# Using Customer Segmentation to Build a Hybrid Recommendation Model

Pedro Camacho[1], Ana de Almeida[1], and Nuno António[1]

[1]ISCTE - Lisbon University Institute, Av. das Forcas Armadas, 1649-026, Portugal,
pafco11@iscte-iul.pt,
WWW home page: https://www.iscte-iul.pt

# Using Customer Segmentation to Build a Hybrid Recommendation Model

Pedro Camacho[1], Ana de Almeida[1], and Nuno António[1]

ISCTE - Lisbon University Institute, Av. das Forças Armadas, 1649-026, Portugal,
pafco11@iscte-iul.pt,
WWW home page: https://www.iscte-iul.pt

**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.
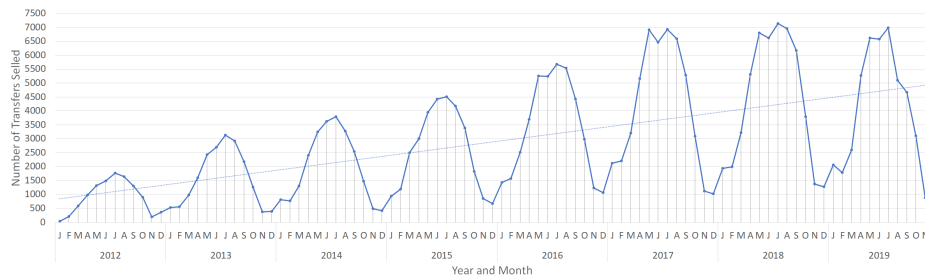


Fig. 1: Transfer service bookings evolution from February 2012 to April 2019

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).



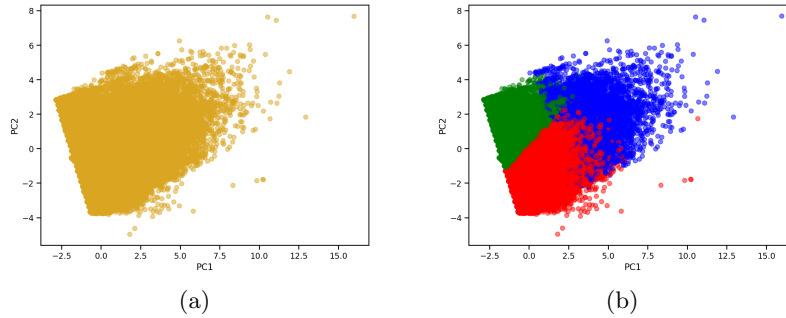(a)                                                        (b)

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.
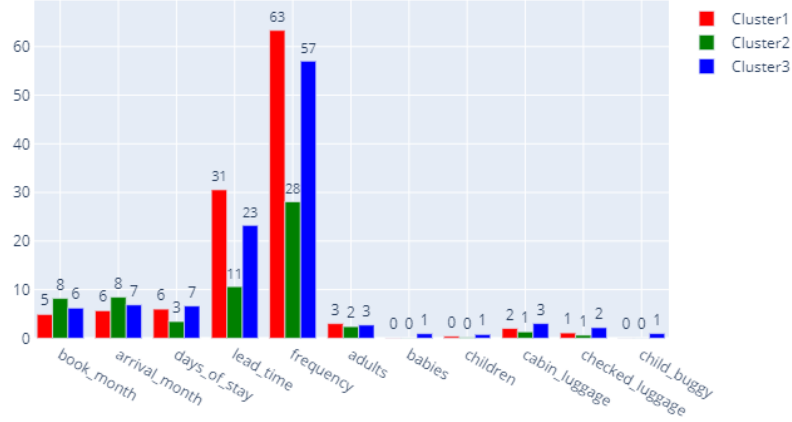


Fig. 3: Customer segmentation of transfer service buyers, divided in three cluster

## 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 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

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 |

*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.

### 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

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 |

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)