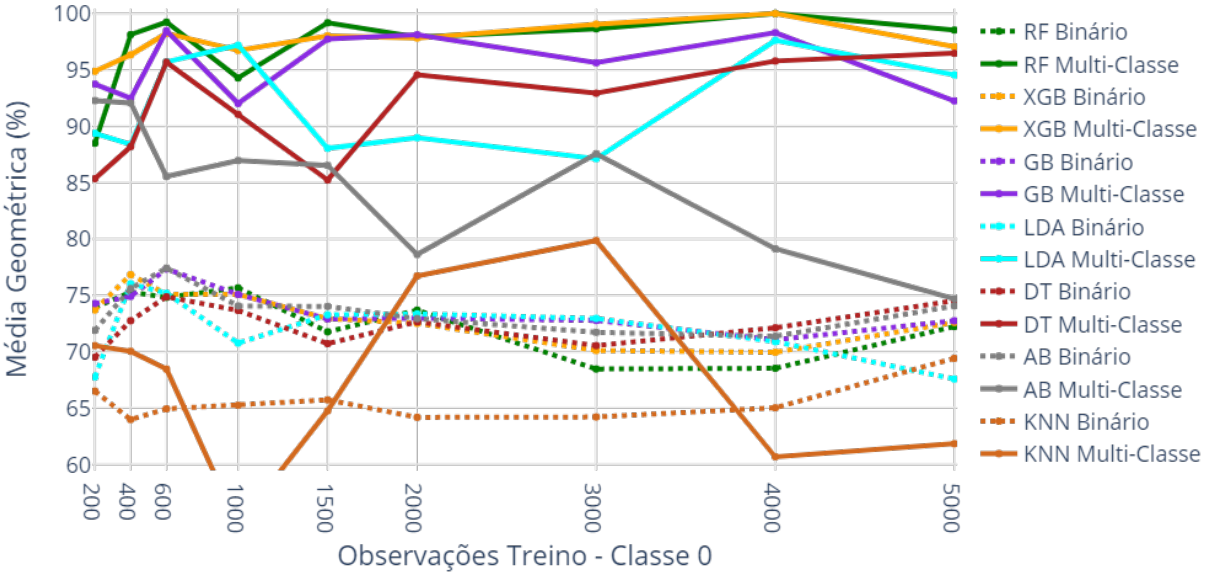


FIGURA 6.4. Comparação entre modelos híbridos com diferentes tamanhos para o treino da classe 0 utilizando diferentes técnicas de *undersampling* e *oversampling*.

Utilizando a técnica de *undersampling*, *NeighbourhoodCleaningRule*, foi possível obter os melhores desempenhos para os modelos utilizando algoritmos AB, GB e XGB, por esta ordem de desempenho. Apesar do modelo AB apresentar uma maior percentagem de

GM (77,41%) este classificou menos dois exemplos de verdadeiros positivos que o modelo utilizando o algoritmo GB, contudo, apresenta uma maior sensibilidade para verdadeiros negativos. Estes dois primeiros modelos diferem-se principalmente na capacidade de identificar qual das experiências o cliente irá comprar, sendo que o modelo com o algoritmo AB obteve um desempenho inferior ao GB. Para o contexto do problema da empresa, provavelmente será de mais valia utilizar o modelo AB, dado que existe um melhor equilíbrio tanto a nível de classificação de verdadeiros positivos como de verdadeiros negativos, partindo do principio que são recomendadas todas as três experiências sem distinção especifica entre elas.

TABELA 4. Melhores resultados utilizando modelos híbridos com técnicas de *oversampling* e *undersampling*.

C0	C1	C2	C3	OS	US	Modelo	VP	VPm	VN	FP	FN	TVP	TVN	GM
600	299	114	120	0	5	AB	86	71	73144	8859	42	67,19%	89,20%	**77,41%**
600	299	114	120	0	5	GB	88	**86**	71350	10653	40	68,75%	87,01%	77,34%
400	299	114	120	0	5	XGB	82	78	75575	6428	46	64,06%	92,16%	76,84%
400	299	114	120	0	4	LDA	89	75	68166	13837	39	69,53%	83,13%	76,03%
1000	2954	991	977	2	1	RF	**90**	83	66769	15234	**38**	**70,31%**	81,42%	75,66%
400	299	114	120	0	5	AB	89	80	67412	14591	39	69,53%	82,21%	75,60%
400	299	114	120	0	6	RF	77	75	77275	4728	51	60,16%	94,23%	75,29%
600	299	114	120	0	6	LDA	78	74	76095	5908	50	60,94%	92,80%	75,20%
600	299	114	120	0	1	XGB	74	72	**80004**	**1999**	54	57,81%	**97,56%**	75,10%
1000	299	114	120	0	5	GB	78	70	75872	6131	50	60,94%	92,52%	75,09%

Os modelos híbridos utilizando técnicas de *undersampling* e *oversampling* e analisando a métrica TVP, não conseguiram obter melhores resultados que os modelos base com a mesma aplicação de técnicas de amostragem. O modelo híbrido que conseguiu obter uma melhor TVP foi o LDA com um valor de 70,31% enquanto que os modelos base aplicando as mesmas técnicas conseguiram obter um melhor resultado de 77,87%. Analisando a métrica TVN os modelos híbridos com técnicas de *undersampling* e *oversampling* conseguiram obter um máximo de 97,56% enquanto que os modelos base com a mesma aplicação de técnicas de amostragem, conseguiram obter um resultado de 97,48%.

De uma perspectiva geral, avaliando os modelos híbridos desenvolvidos como forma de melhorar o desempenho das classificações em conjuntos de dados desequilibrados, estes conseguiram melhorar a média geométrica em 2,32% nos modelos em que apenas foi aplicado *randomundersamplin* e em 0,09% comparando os modelos base com os modelos híbridos utilizando as diferentes técnicas de *undersampling* e *oversampling*. O melhor modelo, do ponto de vista da métrica GM, foi obtido com um modelo híbrido utilizando apenas a técnica de *undersampling*, *NeighbourhoodCleaningRule*, onde foi obtido um total de 86 verdadeiros positivos e 73144 verdadeiros negativos. Se o objetivo da empresa for o de manter tanto as métricas TVP e TVN equilibradas, dando valor tanto aos falsos

positivos como aos falsos negativos, o modelo híbrido AB treinado com 600 observações da classe 0, em que estas foram selecionadas através da técnica *NeighbourhoodCleaningRule*, é o mais indicado. Porém, se a empresa pretende maximizar o número de possíveis conversões, recomendando a um maior número de pessoas os serviços de experiencia e não considerando o facto dos falsos positivos poderem ser contrapruducentes e fazer com que os clientes desistam da compra e saiam do website, o modelo base utilizando a técnica de *undersampling NeighbourhoodCleaningRule* é o mais indicado, onde foi possível maximizar o número de VP e minimizar o número de FN, ainda assim, com uma TVN baixa de 74,04%, resultando em cerca de 21289 FP. Se o objetivo da empresa é conseguir ocasionalmente recomendar um serviço de experiencia, sem colocar em causa a harmonia visual aquando navegação no website com recomendação desnecessária, a métrica TVN deve ser tida como principal fator. Para isso, deve ser escolhido o modelo base LDA treinado com 10000 observações da classe 0, onde foi só aplicada a técnica *randomundersampling*, em que este obteve uma TVN de 98,73%, refletindo-se em 80965 VN e 1038 FP. No entanto, deve ser tida em consideração que este modelo LDA alcançou um dos mais baixos valores para a métrica TVP com 51,97%, traduzindo-se em 66 VP num total de 128 experiências a teste.

CAPíTULO 7

Conclusão, Limitações e Investigação Futura

A capacidade computacional têm alavancado a evolução das técnicas de *machine learning*, ao mesmo tempo que a aplicação das mesmas nos modelos de negócio é cada vez mais frequente nos vários setores, como é o caso do turismo. A grande quantidade de dados gerada pelos clientes quer nos websites das empresas quer nos dados fornecidos diretamente para reservas, são extremamente ricos para definir a jornada do cliente e sugerir novos produtos/serviços baseados nas características dos clientes, assim como, para melhorar e personalizar a experiência online dos mesmos.

No decorrer deste documento foi apresentada uma análise dos clientes de serviços de transfers através de métodos de exploração dos dados e segmentação do clientes, de forma a distinguir como é que os clientes de transfers se correlacionam de forma específica relevante para campanhas de marketing. Analisando os resultados da segmentação de clientes inferiu-se que os clientes de transfers estão divididos em três grupos com diferenças demográficas e comportamentais. O primeiro *cluster* representa os clientes que costumam reservar o serviço de transfer em maio, chegam em junho e têm uma estadia média de seis dias. O segundo *cluster* representa os clientes que reservam mais tarde (agosto) e com um tempo de antecedência mínimo. Por fim, o terceiro *cluster* representa os clientes que viajam com crianças ou bebés e trazem bagagem mais variada.

A percepção e a análise do tipo de clientes que compra serviços de transfers é uma mais-valia para a empresa no sentido em que permite uma melhor alocação de recursos de marketing da empresa para que a venda destes serviços possa ser maximizada tendo em conta cada grupo de clientes, maximizando assim as oportunidades de *cross-selling* e *up-selling*. O envio de emails específicos para as necessidades dos clientes e ofertas especiais para as características/padrões encontrados na segmentação de clientes para cada grupo de clientes, é uma das aplicabilidades desta análise em contexto real. A YellowFish Transfers beneficiará da segmentação de clientes no sentido em que ficará um passo à frente da concorrência, identificando e oferecendo novos serviços/produtos que os existentes ou potenciais clientes possam estar interessados, tendo em conta as características encontradas.

O comportamento dos clientes na navegação e reserva de serviços no website da Yellowfish caracterizam-se em termos de número de visitas no website, o país dos visitantes, a fonte proveniente dos visitantes e a taxa de conversão por mês, por sistema operativo utilizado pelos clientes, e por período do dia. A maior afluência ao website acontece durante as segundas, terças e domingos onde são registados o maior número de visitas. O Reino Unido é o país de onde são originários a maior fatia de visitantes das páginas da

Yellowfish e logo de seguida, a Irlanda, Portugal, Estados Unidos, Alemanha, Holanda e França. Cerca de um quarto do número total de visitantes na página da YellowFish Transfers converte através da compra de um serviço de transfer. Nos meses de março, abril, maio, junho e julho, foram registadas taxas de conversão de 13%, 15%, 17%, 16% e 22%, respetivamente. Os visitantes do website da Yellowfish que utilizam sistemas operativos Windows ou iOS são os que mais reservas fazem no site website e a parte do dia em que é gerada mais receita acontece durante o período da manhã. Contudo, de um ponto de vista geral, as melhores alturas do dia a nível de vendas acontecem durante o período da manhã e durante o anoitecer, onde são geradas mais receitas. Grande parte dos utilizadores que visita o website é proveniente do motor de busca da Google, através da procura voluntária e sem pagamento pela empresa para tal. A segunda origem da pesquisa aparece no canal (direct), que indica tráfego onde a referência ou fonte é desconhecida.

Foi ainda proposto e desenvolvido um modelo de classificação híbrido capaz de identificar clientes potencialmente interessados na aquisição de experiências de lazer. Dadas as características de cada reserva de transfer e as características de cada cliente, foram desenvolvidos modelos capazes de identificar possíveis clientes de serviços adicionais e sugerir um destes serviços em especifico a estes clientes. Os modelos de recomendação (híbridos) desenvolvidos melhoraram as taxas de verdadeiros positivos e negativos, assim como da métrica utilizada para comparar os diferentes modelos, média geométrica. Numa primeira fase, o modelo binário consegue identificar o comportamento dos clientes que realmente pretendem reservar serviços, dos restantes, que apenas visitam o site. De um ponto de vista do negócio, esta primeira fase, que identifica se um cliente irá comprar uma experiência é mais importante do que a segunda fase do modelo, que classifica o tipo de experiência que o cliente vai comprar. Como existem apenas três tipos de experiências e supondo que o modelo indica que um determinado cliente é provável que compre uma experiência, simplesmente pode ser recomendado os três tipos em vez de uma específica. Os resultados obtidos com o modelo de classificação híbrida são significativamente afetados pelo fato de haver menos observações para as classes de experiência. Ainda de salientar que a identificação do *cluster* a que pertence cada cliente na segmentação de clientes, não afetou o desempenho dos modelos.

A utilização de uma abordagem híbrida melhorou a classificação (multi-classe) dos tipos de experiências que um determinado cliente de transfers pode vir a comprar. Os algoritmos que de uma forma geral obtiveram melhores resultados em contexto de dados desequilibrados foram: XGB, GB, RF e AB. Quanto às técnicas de *undersampling* e *oversampling* que mais contribuíram para melhorar o desempenho dos modelos, destacam-se a *ADASYN*, para *oversampling* e *TomekLinks*, *NeighbourhoodCleaningRule* para *undersampling*.

Do ponto de vista do negócio, o sistema de recomendação desenvolvido através dos vários modelos apresentados, pode ser aplicado para que exista uma fomentação da estratégia de *cross-selling* da empresa, ao sugerir a compra de uma experiência aos clientes

de transfers da YellowFish Transfers. A oferta cada vez maior de serviços terceiros de transfers e experiências lúdicas, faz com que os clientes possam ser divergidos para essas empresas. Dado que apenas uma pequena percentagem dos clientes de transfers (0,20%) adquire experiências, ao ser aplicado o sistema de recomendação a probabilidade dos clientes de transfers também comprarem um serviço de experiência com a empresa é maior, assim como, é maior a probabilidade destes clientes não divergirem nem procurarem noutro fornecedor este tipo de experiências. O sistema de recomendação desenvolvido filtra os clientes que devem ser selecionados para serem alvo da recomendação de experiências, sendo estes apenas os visados na apresentação dos anúncios para a compra de experiências, minimizando assim o fator incomodativo para os clientes que não pretendem adquirir qualquer tipo de outros serviços adicionais.

7.1. Contribuições

Com esta dissertação foi possível analisar e identificar os comportamentos e características dos clientes de serviços de transfer e serviços de experiências. O desequilíbrio no número total de observações entre as classes de transfer e experiências, originou uma pesquisa exaustiva pelas melhores técnicas que pudessem lidar com dados desequilibrados. Foi feita uma análise das técnicas de *undersampling* e *oversampling* através da sua implementação na preparação de diferentes conjuntos de dados para treino. A técnica *ADASYN* para *oversampling* e *TomekLinks*, *NeighbourhoodCleaningRule* para *undersampling* destacaram-se pela otimização do desempenho dos modelos ao serem utilizadas. O modelo de classificação híbrido desenvolvido foi capaz de dar resposta ao objetivo inicial, melhorando os resultados dos modelos base que incluem todas as classes numa só vez. Desta forma, este modelo pode ser considerado em futuras abordagens em que existam dados desequilibrados, juntamente com técnicas de equilíbrio dos dados. Parte do desenvolvimento desta dissertação originou a escrita de um artigo (Anexo A), denominado de "Using Customer Segmentation to Build a Hybrid Recommendation Model"aceite e apresentado na conferência "ICOTTS'20 - International Conference on Tourism, Technology Systems"em que se apresentou os resultados para a segmentação dos clientes e os resultados dos modelos base e híbridos utilizando apenas a técnica de *undersampling* inicial *randomundersampler*. O artigo encontra-se ainda em no livro "Advances in Tourism, Technology and Systems", capítulo número vinte e sete e capítulo com o DOI 10.1007/978-981-33-4256-9_27. Pretendemos ainda submeter outro artigo com os novos resultados para os modelos utilizando as diferentes técnicas de *undersampling* e *oversampling*.

7.2. Limitações e Investigação Futura

O baixo número de registos de experiências limitou o âmbito do sistema de recomendação, não sendo possível os modelos treinarem com mais exemplos e assim, por consequência, aprender mais acerca destes registos e melhorar o desempenho na classificação de possíveis compradores de experiências.

Dado que o Google Analytics limita os dados de *clickstream* que podem ser extraídos das suas bases de dados, impossibilitou-nos de utilizar estes mesmos dados no sistema de recomendação e ser um sistema holístico complementado com os dados característicos da navegação dos clientes no website.

O objetivo dos modelos construídos seria de aplicar online no website da YellowFish Transfer, mas o seu *deployment* não estava no âmbito da tese, prevendo-se que isso possa ser feito posteriormente pela própria empresa.

Como investigação futura pretendemos fazer a integração com os dados de *clickstream*, utilizando para isso, métodos próprios da empresa para registar a jornada do cliente no seu website. Assim não será necessário pagar valores avultados pela subscrição do serviço Google 360 e ter acesso a todos os registos de *clickstream* no Google Analytics.

Referências

[1] Aluri, A., Price, B.S., McIntyre, N.H.: Using machine learning to cocreate value through dynamic customer engagement in a brand loyalty program. Journal of Hospitality & Tourism Research, 10963480, 43(1), 78–100 (2019)

[2] Mariani, M., Baggio, R., Fuchs, M., Höepken, W.: Business intelligence and big data in hospitality and tourism: a systematic literature review. International Journal of Contemporary Hospitality Management, 0959-6119, 30(12), 3514–3554 (2018)

[3] Li, S.S., Karahanna, E.: Online recommendation systems in a b2c e-commerce context: a review and future directions. Journal of the Association for Information Systems, 1536-9323, 16(2), 72 (2015)

[4] Prideaux, B.: Tourism and surface transport. The Sage handbook of tourism management pp. 297–313 (2018)

[5] INE, I.N.d.E.: Estatísticas do turismo 2018. Estatísticas do Turismo 2018: 978-989-25-0497-1 p. 125 (2019)

[6] Çetin, T.: The rise of ride sharing in urban transport: Threat or opportunity? Urban Transport Systems p. 191 (2017)

[7] ECO, S.: Uber vai ter mais concorrência no verão. cabify estreia-se no algarve. `https://eco.sapo.pt/2018/06/07/uber-vai-ter-mais-concorrencia-no-verao-cabify-estreia-se-no-algarve/` (2019), accessed: 2019-05-01

[8] More, A.: Survey of resampling techniques for improving classification performance in unbalanced datasets. arXiv preprint arXiv:1608.06048 (2016)

[9] Lemon, K.N., Verhoef, P.C.: Understanding customer experience throughout the customer journey. Journal of Marketing, 00222429, 80(6), 69–96 (2016), `https://doi.org/10.1509/jm.15.0420`

[10] Bolton, R.N., McColl-Kennedy, J.R., Cheung, L., Gallan, A., Orsingher, C., Witell, L., Zaki, M.: Customer experience challenges: bringing together digital, physical and social realms. Journal of Service Management, 1757-5818, (2018)

[11] Alaei, A.R., Becken, S., Stantic, B.: Sentiment analysis in tourism: capitalizing on big data. Journal of Travel Research, 00472875, 58(2), 175–191 (2019)

[12] Dalessandro, B., Hook, R., Perlich, C., Provost, F.: Evaluating and optimizing online advertising: Forget the click, but there are good proxies. Big data 3(2), 90–102 (2015)

[13] Hu, Y.H., Lee, P.J., Chen, K., Tarn, J.M., Dang, D.V.: Hotel recommendation system based on review and context information: a collaborative filtering appro. In: PACIS. p. 221 (2016)

[14] Takuma, K., Yamamoto, J., Kamei, S., Fujita, S.: A hotel recommendation system based on reviews: What do you attach importance to? In: 2016 Fourth International Symposium on Computing and Networking (CANDAR). pp. 710–712. IEEE (2016)

[15] Zhang, K., Wang, K., Wang, X., Jin, C., Zhou, A.: Hotel recommendation based on user preference analysis. In: 2015 31st IEEE International Conference on Data Engineering Workshops. pp. 134–138. IEEE (2015)

[16] Lu, J., Wu, D., Mao, M., Wang, W., Zhang, G.: Recommender system application developments: a survey. Decision Support Systems 74, 12–32 (2015)

[17] Masiero, L., Zoltan, J.: Tourists intra-destination visits and transport mode: A bivariate probit model. Annals of Tourism Research 43, 529–546 (2013)

[18] Loi, L.T.I., So, A.S.I., Lo, I.S., Fong, L.H.N.: Does the quality of tourist shuttles influence revisit intention through destination image and satisfaction? the case of macao. Journal of Hospitality and Tourism Management 32, 115–123 (2017)

[19] Lemon, K.N., Verhoef, P.C.: Understanding customer experience throughout the customer journey. Journal of marketing, 00222429, 80(6), 69–96 (2016)

[20] Richardson, A.: Using customer journey maps to improve customer experience. Harvard business review, 0017-8012, 15(1), 2–5 (2010)

[21] Zomerdijk, L.G., Voss, C.A.: Service design for experience-centric services. Journal of Service Research, 10946705, 13(1), 67–82 (2010)

[22] Følstad, A., Kvale, K.: Customer journeys: a systematic literature review. Journal of Service Theory and Practice 28(2), 196–227 (2018)

[23] Plaza, B.: Google analytics for measuring website performance. Tourism Management, 0261-5177, 32(3), 477–481 (2011)

[24] Hyken, S.: Personalized customer experience increases revenue and loyalty (2017)

[25] Sheil, H., Rana, O., Reilly, R.: Predicting purchasing intent: automatic feature learning using recurrent neural networks. arXiv preprint arXiv:1807.08207 (2018)

[26] Micarelli, A., Gasparetti, F., Sciarrone, F., Gauch, S.: Personalized search on the world wide web. vol. 4321, pp. 195–230 (01 2007)

[27] Massoud, M., Abo-Rizka, M.: A conceptual model of personalized pricing recommender system based on customer online behavior. International Journal of Computer Science and Network Security (IJCSNS) 12(6), 129 (2012)

[28] Chen, L., Su, Q.: Discovering user's interest at e-commerce site using clickstream data. In: 2013 10th International Conference on Service Systems and Service Management. pp. 124–129. IEEE (2013)

[29] Oard, D.W., Kim, J., et al.: Implicit feedback for recommender systems. In: Proceedings of the AAAI workshop on recommender systems. vol. 83. WoUongong (1998)

[30] Rashid, A.M., Albert, I., Cosley, D., Lam, S.K., McNee, S.M., Konstan, J.A., Riedl, J.: Getting to know you: learning new user preferences in recommender systems. In: Proceedings of the 7th international conference on Intelligent user interfaces. pp. 127–134. ACM (2002)

[31] Schafer, J.B., Konstan, J., Riedi, J.: Recommender systems in e-commerce. inproceedings of the 1st acm conference on electronic commerce (denver, colorado, united states, november 03-05, 1999). ec'99 (1999)

[32] Cho, Y.H., Kim, J.K., Kim, S.H.: A personalized recommender system based on web usage mining and decision tree induction. Expert systems with Applications 23(3), 329–342 (2002)

[33] Thorat, P.B., Goudar, R., Barve, S.: Survey on collaborative filtering, content-based filtering and hybrid recommendation system. International Journal of Computer Applications 110(4), 31–36 (2015)

[34] Resnick, P., Varian, H.R.: Recommender systems. Communications of the ACM 40(3), 56–59 (1997)

[35] Ekstrand, M.D., Riedl, J.T., Konstan, J.A., et al.: Collaborative filtering recommender systems. Foundations and Trends® in Human–Computer Interaction 4(2), 81–173 (2011)

[36] Lops, P., De Gemmis, M., Semeraro, G.: Content-based recommender systems: State of the art and trends. In: Recommender systems handbook, pp. 73–105. Springer (2011)

[37] Kaššák, O., Kompan, M., Bieliková, M.: Personalized hybrid recommendation for group of users: Top-n multimedia recommender. Information Processing & Management 52(3), 459–477 (2016)

[38] Codina, V., Ceccaroni, L.: A recommendation system for the semantic web. In: Distributed Computing and Artificial Intelligence, pp. 45–52. Springer (2010)

[39] Choi, K., Yoo, D., Kim, G., Suh, Y.: A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis. electronic commerce research and applications 11(4), 309–317 (2012)

[40] Xiao, Y., Ezeife, C.I.: E-commerce product recommendation using historical purchases and clickstream data. In: Ordonez, C., Bellatreche, L. (eds.) Big Data Analytics and Knowledge Discovery. pp. 70–82. Springer International Publishing, Cham (2018)

[41] Liu, J., Dolan, P., Pedersen, E.R.: Personalized news recommendation based on click behavior. In: Proceedings of the 15th international conference on Intelligent user interfaces. pp. 31–40. ACM (2010)

[42] Rawat, M., Goyal, N., Singh, S.: Advancement of recommender system based on clickstream data using gradient boosting and random forest classifiers. In: 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). pp. 1–6. IEEE (2017)

[43] Bennett, J., Lanning, S., et al.: The netflix prize. In: Proceedings of KDD cup and workshop. vol. 2007, p. 35. New York, NY, USA. (2007)

[44] Utku, A., Aydogan, E., Mutlu, B., Akcayol, M.A.: A new recommender system based on multiple parameters and extended user behavior analysis. In: Proceedings of the 9th International Conference on Information Management and Engineering. pp. 133–138. ACM (2017)

[45] Wei, W., Li, J., Cao, L., Ou, Y., Chen, J.: Effective detection of sophisticated online banking fraud on extremely imbalanced data. World Wide Web 16(4), 449–475 (2013)

[46] Ganganwar, V.: An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering 2(4), 42–47 (2012)

[47] Crone, S.F., Finlay, S.: Instance sampling in credit scoring: An empirical study of sample size and balancing. International Journal of Forecasting 28(1), 224–238 (2012)

[48] Drummond, C., Holte, R.C., et al.: C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In: Workshop on learning from imbalanced datasets II. vol. 11, pp. 1–8. Citeseer (2003)

[49] Elkan, C.: The foundations of cost-sensitive learning. In: International joint conference on artificial intelligence. vol. 17, pp. 973–978. Lawrence Erlbaum Associates Ltd (2001)

[50] Manevitz, L.M., Yousef, M.: One-class svms for document classification. Journal of machine Learning research, 1532-4435, 2(Dec), 139–154 (2001)

[51] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority oversampling technique. Journal of artificial intelligence research, 1076-9757, 16, 321–357 (2002)

[52] He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). pp. 1322–1328. IEEE (2008)

[53] Wu, X., Meng, S.: E-commerce customer churn prediction based on improved smote and adaboost. In: 2016 13th International Conference on Service Systems and Service Management (ICSSSM). pp. 1–5. IEEE (2016)

[54] Arefeen, M., Nimi, S.T., Rahman, M.S., et al.: Neural network based undersampling techniques. arXiv preprint arXiv:1908.06487 (2019)

[55] Yap, B.W., Rani, K.A., Rahman, H.A.A., Fong, S., Khairudin, Z., Abdullah, N.N.: An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In: Proceedings of the first international conference on advanced data and information engineering (DaEng-2013). pp. 13–22. Springer (2014)

[56] YellowfishTransfers: Sobre nós. https://www.yellowfishtransfers.com/en/about-us (2019), accessed: 2019-04-01

[57] Isinkaye, F., Folajimi, Y., Ojokoh, B.: Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal 16(3), 261 – 273 (2015), http://www.sciencedirect.com/science/article/pii/S1110866515000341

[58] YellowfishAdventures: Sobre nós. https://www.yellowfishadventure.com/sobre-nos/ (2019), accessed: 2019-04-01

[59] Apresentação da reporting api v4 do google analytics: início rápido de python para contas de serviço. https://developers.google.com/analytics/devguides/reporting/core/v4/quickstart/service-py, accessed: 2019-02-15

[60] Park, C., Kim, D., Oh, J., Yu, H.: Predicting user purchase in e-commerce by comprehensive feature engineering and decision boundary focused under-sampling. In: Proceedings of the 2015 International ACM Recommender Systems Challenge. RecSys '15 Challenge, Association for Computing Machinery, New York, NY, USA (2015), https://doi.org/10.1145/2813448.2813517

[61] holidays. https://pypi.org/project/holidays/, accessed: 2019-05-05

[62] Brito, P.Q., Soares, C., Almeida, S., Monte, A., Byvoet, M.: Customer segmentation in a large database of an online customized fashion business. Robotics and Computer-Integrated Manufacturing, 0736-5845, 36, 93–100 (2015)

[63] Anshari, M., Almunawar, M.N., Lim, S.A., Al-Mudimigh, A.: Customer relationship management and big data enabled: Personalization & customization of services. Applied Computing and Informatics (2018)

[64] Weinstein, A.: Customer retention: a usage segmentation and customer value approach. Journal of Targeting, Measurement and Analysis for Marketing, 1479-1862, 10(3), 259–268 (2002)

[65] Afrin, F., Al-Amin, M., Tabassum, M.: Comparative performance of using pca with k-means and fuzzy c means clustering for costumer segmentation. International Journal Of Scientific & Technology Research, 22778616, 4, 70–74 (2015)

[66] Kodinariya, T.M., Makwana, P.R.: Review on determining number of cluster in k-means clustering. International Journal 1(6), 90–95 (2013)

[67] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)

[68] imbalanced-learn. https://imbalanced-learn.readthedocs.io (2017), accessed: 2019-04-22

[69] Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 1532-4435, 18(17), 1–5 (2017), http://jmlr.org/papers/v18/16-365.html

[70] Mani, I., Zhang, I.: knn approach to unbalanced data distributions: a case study involving information extraction. In: Proceedings of workshop on learning from imbalanced datasets. vol. 126 (2003)

[71] Zhang, Y.P., Zhang, L.N., Wang, Y.C.: Cluster-based majority under-sampling approaches for class imbalance learning. In: 2010 2nd IEEE International Conference on Information and Financial Engineering. pp. 400–404. IEEE (2010)

[72] Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics SMC-2(3), 408–421 (1972)

[73] Tomek, I.: Two modifications of cnn. IEEE Trans. Systems, Man and Cybernetics 6, 769–772 (1976)

[74] Laurikkala, J.: Improving identification of difficult small classes by balancing class distribution. In: Conference on Artificial Intelligence in Medicine in Europe. pp. 63–66. Springer (2001)

[75] Smith, M.R., Martinez, T., Giraud-Carrier, C.: An instance level analysis of data complexity. Machine learning 95(2), 225–256 (2014)

Anexos

Anexo A - Artigo Aceite e Apresentado na Conferência ICOTTS'20

Using Customer Segmentation to Build a Hybrid Recommendation Model

Pedro Camacho[1], Ana de Almeida[1,4,5], and Nuno António[1,2,3]

[1] ISCTE - Instituto Universitário de Lisboa, Av. Das Forças Armadas, 1649-026 Lisboa, Portugal
[2] NOVA Information Management School, Universidade Nova de Lisboa - Campus de Campolide, 1070-312 Lisboa, Portugal
[3] CITUR - Centro de Investigação, Desenvolvimento e Inovação em Turismo, Rua Luís de Camões 139, 1300-357 Lisboa, Portugal
[4] ISTAR - Information Sciences and Technologies and Architecture Research Center, Av. Das Forças Armadas, 1649-026 Lisboa, Portugal
[5] CISUC - Center for Informatics and Systems of the University of Coimbra, R. Miguel Bombarda, 3030-790 Coimbra, Portugal

Abstract. The growing trend in leisure tourism has been closely followed by the number of hospitality services. Nowadays, customers are more sophisticated and demand a personalized and simplified experience, which is commonly achieved through the use of technological means for anticipating customer behaviour. Thus, the ability to predict a customer's willingness to buy is also a growing trend in hospitality businesses to reach more customers and consolidate existing ones. The acquisition of a transfer service through website reservation generates data that can be used to perform customer segmentation and enable recommendations for other products or services to a customer, like recreation experiences. This work uses data from a Portuguese private transfer company to understand how its private transfer business customers can be segmented and how to predict their behaviour to enhance services cross-selling. Information extracted from the data acquired with the private transfer reservations is used to train a model to predict customer willingness to buy, and based on it, offer leisure services to customers. For that, a hybrid classifier was trained to offer recommendations to a customer when he/she is booking a transfer. The model employs a two-phase process: first, a binary classifier asserts if the customer who's buying the transfer would eventually buy a service experience. In that case, a multi-class model decides what should be the most likely experience to be recommended.

Keywords: Hospitality, Transfers, Customer Segmentation, Recommendation System

1 Introduction

Hospitality and tourism are areas that have experienced massive growth over the last decades. The vast amount and diversity of data that hospitality generates

provides new perspectives and possibilities to improve customers' journey experience [1]. Hospitality businesses use digital marketing and online services to enhance sales by providing an improved and personalized customer experience.

Customer segmentation helps companies understand what patterns best describe customers in terms of purchased products or services. Companies can use this information in revenue management to provide distinct prices or personalized offers in terms of the different groups identified. Private transfer business is a subarea of hospitality that is responsible for the previously agreed transportation of customers between locations, using the company's private vehicles [2]. With the increase of private transportation service offers, companies are looking into enhancing the customers' experience and anticipating what customers may want to gain an advantage. Works can be found on the analysis of tourists' behaviour regarding movement patterns and transport modes [3], as well as the influence of transports in satisfaction to predict the intention for revisiting a destination [4]. However, there is a lack of research on understanding customers regarding private transfer business. What leads a customer to buy a transfer? How long in advance do customers make reservations? How could the transfer company predict if a website user will purchase a transfer? What products or services can be offered to enhance up-selling or cross-selling targets?

This paper presents a study of customers from a Portuguese private transfer company operating in a holiday resort. Customer segmentation results are discussed and a new classification model for recommending possible types of leisure experiences for customers that bought a transfer service is proposed. To deal with unbalanced data and enhance recommendation performance, the classification model is a two-phase approach.

This paper is organized as follows: next section presents a literature review regarding customer journey and hospitality analytics advances. Section 3 introduces the transfer business understanding, an exploratory data analysis of transfer purchase data, and a customer segmentation of the company's transfer reservations. The following section presents the hybrid classification model results whose targets are the experience/tour services. In the last section, conclusions are drawn, and lines for future work are discussed.

2 Literature Review

When surfing a commercial website connected with a dynamic information system, the website generates large amounts of data, for which techniques have been developed to analyze this information [5].

Customer journey [6] characterizes the set of events into which the customer interacts during his browsing of the website. One of the objectives of the customers' journey is to map diagrams that illustrate the main steps customers take when connecting with a particular company, whether it is a product, an online experience, shopping, services, or other combination [7]. Design and analysis of the customers' journey aim to maximize customer value. Personalization of the customers' experience attracts more visitors and increases customers' loyalty [8].

Experience customization is a growing investment area that requires attention from organizations as customers expect more and more of this kind of interaction. A recent survey found that customers have high expectations for personalized experiences on websites they visit [9], as well as, customers express their disappointment when they see it missing in their online shopping experiences. Online experience customization boosts purchases, profits, customer loyalty, and improves customer satisfaction in general [10].

By understanding the typical profiles of customers, companies can provide a more individual and personalized experience for each of these grouping profiles by exploring the relevant content and aspects that they share [11]. Large amounts of information about customers enable them to personalize the customers' journey, enhancing customer experience, customer loyalty, optimizing sales, and making it possible to reach a more significant number of users [12]. Customer segmentation turns the process of buying into a faster one while helping to build up loyalty if based on relevant interaction in the customers' journey [13].

To avoid overloading the customer, it is essential to gain perception about the customers' intention to complete a purchase while browsing within the website before suggesting recommendations for new products, services, or packages. Sheil *et al.* describe a neural network for the prediction of purchase intention in an e-commerce environment, addressing the significance of investing in feature engineering. Their results show that the model reached 98.4% of the area under the ROC Curve metric performance, predicting customer willingness to buy [14].

3 Private Transfer Company

3.1 Business Understanding

Private transfer business is an agreement for the transportation of customers from one location to another. Most of the time, in terms of ground transfer services, from an airport to a hotel and vice-versa. The company used in this study owns a vehicles' fleet (cars, mini-vans, and buses), and services can be booked either as one way or return service. Although mainly operating in Algarve, the company sells transfers between different locations of Portugal and allows reservations originating from different points of Portugal and, even, Spain.

Transfer reservations are made through the company website and can either be sold by affiliates or associated partners. The booking process starts by specifying the departure and arrival points and dates, hours, number/type of passengers and luggage. Service is confirmed when the payment is made. From this moment on, transfer service details are saved in a database that aggregates the reservations data. Although the company has its core business in transfer sales, it also sells leisure experiences in the Algarve region, using buggy or quad rides while touring around touristic points of interest. The experiences are sold in a parallel website. A customer completes the process of buying an experience, indicating the date/time, type of experience, person's quantity, and type of vehicle to use. The sale of experiences, similarly to the sale of transfers, may also be sold by third parties.

The company sells two types of experiences: Feel Tour and Experience Tour. A tour consists in a trip that explores Algarve typical villages and historical/cultural locations where only an all-terrain vehicle (buggy or motorcycle quad) can circulate. The Experience Tour has a 90-minute duration and the Feel Tour, 120 minutes.

3.2 Dataset Analysis

A brief exploratory analysis of the data is next presented, summarizing the main characteristics, patterns and insights. The dataset disclosed by the company presents 273768 transfer services, ranging between 2012-01-25 to 2019-11-30, and 427 of these observations have an experience sale associated. The dataset have the following features: experiencename, bookpartofday, bookweekday, bookday, bookmonth, bookquarter, bookseason, bookcode, airport, pickup, dropoff, daysofstay, arrivalpartofday, departurepartofday, arrivalweekday, departureweekday, leadtime, arrivalday, departureday, arrivalmonth, departuremonth, arrivalquarter, departurequarter, arrivalseason, departureseason, nearestarrivalholiday, nearestdepartureholiday, arrivalflight, departureflight, arrivalpayment, departurepayment, adults, children, babies, cabinluggage, checkedluggage, childbuggyluggage, golfbagsluggage, bikeboxluggage, wheelchairluggage, surfboardluggage, scootersluggage, petscratesluggage, clientlongtime, clientfrequency, clientconcludedtrips.

The sales of private transfer services show a growing trend since 2012 (Figure 1). However, the recession in 2019 has slowed down this ever-increasing trend.

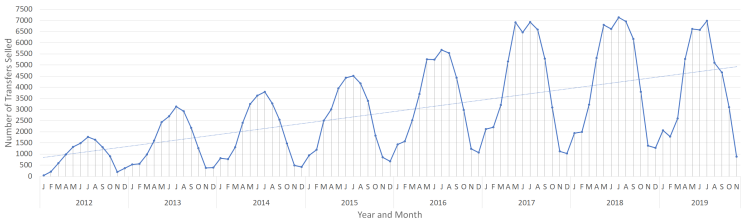


Fig. 1: Transfer service bookings evolution from February 2012 to April 2019

Almost half of the transfer reservations originate from United Kingdom customers. Considering all the bookings, most of the observations are reservations for 2 or 4 adults, with 1 or 2 children and no babies. Reservations associated with a higher number of adults are usually connected to golf players since they have golf luggage and the transfer reservation destiny (mainly Algarve) is a golf courses resort. Regarding airline companies, customers tend to travel using low-cost companies that represent more than 60% of the observations, namely Ryanair, Easyjet, Aer Lingus, and Jet2.com, in this order. The main airport

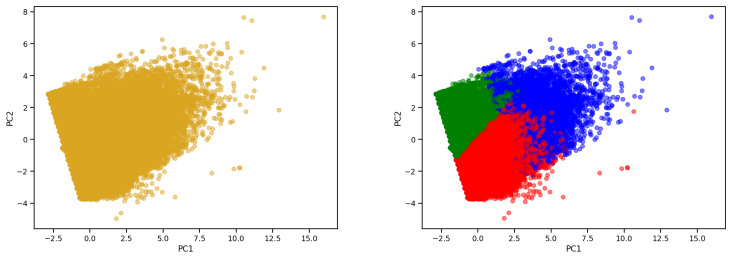
for arrival/departure is Faro airport, located in southeast Algarve and the only major airport in the south of Portugal that also supports southwest Spain.

Regarding experiences booking data, Feel Tour is the experience most sold and motorcycle quad is the vehicle commonly chosen (60% of the cases). Most of the leisure experience services are purchased by United Kingdom and Ireland customers'.

3.3 Customer Segmentation

Considering the data acquired with the transfer services sales, this study aims to distinguish the company's different types of customers. To perform customer segmentation, that is, grouping customers with similar behaviour, demographic data, or interests, PCA (Principal Components Analysis) was employed [15]. The first six components of the PCA were selected since they represent 80% of explained variance ratio to proceed with K-Means technique for segmenting [16].

As seen in Figure 2 (a) plots the first two components showing no significant separation between the data points. To infer how many clusters could exist in data, the Elbow method heuristic was used. The method pointed to $K = 3$ clusters as the best choice and the subsequent K-Means results in PCA data are showed in Figure 2 (b).



(a) (b)

Fig. 2: (a) PCA analysis of two first components; (b) K-Means prediction (K= 3)

Figure 3 illustrates the clusters found by showing different quantitative features that characterize the behaviour and demographic context of customers in transfer service reservations.

From the analysis of the customers in the three groups indicated by PCA and K-means techniques, it was discovered that: the first cluster (in red) represents customers that usually book the service in May, arrive in June and stay six days. These customers often show a lead time of one month and tend to make a new reservation within the next two months. Transfer reservations in this cluster average three adults, no babies or children, and bring three items of baggage. The

second cluster (green) represents customers that book later and with minimum lead time. They usually book and arrive in August with a lead time of eleven days. These customers stay for less time (three days) but buy services with more frequency. These customers travel with fewer pieces of baggage, with an average of two adults without babies or children in their reservations. The last cluster (blue), represents customers that travel with children or babies and bring more varied baggage. These customers usually arrive in July, stay for one week, present a lead time of 23 days and book again in the next two months.

Information about each customers' cluster will be passed as a feature for the dataset used in classification models, in order to help models identify possible experience service buyers.

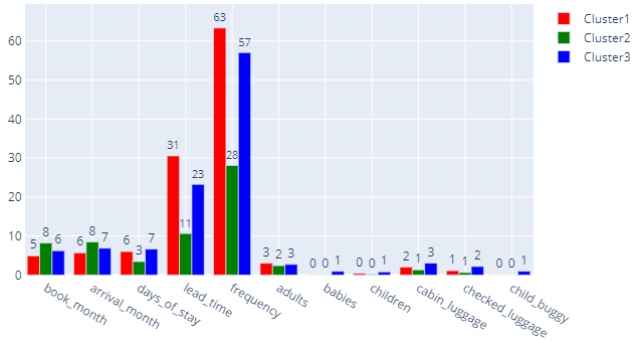


Fig. 3: Customer segmentation of transfer service buyers, divided in three cluster

4 Classification Models

4.1 Data preparation

The dataset used to train the models is divided into four classes. Class 0 represents the transfer service purchase observations with no experience service associated with it. Class 1 is associated with a golf experience service, class 2 is associated with a Feel Tour experience and class 3 with an Experience Tour. Experiences of class 2 and 3 can use different types of vehicles: buggy or quad motorcycle. Since this division was not properly learned by the models that showed consistent underfit, the observations have been grouped by experience independently of the type of the vehicle used, thus reducing the final quantity of target classes, from four to two different classes. Table 1 resumes the distribution of data between the different classes.

A serious unbalance between observations for no purchase and experience purchase can be observed, with 273341 observations for class 0 (no experience

Table 1: Number of observations in dataset per type of service bought.

Experience	Class	Total (original)	Train Set	Test Set
Only Transfer	0	273384	200-10000	82003
Golf	1	172	120	52
Feel Tour	2	162	114	48
Experience Tour	3	93	65	28

bought) and 427 for the remaining classes (transfer plus experience). Therefore, for models to work properly, it was necessary to balance the data. Undersampling over class 0 was performed, selecting a quantity closer to the number of class 1 samples. Since there are many more transfer purchases than experiences being bought, this information was passed implicitly to the model, through a selection of class 0 samples slightly superior to those of class 1, to a minimum of 200 examples comparing to 120 examples from class 1. Undersampling used *RandomUnderSampler* from *imblearn* [17] library, that randomly and uniformly undersampled the majority class to a different number of observations 200, 400, 600, 1000, 2000, 3000, 4000, 5000, 10000.

4.2 Baseline models

First, classification models including all classes at once, were developed to predict the likelihood of a customer who was buying a service transfer would also buy an experience.

The following algorithms were used to train the models: decision tree, random forest, k-nearest neighbours, naive bayes, support vector clustering, ada boost, gradient boosting, linear discriminant analysis, and XGBoost. To construct the models, the open-source Scikit-learn software was used [18]. The dataset was split in 70% for training and 30% for test. Models were trained using a hyperparameter heuristic procedure, GridSearchCV from scikit-learn [18], to identify the best parameters of different algorithms for best performance. The tunning was performed using five cross-validations for hyperparameter optimisation to find the best parameters and estimate the performance of the models on unseen data.

Using different sizes of undersampling for class 0 it was possible to analyse and achieve better performance for the different algorithms used. Table 2 resumes the results for the five best baseline models built. XGBoost (XGB) was the algorithm that performed better. With 200 class 0 observations, the model correctly identified 80 experience observations in a total of 128 observations in the test. The second-best approach was obtained with XGB using 400 observations from class 0 that improved the true negative rate (correctly identified only transfer observations) but reduced the true positive rate (class 1, 2 and 3) with 71 observations corrrectly identified.

Table 2: Best results for baseline multi-class models

Class 0	Algorithm	Accuracy	Recall	Precision	Sensivity	TP	TN
200	XGB	85.7%	68.1%	25.9%	68.1%	80	70319
400	XGB	95.8%	67.4%	32.8%	67.4%	71	78615
400	GB	86.5%	67.1%	25.6%	67.1%	76	70936
400	LDA	89.6%	66.3%	25.8%	66.3%	71	73492
200	LDA	93.6%	66.1%	32.5%	66.1%	72	76764

4.3 Building the Hybrid Models

The unbalanced number of observations contributed to a limited performance, even when using undersampling. As the company's main objective is to enhance the true positive rate and minimize the false-negative rate so that they effectively can recommend experiences services for their transfer customers, a hybrid model was built based on a two-step procedure.

In a first step, a binary model is used to predict if a transfer purchase observation will buy an experience or not. The second step employs a multi-class classification model, trained exclusively with observations that purchased experiences, to classify which type of experience should be recommended for each customer identified in the first step as a potential experience buyer. For performance measure purposes of the hybrid model, observations classified as true positives by the binary model (predicted that transfer customer would purchase experience) are sent to the multi-class model to classify which type of leisure experience service customers will buy, which will then be recommended to the customer.

For the binary classifier, the transfer observations associated with an experience bought were grouped and labelled as class 1 and the remaining observations were labelled as class 0. As previously stated, in a first phase, a binary classification model is used. A binary model usage is essential to identify which patterns may be used to predict which customers can potentially buy additional services. The main objective is to construct a model that can predict if a customer will purchase an experience. The models were tuned to maximize the true positives rates and the sensitivity rate.

In the different binary evaluations, random forest and XGB algorithms presented the best results. Results from the former reveal better sensitivity regarding true positives, 78.9%. XGB, on the other side, showed a lower sensitivity score of 74.2%, with 95 observations correctly identified. However, binary models are not suitable to accurately identify class 0 observations as the baseline models with a lot of false positives. Table 3 summarizes the best results for all the models that were tested in a hybrid context. Besides the high rate of true positives identified in binary models, when these results are passed to multi-

class model (that was trained only with class 1,2 and 3), the true positives for the right experience type drop. For instance, in the first line of Table 3 we can see that the binary model correctly identified 101 transfer services with experience observations. Still, when passed to the multi-class model, it only correctly identified 88 of experience types.

Table 3: Best results for hybrid models

Class 0	Algorithm	Accuracy	Recall	Precision	Sensivity	TP	TN	FP	FN	TPhm
200	RF	65.7%	72.3%	35.7%	78.9%	101	53827	28176	27	88
200	XGB	62.9%	68.5%	33.6%	74.2%	95	51541	30462	33	86
200	NB	64.4%	67.8%	33.5%	71.1%	91	52828	29175	37	76
400	GB	78.3%	73.5%	32.2%	68.8%	88	64224	17779	40	77
200	LDA	59.9%	63.9%	31.0%	68.0%	87	49095	32908	41	78

From a comparison point of view, hybrid models were able to improve performance regarding identifying experience buyer and classifying types of experiences when compared to all-in-one models with non-purchase experience targets and different experience types targets. However, hybrid models have the limitation of classifying a large number of observations has false positives. Using a baseline multi-class model with 200 observations for class 0 and XGB algorithm, it was possible to classify correctly 80 observations of experiences classes (1-3). Compared to a hybrid model, using the same algorithm and the same sample, it correctly classified 95 observations of experiences classes (1-3).

5 Conclusions

This work present an analysis of private transfer company customers' through customer segmentation and a hybrid classification model to recommend new services to transfer customers. It shows that transfer customers are divided into three groups with demographically and behavioural differences. The hybrid model developed improved the sensitivity rates compared to baseline models, but, increased the number of false positives. Correctly identify if a customer will purchase a leisure experience is more important than knowing which type of experience the customer will buy. As there are only three types of experiences and the model indicates willingness to buy, the model can recommend the three types. The results obtained with the hybrid classification model are affected significantly by the fact that there are fewer observations for the experience classes. Cluster labelling for each customer did not affect the models' performance.

For future work, a hybrid recommendation model with different techniques of over and undersampling should be tested.

References

1. Mariani, M., Baggio, R., Fuchs, M., Höepken, W.: Business intelligence and big data in hospitality and tourism: a systematic literature review. International Journal of Contemporary Hospitality Management 30(12), 3514–3554 (2018)
2. Prideaux, B.: Tourism and surface transport. The Sage handbook of tourism management pp. 297–313 (2018)
3. Masiero, L., Zoltan, J.: Tourists intra-destination visits and transport mode: A bivariate probit model. Annals of Tourism Research 43, 529–546 (2013)
4. Loi, L.T.I., So, A.S.I., Lo, I.S., Fong, L.H.N.: Does the quality of tourist shuttles influence revisit intention through destination image and satisfaction? the case of macao. Journal of Hospitality and Tourism Management 32, 115–123 (2017)
5. Hanamanthrao, R., Thejaswini, S.: Real-time clickstream data analytics and visualization. In: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). pp. 2139–2144. IEEE (2017)
6. Lemon, K.N., Verhoef, P.C.: Understanding customer experience throughout the customer journey. Journal of marketing 80(6), 69–96 (2016)
7. Richardson, A.: Using customer journey maps to improve customer experience. Harvard business review 15(1), 2–5 (2010)
8. Zomerdijk, L.G., Voss, C.A.: Service design for experience-centric services. Journal of Service Research 13(1), 67–82 (2010)
9. Hyken, S.: Personalized customer experience increases revenue and loyalty (2017)
10. Abrar, K., Zaman, S., Satti, Z.W.: Impact of online store atmosphere, customized information and customer satisfaction on online repurchase intention. Global Management Journal for Academic & Corporate Studies 7(2), 22–34 (2017)
11. Brito, P.Q., Soares, C., Almeida, S., Monte, A., Byvoet, M.: Customer segmentation in a large database of an online customized fashion business. Robotics and Computer-Integrated Manufacturing 36, 93–100 (2015)
12. Anshari, M., Almunawar, M.N., Lim, S.A., Al-Mudimigh, A.: Customer relationship management and big data enabled: Personalization & customization of services. Applied Computing and Informatics (2018)
13. Weinstein, A.: Customer retention: a usage segmentation and customer value approach. Journal of Targeting, Measurement and Analysis for Marketing 10(3), 259–268 (2002)
14. Sheil, H., Rana, O., Reilly, R.: Predicting purchasing intent: automatic feature learning using recurrent neural networks. arXiv preprint arXiv:1807.08207 (2018)
15. Ding, C., He, X.: K-means clustering via principal component analysis. In: Proceedings of the twenty-first international conference on Machine learning. p. 29 (2004)
16. Afrin, F., Al-Amin, M., Tabassum, M.: Comparative performance of using pca with k-means and fuzzy c means clustering for costumer segmentation. International Journal Of Scientific & Technology Research 4, 70–74 (2015)
17. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research 18(17), 1–5 (2017), http://jmlr.org/papers/v18/16-365.html
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)