



INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Modelos de Previsão e Análise de Dados para otimizar o desempenho de um Grupo de Media

André Filipe Carloto Bernardes

Mestrado em **Sistemas Integrados de Apoio à Decisão**

Orientador:

PhD João Carlos Amaro Ferreira, Professor Associado com
Habilitação,
ISCTE-IUL

Co-Orientador:

Nuno Alexandre Lopes Ramos, Diretor

Julho, 2023

Departamento de Ciência e Tecnologias de Informação

Modelos de Previsão e Análise de Dados para otimizar o desempenho de um Grupo de Media

André Filipe Carloto Bernardes

Mestrado em **Sistemas Integrados de Apoio à Decisão**

Orientador:

PhD João Carlos Amaro Ferreira, Professor Associado com
Habilitação,
ISCTE-IUL

Co-Orientador:

Nuno Alexandre Lopes Ramos, Diretor

Julho, 2023

"Opportunities don't happen. You create them."

Chris Grosser

Agradecimentos

A conclusão desta dissertação dita o encerrar de uma fase da minha vida que representa anos de extrema dedicação tanto em termos profissionais como pessoais, enquanto trabalhador-estudante. A concretização deste objetivo que não seria possível sem os conhecimentos adquiridos ao longo dos últimos dois anos, nem sem as pessoas cujos caminhos se cruzaram com o meu, às quais quero deixar umas palavras de apreço.

Quero agradecer ao meu orientador Professor João Carlos Ferreira por toda a sua orientação indispensável não só na concretização desta dissertação como ao longo do mestrado e por dispensar o seu tempo ao ajudar-me a concluir esta etapa.

Ao meu co-orientador Nuno Alexandre Lopes Ramos um especial obrigado por todo o seu apoio na concretização deste trabalho, bem como todos os conhecimentos e conselhos que partilhou comigo e que me permitiram crescer tanto ao nível profissional como pessoal.

Aos meus amigos que sempre tanto me apoiaram e ajudaram ao longo de todo o meu percurso académico, um grande obrigado.

Quero também agradecer à Global Media Group por me confiar a oportunidade de desenvolver este trabalho.

Por último, quero deixar o meu maior agradecimento à minha família, em especial à minha mãe, ao meu pai e aos meus avós, por todo o vosso apoio, paciência e motivação ao estarem sempre ao meu lado e acreditarem em mim, não só nesta etapa, mas ao longo de toda a minha vida. Acredito genuinamente que sem vocês não conseguiria ter alcançado nada disto sozinho.

André Filipe Carloto Bernardes

Resumo

O aumento de custos no mercado de imprensa tem vindo a crescer de forma exponencial após o surgimento da epidemia Covid-19, provocando um nítido decréscimo nas vendas de jornais e respetiva circulação no mercado português. Com o aumento da competitividade e o decréscimo da procura, algumas publicações passaram a oferecer exclusivamente os seus serviços online, enquanto os restantes têm vindo a desenvolver mecanismos e métricas que permitem prever, de forma objetiva e o mais corretamente possível, a quantidade necessária de exemplares em circulação por forma a, simultaneamente, satisfazer não só as necessidades da procura como também a minimizar o número de sobras diárias.

No âmbito desta nova realidade, desenvolveu-se o presente estudo com dados fornecidos pela Global Media Group, com particular enfoque em três grandes jornais diários portugueses, - Diário de Notícias, Jornal de Notícias e O Jogo, tendo como princípio norteador dois principais objetivos. O primeiro desses objetivos, com recurso a séries temporais, prever as vendas e tiragens de cada um dos jornais, aplicando o algoritmo *Prophet* em *Python*, bem como explorar possíveis padrões, tendências e sazonalidades que possam existir. O segundo consiste na criação de um *dashboard* de apoio à tomada de decisão que se revele capaz de auxiliar no processo visualização e análise de dados referentes a vendas. Por conseguinte, foi adotada uma metodologia CRISP-DM, sendo que os dados disponibilizados são referentes aos anos de 2021 e 2022.

Palavras-Chave: Circulação de jornais, Vendas de jornais, Previsão de séries temporais, Sistemas de apoio à decisão, Business Intelligence.

Abstract

The increase in costs in the press market has been growing exponentially after the outbreak of the Covid-19 epidemic, causing a clear decrease in newspaper sales and circulation in the Portuguese market. With the increase in competition and the decrease in demand, some publications began to offer their services exclusively online, while the others have been developing mechanisms and metrics that allow them to predict, objectively and as accurately as possible, the necessary quantity of copies in circulation in order to simultaneously meet not only the needs of demand but also to minimise the number of daily leftovers.

In the scope of this new reality, this study was developed with data provided by Global Media Group, with particular focus on three major Portuguese daily newspapers, - Diário de Notícias, Jornal de Notícias and O Jogo, having as guiding principle two main objectives. The first of these objectives, using time series, is to predict the sales and circulation of each newspaper, applying the Prophet algorithm in Python, as well as to explore possible patterns, trends and seasonalities that may exist. The second consists in the creation of a decision support dashboard that is able to assist in the process of visualization and analysis of sales data. Therefore, a CRISP-DM methodology was adopted, and the data provided refers to the years 2021 and 2022.

Keywords: Newspaper Circulation, Newspaper Sales, Time Series Forecasting, Decision Support Systems, Business Intelligence.

Índice

Agradecimentos.....	iii
Resumo.....	v
Abstract	vii
Índice.....	ix
Índice de figuras	xi
Índice de tabelas	xiii
Lista de abreviações	xv
Introdução.....	1
1.1. Contexto temático	1
1.2. Objetivos do estudo.....	2
1.3. Abordagem metodológica	3
1.4. Estrutura da dissertação.....	4
Revisão da Literatura	5
2.1. Contexto	5
2.2. Metodologia para revisão de literatura.....	7
2.3. Resultados para revisão de literatura.....	8
2.4. Fluxograma PRISMA.....	8
2.5. Identificação e aplicação de publicações	9
Processo de análise para dados de jornais	15
3.1. Dados e conhecimento do negócio.....	15
3.2. Preparação de dados	16
3.3. Visualização de dados	17
3.4. Prophet – Aplicação	24
3.4.1. Modelo Univariável.....	25
3.4.2. Modelo Multivariável	30
3.4.3. Comparação Modelo Multivariável vs Univariável	34
3.5. Avaliação de resultados.....	37
Implementação do <i>Dashboard</i>	39
4.1. Dados e conhecimento do negócio.....	39
4.2. Dashboard.....	41
4.3. Avaliação de resultados.....	45
Conclusões	47

5.1. Discussão de resultados.....	47
5.2. Limitações e trabalho futuro do estudo	48
Referências Bibliográficas	49

Índice de figuras

Figura 1. Processo CRISP-DM [10].....	3
Figura 2. Fluxograma PRISMA	9
Figura 3. Evolução da publicação dos estudos analisados entre 2001 e 2022.....	10
Figura 4. Peso por tema de cada artigo	10
Figura 5. Evolução de tiragens.....	18
Figura 6. Evolução de vendas.....	19
Figura 7. Evolução de vendas por publicação	19
Figura 8. Evolução de tiragens por publicação	20
Figura 9. Tendência semanal de vendas por publicação	20
Figura 10. Divisão das tiragens em percentagem.....	21
Figura 11. Invasão Russa na Ucrânia	22
Figura 12. Morte da Rainha Isabel II	23
Figura 13. Conquista do campeonato nacional de futebol.....	23
Figura 14. Matriz de correlação	25
Figura 15. Séries temporais de vendas e tiragens do DN	26
Figura 16. Real vs Previsões DN.....	26
Figura 17. Séries temporais de vendas e tiragens do JN	27
Figura 18. Real vs Previsões JN.....	28
Figura 19. Séries temporais de vendas e tiragens do JG	29
Figura 20. Real vs Previsões JG.....	29
Figura 21. Real vs Previsões DN.....	30
Figura 22. Real vs Previsões JN.....	31
Figura 23. Real vs Previsões JG.....	32
Figura 24. Gráficos mensais e semanais com previsão de dezembro para o JN	33
Figura 25. Gráficos mensais e semanais com previsão de dezembro para o DN	33
Figura 26. Gráficos mensais e semanais com previsão de dezembro para o JG	34
Figura 27. Comparação dos modelos de vendas	35
Figura 28. Comparação dos modelos de tiragens.....	36
Figura 29. Circulação Total Paga	42
Figura 30. Vendas em Banca.....	43
Figura 31. Vendas por Zona Geográfica	44
Figura 32. Evolução de assinaturas	45

Índice de tabelas

Tabela 1. Resumo da revisão de literatura classificado por citação	11
Tabela 2. Variáveis dos conjuntos de dados referentes aos jornais em estudo.....	15
Tabela 3. Variáveis do conjunto de dados.....	17
Tabela 4. Métricas de análise para modelos de previsão - vendas	37
Tabela 5. Métricas de análise para modelos de previsão - tiragens.....	37
Tabela 6. Questionário Modelos de Previsão	38
Tabela 7. Ano corrente	39
Tabela 8. Ano homólogo.....	40
Tabela 9. Vendas por distrito	41
Tabela 10. Questionário Dashboard	46

Lista de abreviações

APCT - Associação Portuguesa Para O Controlo De Tiragem E Circulação

GMG – Global Media Group

CRISP-DM - Cross Industry Standard Process for Data Mining

PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RSL - Revisão Sistemática da Literatura

DN - Diário de Notícias

JN – Jornal de Notícias

JG - OJogo

Introdução

1.1. Contexto temático

Os jornais diários têm constituído uma constante básica para o consumo de notícias ao longo de gerações. Contudo, a sua popularidade tem vindo a diminuir nos últimos anos à medida que outras fontes de informação, nomeadamente de natureza digital, eclodiram e se afirmaram no mercado e que pela sua atratividade, em que se conjugam modernidade, velocidade e gratuidade, tornaram os antigos meios de comunicação algo antiquados.

Um dos principais problemas com que os jornais diários se debatem é o declínio do número de leitores. Esta realidade deve-se em parte ao acréscimo dos meios digitais que em muito facilitou a obtenção rápida de notícias e de informação em geral. Como resultado, os jornais tiveram de se adaptar, contudo esta transição tem-se revelado difícil para muitos dos títulos outrora pujantes, uma vez que estes têm lutado para rentabilizar o seu conteúdo online, procurando recolher o maior número possível de assinantes, da mesma forma que o fizeram com as edições impressas.

O preço das publicações diárias é outro problema a equacionar. Os jornais tiveram de aumentar os seus preços de assinatura e as taxas de publicidade devido ao aumento dos custos do papel, da tinta e da distribuição. A este ajuste, veio somar-se a retração do público consumidor que passou a achar difícil sustentar o custo de uma assinatura diária de um jornal, quando poderia obter informação de forma mais célere e gratuita.

De acordo com [1] aponta-se o ano de 2029 como uma data hipotética para o fim das vendas de jornais em papel no mercado português, previsão esta com base nos resultados apresentados pela Associação Portuguesa Para O Controlo De Tiragem E Circulação (APCT) [2] referente ao período 2008-2017, data essa que poderá ter sido antecipada ou não pelos efeitos e constrangimentos provocados pela pandemia Covid-19.

Em face de todos os constrangimentos anteriormente enunciados, os jornais devem-se focar em encontrar métodos para se adaptarem ao ambiente em mudança e oferecer aos leitores valor de formas novas e criativas, se optarem pela sobrevivência e crescimento na era digital, de uma forma economicamente viável.

Este estudo pretende, portanto, contribuir para essa adaptação ao melhorar o planeamento e análise estratégica, tanto a médio como a longo prazo, minimizando custos de produção e de distribuição de jornais, ao oferecer uma forma que permite prever, corretamente, a quantidade necessária de jornais para um determinado dia, satisfazendo as necessidades dos agentes de venda e minimizando o número de sobras diária por edição.

Como referido em [3], as despesas para o negócio tenderão, indubitavelmente, a aumentar à medida que aumentarem os jornais não vendidos. Por esta razão, a empresa necessita de um algoritmo eficaz que permita prever o volume de vendas.

Além disso, os autores anteriormente referidos defendem, igualmente, que a quantidade de jornais enviados para determinado ponto de venda deve ser estimada com a menor quantidade de imprecisão, de forma a não resultar no esgotamento do stock para aquele determinado dia, o que sugeriria que se poderiam ter vendido mais jornais e, conseqüentemente, ter obtido um maior lucro.

Após a fase de determinação do modelo melhor habilitado a fazer previsões corretas das quantidades, pretende-se oferecer uma perspetiva de visualização global e rápida com um *dashboard*, através da ferramenta *Power BI*, que permite auxiliar no processo de tomada de decisão ao identificar as principais zonas de venda por cada título, bem como possíveis padrões e sazonalidades associados às vendas.

1.2. Objetivos do estudo

O presente tema de investigação foi inicialmente proposto pela Global Media Group (GMG) [4], com os pressupostos de minimizar custos de produção e distribuição, e de oferecer uma perspetiva de análise de dados, bem como de visualização passível de auxiliar no processo de tomada de decisão.

Desta forma, o estudo propõe-se responder aos seguintes objetivos

1. Construir modelos de previsões de vendas e de tiragens de jornais diários, de forma a minimizar custos de produção e de distribuição de jornais, utilizando uma abordagem de aprendizagem automática através de séries temporais
2. Desenvolver um *dashboard* de apoio à decisão que permita analisar tendências de evolução temporal e geográfica de vendas de diferentes jornais e revistas, para avaliar e tomar decisões, utilizando uma abordagem de *Business Intelligence* e as melhores práticas de visualização de informação

Para responder, de forma eficaz e correta aos objetivos enunciados anteriormente, foi desenvolvido numa primeira fase uma análise e visualização aos dados fornecidos, de forma a extrair conhecimento através de padrões e sazonalidades, para melhor compreender todo o comportamento de vendas. Ora, após esta análise, foi iniciada a criação e aplicação de um algoritmo de previsão, através da implementação do *Prophet* [5] em *Python*[6].

Na segunda fase deste estudo, foi iniciada a construção de um *dashboard* através da ferramenta *Power BI* [7], de forma a ser útil tanto no processo de tomada de decisões como no planeamento e análise estratégica a curto, médio e longo prazo.

1.3. Abordagem metodológica

Para este estudo foi aplicada a metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM) [8], que consiste em seis fases distintas como descrito por [9], e exemplificado na Figura 1.

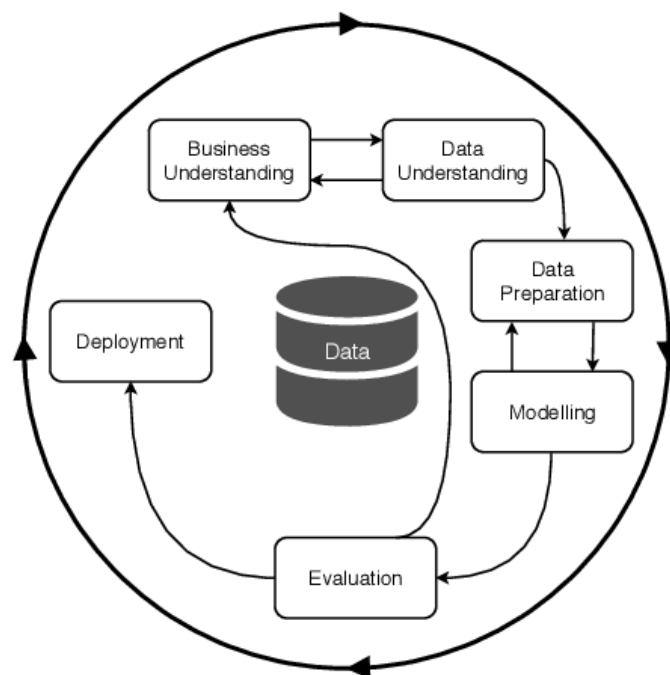


Figura 1. Processo CRISP-DM [10]

Iniciando pela fase de “*Business Understanding*”, a mesma consistiu em elaborar, previamente, isto é, antes de se iniciar o estudo, uma compreensão geral dos requisitos necessários, de forma a estabelecer objetivos claros e exequíveis.

Numa segunda fase, a da “*Data Understanding*”, procedeu-se à recolha, exploração e descrição dos dados, seguida da “*Data Preparation*” que consistiu em identificar os conjuntos de dados fornecidos pela GMG, para limpar e formatar os mesmos.

A fase de “*Data Modelling*” consistiu, essencialmente, no desenvolvimento e comparação de modelos com algoritmo *Prophet*, bem como a criação de um *dashboard* interativo, progredindo assim para a fase de “*Evaluation*” onde foi realizada uma avaliação geral e onde foram determinados os passos seguintes com base nos resultados obtidos.

Por fim, na última etapa, a do “*Deployment*”, foram apresentados os modelos e resultados já finalizados, com um breve resumo do estudo e um planeamento da sua implementação e monitorização, bem como um breve inquérito aos responsáveis da empresa relativamente à concretização deste projeto.

1.4. Estrutura da dissertação

Esta dissertação, cuja estrutura é apresentada infra, juntamente com uma breve descrição de cada um dos capítulos, encontra-se dividida em cinco capítulos.

No capítulo 1, é feita uma introdução ao estudo, e nela se referem os objetivos que se propõe cumprir, bem como a abordagem metodológica.

Quanto ao capítulo 2 apresenta uma descrição das técnicas de investigação utilizadas para a revisão bibliográfica e os seus respetivos tópicos de investigação.

O capítulo 3 destina-se à apresentação, através da metodologia CRISP-DM, de todo o trabalho referente à análise dos dados, construção e avaliação dos modelos aos quais foram aplicados o algoritmo *Prophet*, integrando ainda um questionário a elementos da empresa para avaliação do trabalho realizado.

No capítulo 4 é exposta a criação e implementação do *dashboard*, com base nos requisitos impostos e acordados com a empresa, integrando ainda um questionário a elementos da empresa para avaliação do trabalho realizado.

Por último, o capítulo 5 aborda as conclusões da dissertação, discutindo-se os resultados obtidos. Dele constam, também, todas as limitações e recomendações para trabalho a realizar no futuro.

Revisão da Literatura

2.1. Contexto

Atualmente, alguns jornais começaram a prestar os seus serviços somente em formato *online*, devido ao crescente aumento de custos no mercado e à competitividade crescente no setor da imprensa.

Desta forma, é essencial estimar a quantidade necessária de publicações a serem disponibilizadas nos pontos de venda através de métodos e técnicas analíticas adequadas, bem como analisar tendências de evolução temporal e geográfica.

A maioria dos artigos analisados referentes à previsão de vendas de jornais, todos eles de autores estrangeiros que se debruçaram sobre o problema, mostra não só diferentes contextos, bem como diferentes abordagens e soluções para o problema consoante a sua proveniência. Assim, será apresentada a forma como a comunidade científica tem estudado o tema, através de abordagens tanto analíticas, como de previsibilidade.

Relativamente ao estudo [11] revela elementos e variáveis importantes, como por exemplo a utilização de condições meteorológicas para uma melhor previsão de resultados, e a sua interrelação com as vendas, a fim de compreender, através de uma forma estruturada, o comportamento e a sua influência no volume de vendas de jornais, de forma a minimizar as sobras diárias. Tomando como exemplo a Turquia e o consumo de informação impressa, constata-se que algumas dessas variáveis têm em linha de conta, por exemplo, as condições meteorológicas, assim como o dia da semana para que remete a análise e que influenciam a compra de jornais nas cidades analisadas.

Quanto a estudos específicos focados no mercado português, e devido à especificidade do tema, revelou-se bastante difícil analisar artigos relevantes para a dissertação. Como tal, somente o estudo realizado por Quintanilha [1] em 2018, onde, através de uma revisão crítica, se explora a forma como o aparecimento dos *media* digitais teve um enorme impacto no sector dos jornais em Portugal e se identifica as dificuldades com que os jornais impressos se confrontam e os benefícios dos jornais digitais, prevendo-se o desaparecimento dos jornais impressos em Portugal até 2029.

Akbaş et al. [3] identifica o processo de previsão de jornais como o grande problema do seu estudo, propondo a automatização do processo de previsão de vendas, e a sua resolução computacional, preterindo o método manual, e recorrendo a uma abordagem de previsão com

os testes de Holt-Winter's, Regressão linear, Fuzzy C-Means e Adaptive Neuro Fuzzy inference system.

O artigo [12] aborda novamente o problema na Turquia, cruzando dados históricos de vendas disponibilizados por uma das maiores empresas de distribuição da Turquia e informação meteorológica para prever vendas futuras. O sistema foi considerado preciso e eficaz na previsão de vendas para uma variedade significativa de jornais e regiões, podendo ser utilizado por empresas de jornais a fim de otimizar a sua cadeia de abastecimento, reduzir o desperdício e a ineficiência, e melhorar a satisfação do cliente.

Kovačević et al. [13] focaram-se em desenvolver um processo simplificado de distribuição e venda de jornais estrangeiros durante a época turística, numa região da Croácia, com o objetivo de formular modelos de operadores destinados a lidar com o controlo da distribuição e venda de jornais através de um modelo de *software* e *hardware*.

No artigo [14] os autores optaram por uma abordagem em que se recorreu a redes neuronais artificiais para prever as vendas de jornais, bem como tendências futuras diárias, de forma a otimizar inventários e cadeias de distribuição. A rede neural foi constituída por três camadas e treinada com dados referentes a um ano de vendas.

Bakker et al. [15] abordam, ao longo do artigo, três casos de estudo distintos, sendo que o que foi analisado e considerado relevante para esta dissertação refere-se à base de dados do número de vendas de um dos principais jornais holandeses. Os autores utilizam *task clustering* com base em semelhanças e dependências de variáveis, com o intuito de melhorar o desempenho de algoritmos através de uma abordagem de aprendizagem multitarefa.

No estudo [16] é proposta a utilização de redes neuronais para aprender as relações entre vários fatores que afetam as vendas de jornais, tais como o preço, o clima e a publicidade. A rede neural é então combinada com um modelo Bayesiano, permitindo assim gerar previsões de vendas para diferentes cenários. Os resultados mostram uma melhoria face aos métodos estatísticos utilizados anteriormente, aperfeiçoando-se, assim, a precisão de venda de jornais, bem como a identificação dos fatores mais importantes que afetam as vendas, e que podem ajudar as empresas jornalísticas a tomar decisões mais informadas.

Ainda na utilização de redes neuronais para a resolução do problema de previsões de vendas para jornais, o estudo [17] adota uma metodologia baseada na utilização de um exemplo prático de um jornal alemão para avaliação do modelo, através de uma abordagem para optimização de etapas de redes neuronais, melhorando assim a precisão dos modelos criados.

Prosseguindo o rastreio e passando aos artigos referentes à implementação e visualização de dados, a escolha dos mesmos para análise nesta dissertação revelou-se uma tarefa inegavelmente árdua devido à similaridade dos dados, às técnicas de *Business Intelligence* utilizadas para trabalhar a especificidade dos dados fornecidos a que se acrescenta a necessidade de cumprir os requisitos acordados com a GMG.

O artigo [18] destaca, ainda que de um modo geral, a importância da *Business Intelligence* para a gestão de dados de vendas para uma empresa e os benefícios que podem ser obtidos através da utilização de visualizações interativas, ao proceder-se à sua implementação recorrendo a técnicas e ferramentas de *Business Intelligence*, de forma a melhorar e agilizar a tomada de decisões.

Sharaf Addin et al. [19] implementaram um *dashboard* interativo que permite visualizar agregados com base na demografia, no comportamento e nas características de uma determinada região, utilizando uma abordagem de aprendizagem automática, o que permitiu identificar segmentos de clientes.

No estudo [20] é proposto um processo de implementação de um sistema de *Data Warehouse* e a construção de um *dashboard* de *Business Intelligence* que visa monitorizar o desempenho das vendas para uma empresa.

O artigo [21] discute os desafios enfrentados por uma empresa na gestão e análise dos seus dados de vendas, em que o sistema de *Business Intelligence* auxiliou a ultrapassar estes desafios, permitindo à empresa adquirir conhecimentos sobre os seus dados de vendas em tempo real e tomar decisões orientadas por esses mesmos dados de forma a melhorar o desempenho das suas vendas, através da implementação de um *dashboard* de visualização interativa.

Purnamasari et al. [22] propõem um armazém de dados que pode ser utilizado para construir um negócio inteligente com a utilização de informação obtida a partir do histórico de transacções, permitindo uma análise de dados apresentada através de um painel de controlo utilizando o *Power BI*.

2.2. Metodologia para revisão de literatura

De forma a realizar uma Revisão Sistemática da Literatura (RSL), através de uma pesquisa bem estruturada, bem como responder aos objetivos definidos anteriormente na secção 1.2, foi usada a metodologia PRISMA [23]. Como resultado, a metodologia PRISMA foi utilizada tendo em conta a lista de verificação de vinte e sete itens, como vai ser descrito na figura 2, onde está

representado o fluxograma do PRISMA que inclui itens considerados cruciais para a explicação clara de uma revisão sistemática, e um fluxograma de três fases.

Este estudo foi desenvolvido ao longo do mês de janeiro de 2023, e abrangeu artigos publicados durante os anos de 2001 a 2022.

2.3. Resultados para revisão de literatura

A pesquisa bibliográfica iniciou-se pela definição de palavras-chave, com base nos objetivos previamente definidos, que originaram as *queries* de pesquisa utilizadas no repositório SciVerse Scopus. De seguida, foi feita a criação das duas *queries* que melhores resultados de pesquisa alcançaram, podendo assim abranger o máximo possível de estudos para seleção nesta fase inicial, obtendo-se as seguintes:

1. (("Time series forecasting" OR "Newspaper circulation" OR "Circulation forecasting" OR "Forecasting" OR "Newspapers") AND ("Newspaper sales")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "ch") OR LIMIT-TO (DOCTYPE , "cp"))
2. ("Dashboards") AND ("Sales") AND ("Analysis")) AND (LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cr")) AND (LIMIT-TO (PUBYEAR , 2023) OR LIMIT-TO (PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019))

2.4. Fluxograma PRISMA

Para ser possível realizar uma análise explicativa e objetiva na RSL, é representado o fluxograma da metodologia PRISMA, constituído por três fases, processo que se inicia pela sua identificação, com base nos artigos encontrados no repositório Scopus e nas *queries* de pesquisa utilizadas, de onde sobressaíram 75 publicações, 32 relativas à primeira *query* e 43 relativas à segunda.

Dessas 75 publicações iniciais, 23 foram excluídas por não estarem relacionadas com o tema desta dissertação. Continuando a investigação referente à elegibilidade dos 52 artigos, 21 deles foram excluídos após leitura do resumo e 17 após uma leitura completa do artigo.

Assim, após a aplicação desta metodologia, somente 14 artigos vão ser incluídos neste estudo, como representado na Figura 2.

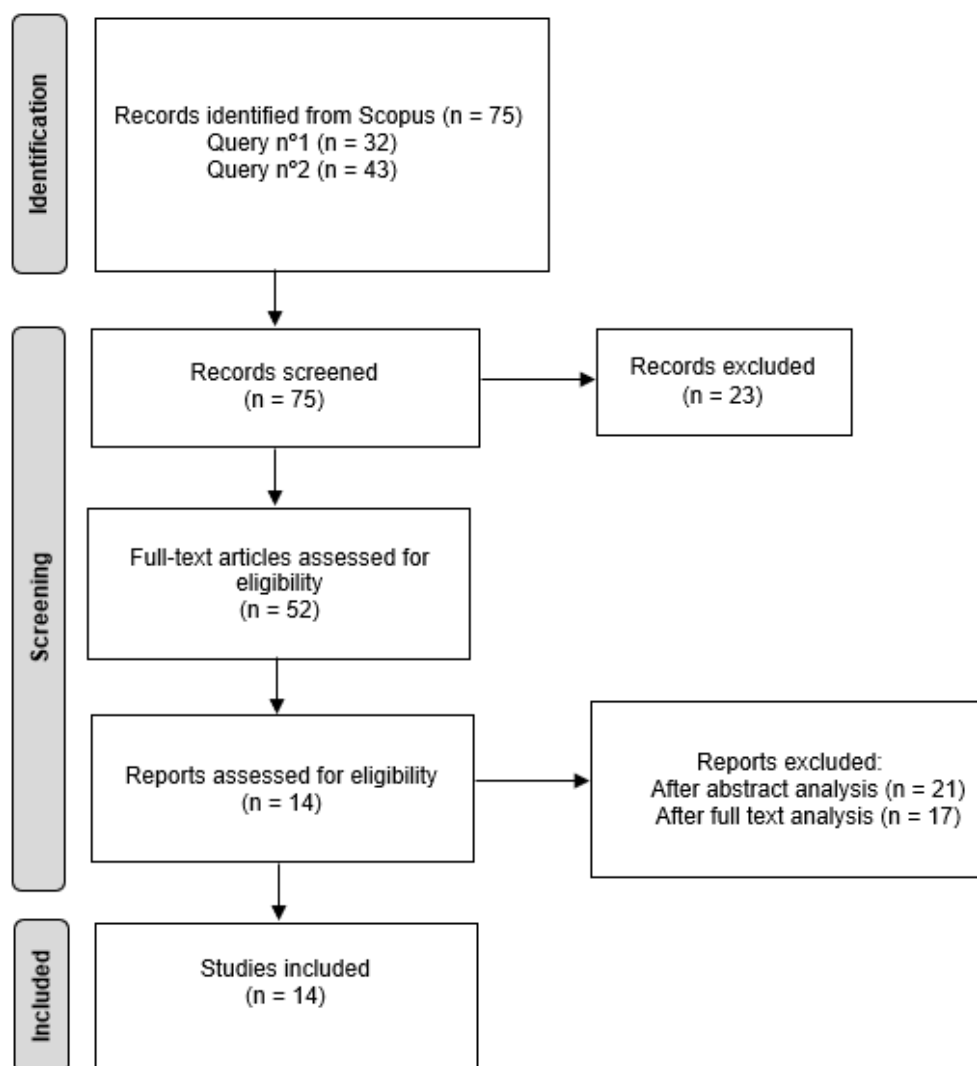


Figura 2. Fluxograma PRISMA

2.5. Identificação e aplicação de publicações

Após completar-se a análise dos 14 artigos previamente elegíveis para o estudo, os mesmos foram armazenados utilizando o software Mendeley, tendo-se aplicado, em seguida, o Microsoft Excel, para um tratamento inicial dos dados, com especial incidência nos títulos, autores, palavras-chave e resumos, como forma de aprofundamento do modo como o tema tem sido abordado ao longo do tempo. Desde logo, se destacou uma linha de tendência cada vez mais crescente quanto aos artigos publicados pelos diversos autores, o que remete para um aumento da sua importância nos últimos anos, representado na Figura 3.

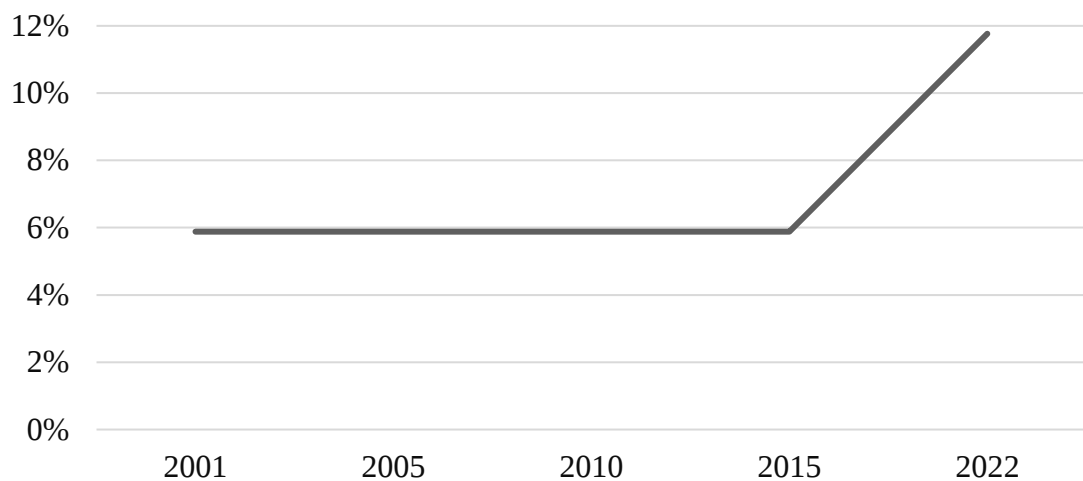


Figura 3. Evolução da publicação dos estudos analisados entre 2001 e 2022

Nesta etapa, e numa fase posterior à determinação de um processo manual para estabelecer os temas gerais abordados por cada um dos artigos, foi representada a sua distribuição através do seguinte gráfico, de forma a providenciar uma visualização global dos temas mais relevantes mencionados nos estudos. Como podemos observar, temos uma distribuição bastante equitativa, o que sugere que os tópicos se relacionam entre si e abrangem, de uma maneira geral, os objetivos de estudo abordados nesta dissertação, como se pode observar pela Figura 4.

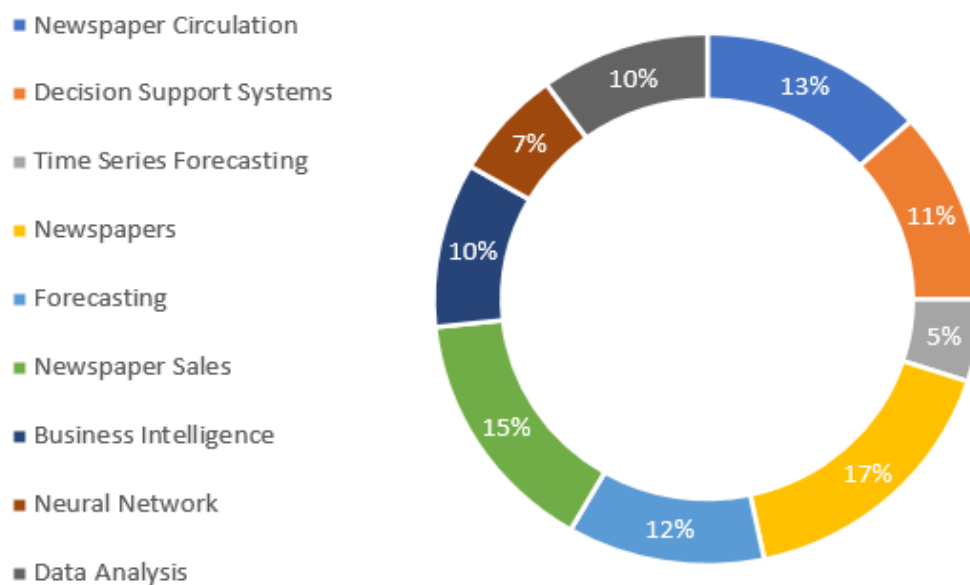


Figura 4. Peso por tema de cada artigo

Em seguida, e com base nos resultados obtidos na secção referente à revisão de literatura deste estudo, procedeu-se à elaboração da Tabela 1, que apresenta a devida identificação do número de citações referentes a cada um dos artigos analisados, ordenados por ordem decrescente, permitindo, assim, identificar a sua influência e notoriedade no meio científico. Acrescentou-se uma breve síntese acerca do método utilizado por cada um dos autores e a respetiva aplicação nos seus estudos.

Tabela 1. Resumo da revisão de literatura classificado por citação

Citações	Título	Autores	Método	Aplicação
394	Task clustering and gating for bayesian multitask learning	Bakker B. Heskes T.	Task Clustering; Task Gating; Multitask learning algorithms	Task Clustering e Task Gating para melhorar o desempenho de algoritmos
6	Newspaper vendor sales prediction using artificial neural networks	Fakharudin A.S. Hamza M.A.M. Johan M.U.	Foram utilizados três métodos para desenvolver o modelo ANN: Back Propagation Neural Network (BPNN), Radial Basis Function Neural Network (RBFNN), and Adaptive Neuro Fuzzy Inference System (ANFIS)	Previsão de vendas de jornais, através de um modelo de redes neuronais (ANN)
5	Implementation of Business Intelligence for Sales Data Management Using Interactive Dashboard Visualization in XYZ Stores	Akbar R. Silvana M. Hersyah M.H. Jannah M.	Implementação de ferramentas e técnicas de Business Intelligence e de um <i>dashboard</i> interativo	Implementação de uma aplicação de Business Intelligence, utilizando um <i>dashboard</i> de visualização interativa, de forma a produzir relatórios que podem ser utilizados pelos gestores da loja para tomar melhores decisões.

4	Sales forecasting system for newspaper distribution companies in Turkey	Incesu G. Asikgil B. Tez M.	Exponential smoothing e ARIMA	Criação de um sistema de previsão de vendas para auxiliar as organizações turcas de distribuição de jornais a planear e a tomar melhores decisões
3	Optimising newspaper sales using neural-Bayesian technology	Heskes T. Spanjers J.-J. Bakker B. Wiegerinck W.	Redes Neurais; Modelos Bayesianos	Desenvolvimento de um software, através da utilização de redes neurais para aprender as relações entre vários factores que afectam as vendas de jornais, combinado com um modelo Bayesiano para prever vendas em diferentes cenários
1	2029 - The end of print newspapers in Portugal? A longitudinal study on the main performance indicators in Portugal's traditional print-newspaper industry	Quintanilha T.L.	Revisão sistemática, através de revistas e documentos de conferências, com o intuito de desenvolver uma investigação sobre o fim de jornais impressos em Portugal	Revisão crítica da investigação, bem como conclusões e as suas implicações para determinar um possível desaparecimento de jornais impressos em Portugal até 2029
1	Business Intelligent in an E-Commerce Industry	Purnamasari A.M. Pah C.E.A. Yoga M.D.I. Girsang A.S. Isa S.M.	Metodologia de Kimball; Implementação de ferramentas de Business Intelligence e de um <i>dashboard</i>	Implementação de um sistema de Data Warehouse e construção um <i>dashboard</i> de Business Intelligence para análise de dados referentes ao negócio

	Realising Newspaper Sales by Using Statistic Methods	Dogan O. Gurcan O.F.	One-way Anova; Two-way Anova; t-test; Levene's, Kolmogorov-Smirnov; Kruskal-Wallis tests	Identificar variáveis relevantes na influência de vendas de jornais, bem como fornecer uma análise estruturada do comportamento de vendas ao longo do tempo
	Predicting newspaper sales by using data mining techniques	Akbaş B. Karahoca D. Karahoca A. Güngör A.	Holt-Winter's, Linear Regression e Fuzzy C-Means (FCM)+ Adaptive Neuro Fuzzy inference system (ANFIS)	Automatizar o processo de previsão de vendas, e resolvê-lo computacionalmente em vez de manualmente
	Modeling human operator and process in different environments	Kovačević D. Poša A. Pribacić N. Antonić R. Kovačević A.	Criação de um software especificamente adaptado ao caso em estudo	Criação um processo simplificado de distribuição e venda de jornais estrangeiros durante a época turística numa região da Croácia
	Building committees by clustering models based on pairwise similarity values	Ragg T.	Optimização de etapas de redes neurais	Desenvolvimento de uma abordagem que pretende otimizar etapas de redes neurais para dados limitados
	The Implementation of Business Intelligence on Visualisation of Transaction Data Analysis using Dashboard System Case Study: XYZ Convenience Store	Akbar R. Wahyudi Rahmadoni J. Lubis H.A.	Implementação de ferramentas de Business Intelligence e de um dashboard interativo	Implementação com o intuito de melhorar a precisão dos dados, agilizar a tomada de decisões e ganhar vantagem competitiva no mercado

	Customer Mobile Behavioral Segmentation and Analysis in Telecom Using Machine Learning	Sharaf Addin E.H. Admodisastro N. Mohd Ashri S.N.S. Kamaruddin A. Chong Y.C.	Aprendizagem Automática; Implementação de ferramentas de Business Intelligence e de um <i>dashboard</i> interativo	Visa fazer uma segmentação de clientes utilizando aprendizagem automática e a implementação de um <i>dashboard</i> interativo que permita visualizar agregados com base na demografia, comportamento e características da região
	Implementation of Data Warehouse in Making Business Intelligence Dashboard Development Using PostgreSQL Database and Kimball Lifecycle Method	Ramadhani P.P. Hadi S. Rosadi R.	Método Kimball Lifecycle; Implementação de ferramentas de Business Intelligence e de um <i>dashboard</i> interativo	Conceber um sistema de Data Warehouse e construir um <i>dashboard</i> de Business Intelligence que visa monitorizar o desempenho das vendas e também a entrega de mercadorias.

Processo de análise para dados de jornais

3.1. Dados e conhecimento do negócio

Esta fase visa compreender as exigências e os objetivos do estudo através das necessidades empresariais e delineando objetivos para cumprir. Desta forma realizou-se em conjunto com a GMG uma reunião inicial onde foi apresentado o tema desta dissertação, bem como os seus desafios, permitindo uma maior compreensão tanto dos dados como de aspetos específicos e fundamentais do negócio para os resultados obtidos.

Como referido anteriormente, esta etapa do estudo foca-se com recurso a séries temporais, prever as vendas e tiragens de cada um dos jornais em questão, aplicando o algoritmo *Prophet*, bem como explorar possíveis padrões, tendências e sazonalidades que possam existir, bem como uma maior extração de conhecimento possível para beneficiar o conhecimento do negócio.

Foram fornecidos por parte do departamento de circulação da GMG, três ficheiros em *Excel* que abrangiam dois anos completos de dados internos referentes a 2021 e 2022, com a mesma matriz estrutural para cada um dos jornais em estudo, apresentando bastantes limitações ao nível de organização e variedade de dados, com somente 9 colunas e 2352 linhas, representado através da Tabela 2.

Tabela 2. Variáveis dos conjuntos de dados referentes aos jornais em estudo

Variável	Descrição
Edição	Edição atribuída ao jornal
Data	Data
Expedição	Representa a tiragem total de uma determinada edição
Vnd Apurada	Representa as vendas de uma determinada edição
Vendas diretas	Representa as vendas diretas de uma determinada edição
Sobras	Representa as sobras de uma determinada edição
Permutas	Representa as permutas de uma determinada edição
Ofertas	Representa as ofertas de uma determinada edição
Título	Título da edição

3.2. Preparação de dados

A preparação e fusão dos dados fornecidos, bem como o seu pré-processamento foram realizados em *Python*, utilizando as bibliotecas *Numpy*[24], *Pandas*[25] e *Seaborn*[26].

Iniciando a recolha, exploração e verificação da qualidade dos dados, conclui-se que a empresa não apresentava uma base de dados que permitia uma rápida extração e acesso às informações relevantes ao estudo, como tal, foi feito a fusão relativamente aos três ficheiros em apenas uma base de dados para facilitar o seu tratamento e análises.

Desta forma, foi necessário numa primeira fase reformular a estrutura da base de dados obtida através da eliminação de variáveis, nomeadamente Edição, Vendas diretas, Permutas, Ofertas e Título, sendo que as referidas não acrescentavam valor para análise, bem como apresentavam um elevado número de valores nulos.

Referentemente à aplicação de métodos para limpeza de dados para a nova base de dados, foram aplicados os seguintes:

- Remoção de valores nulos das variáveis em estudo, uma vez que representavam uma pequena percentagem no conjunto de dados;
- Criação de uma nova coluna intitulada como Dia_semana, que permite classificar o dia da semana a que se refere na coluna Data, de forma a facilitar uma futura visualização e agrupamento de dados;
- Criação de uma nova coluna intitulada como Feriado que classificou todos os feriados referentes ao longo do ano, bem como a classificação de três acontecimentos relevantes no ano 2022 que originaram em um aumento da procura de jornais;
- Aplicação da função *unique* em *Python* para verificar se na coluna Publicação apenas se encontravam os títulos dos jornais considerados necessários para o estudo. Ora, após a sua aplicação, verificou-se que a coluna integrava 29 publicações que não deveriam constar do conjunto de dados. Como tal, procedeu-se à sua eliminação imediata;

A fase de pré-processamento dos dados permitiu simplificar e agregar a informação, tendo sido identificadas as principais variáveis necessárias e úteis para o estudo, o que resultou num único conjunto de dados que contém sete colunas e 2190 linhas com registos.

Esta etapa permitiu corrigir erros de inserção de informações, bem como fazer uma limpeza prévia e uma organização estrutural aos dados e ainda criar uma base de dados para apoio à tomada de decisão por parte da GMG.

Denotou-se ainda a ausência de variáveis que impossibilitavam uma maior abrangência para a aplicação de métodos de extração de conhecimento e de análise aos dados, devido à débil estrutura de dados fornecida. A título de exemplo, variáveis referentes a condições meteorológicas poderiam ser bastante interessantes de estudar, tal como foi abordado no estudo [11].

Desta forma, com base nas alterações aplicadas aos dados iniciais fornecidos, foi elaborada a Tabela 3, que permite simplificar a identificação e a classificação das variáveis na base de dados, bem como uma breve descrição de cada uma.

Tabela 3. Variáveis do conjunto de dados

Variável	Classificação	Descrição
Data	Númerica Discreta	Data
Dia da semana	Categórica Nominal	Categoriza o dia da semana
Publicação	Categórica Nominal	Categoriza a publicação jornalística
Expedição	Númerica Discreta	Representa a tiragem total de uma determinada publicação e respetiva data
Venda Apurada	Númerica Discreta	Representa as vendas de uma determinada publicação e respetiva data
Sobras	Númerica Discreta	Representa as sobras de uma determinada publicação e respetiva data
Feriado	Categórica Nominal	Categoriza, caso seja, o feriado associado ao dia

3.3. Visualização de dados

A modelação inicial dos dados foi executada através de bibliotecas *Python*, tais como *Numpy*, *Pandas* e *Seaborn* para a análise estatística. Para visualização gráfica dos dados foram utilizados o *Matplotlib* e o *PowerBI*.

Tendo este trabalho como principal objetivo a previsão de vendas e de tiragens para cada publicação, comecei, primeiramente, por verificar o comportamento de cada uma entre os anos de 2021 e 2022.

Neste sentido, procurei, numa primeira fase, compreender a forma como as tiragens dos jornais se comportaram ao longo do tempo. Através da Figura 5, podemos verificar que, de um modo geral, foram diminuindo gradualmente, apresentando uma sazonalidade nos meses referentes aos primeiros meses do ano e nos meses de verão, o que se reflete numa quebra na produção de jornais a rondar os 9,05% de 2021 para 2022.

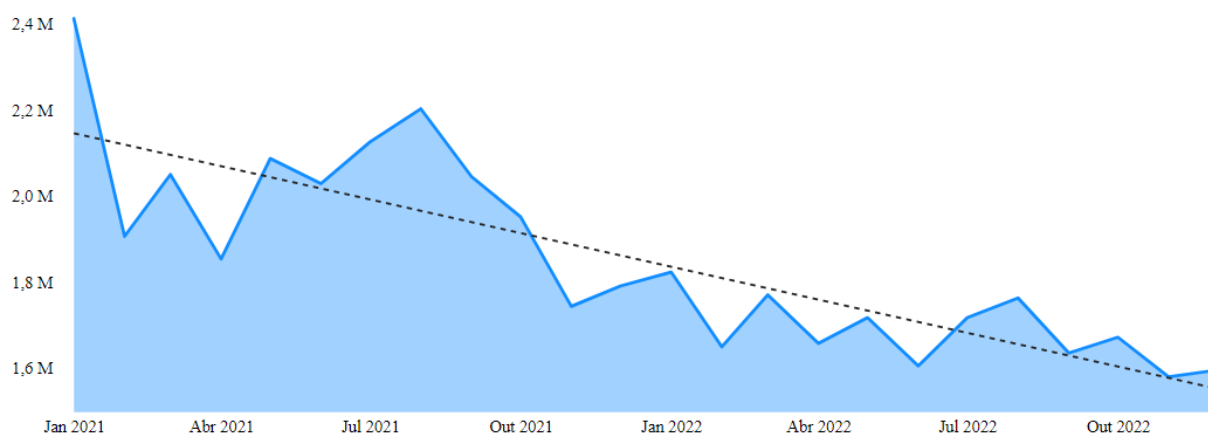


Figura 5. Evolução de tiragens

De seguida, repetiu-se a mesma visualização para as vendas de jornais de forma a poder comparar a diferença entre os exemplares produzidos e os vendidos, sendo que, de um modo geral, apresentaram idêntica sazonalidade, e uma quebra ligeiramente inferior a rondar os 7,11% de 2021 para 2022.

Esta descida contínua pode ser justificada em parte devido a uma contínua transição do jornal em papel para o digital e da crescente mutação dos hábitos de leitura, bem com uma influência de um período de crise mundial, mais especificamente a Covid-19, sendo que, no início de 2021, foram impostas várias restrições, que, de certa forma, não eram as mais vantajosas no sentido de influenciar a procura de jornais por parte dos leitores, como se pode observar pela Figura 6.

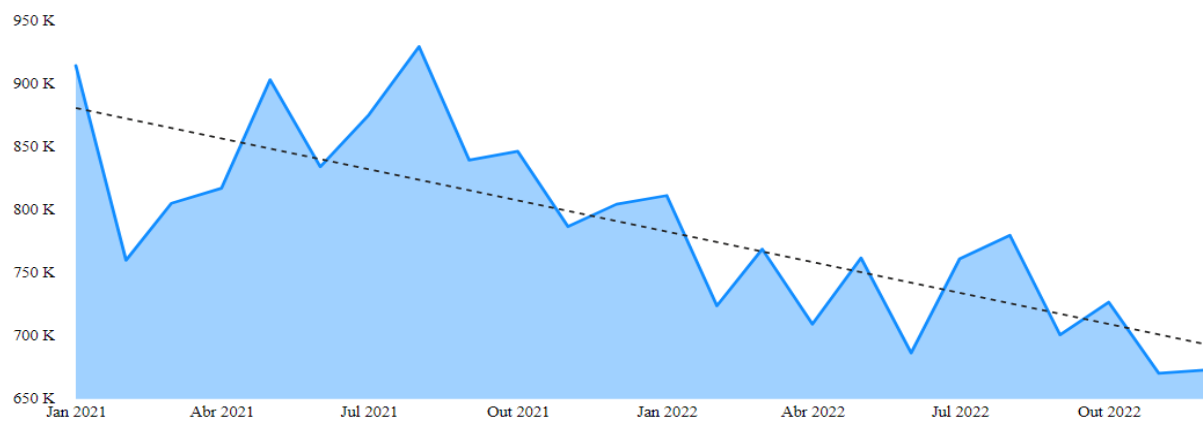


Figura 6. Evolução de vendas

Indo um pouco mais ao detalhe, foi necessário perceber a variação tanto das vendas como das tiragens consoante cada publicação, verificando-se que, e a partir da análise das Figuras 7 e 8, existe uma grande discrepância entre as três publicações, substancialmente notável no JN, onde foi notada a maior quebra, bem como o maior volume de vendas.

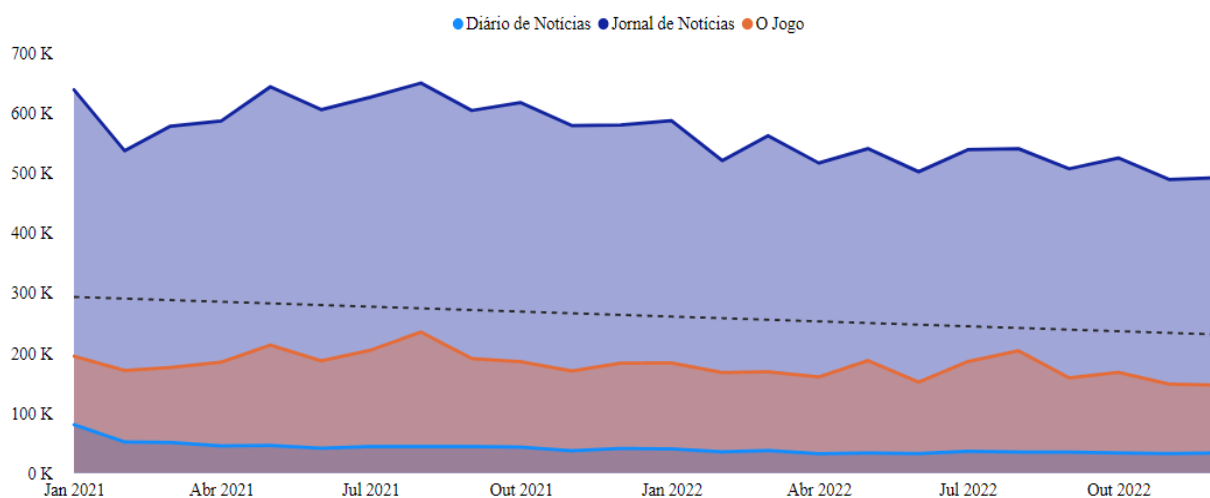


Figura 7. Evolução de vendas por publicação

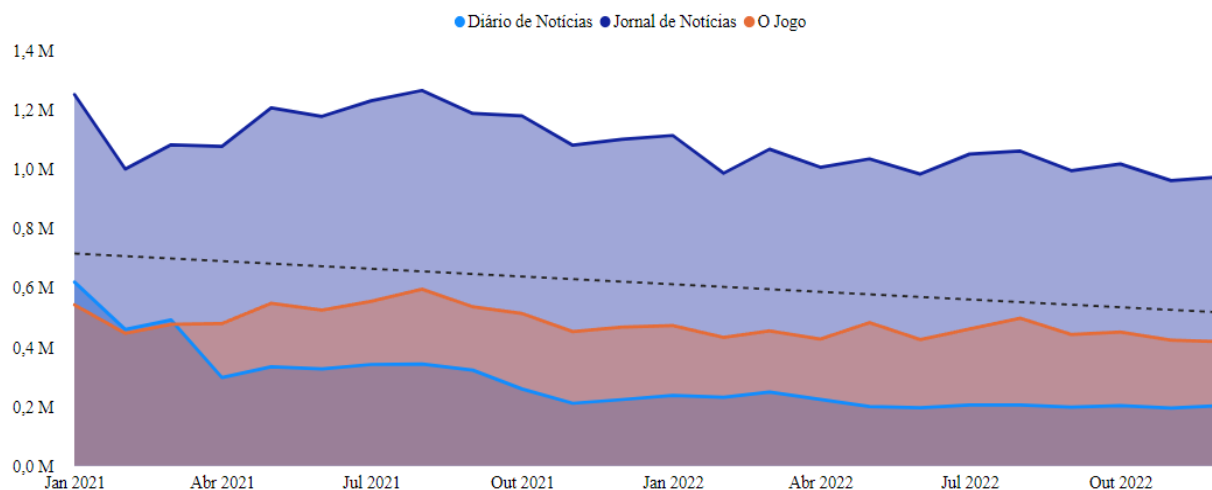


Figura 8. Evolução de tiragens por publicação

Explorando um pouco mais a sazonalidade das vendas, procedeu-se à sua análise em função do dia da semana para procurar padrões e perceber quais os dias que apresentam um maior volume de vendas médio.

Tendo em conta a Figura 9 que apresenta os dias numa escala decrescente, ordenado do dia que apresenta uma maior venda média diária, onde podemos verificar que o domingo é o dia preferencial para aquisição do jornal, seguido do sábado e da sexta-feira, o que leva a concluir que as vendas são bastante influenciadas pelo fim de semana e que, em média, seguem um padrão bastante similar ao longo do tempo.

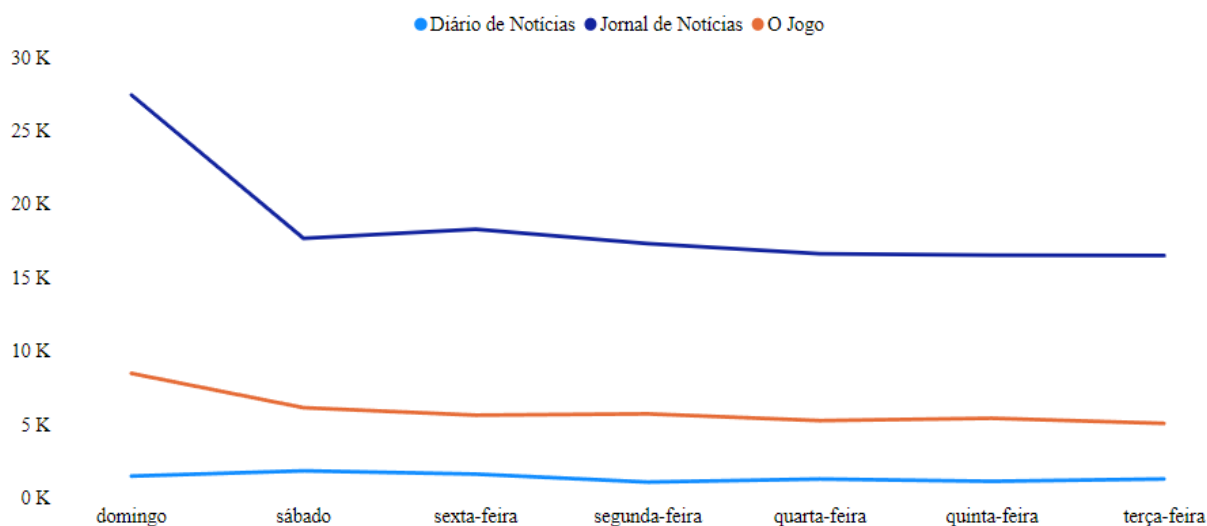


Figura 9. Tendência semanal de vendas por publicação

Outro ponto importante a analisar é a percentagem de sobras de jornais face à tiragem. A quantidade de jornais produzidos e distribuídos deve ser estimada com a menor imprecisão possível, de forma a não resultar em esgotamento de stock para determinado dia, pois sugere que poderiam ter sido vendidos mais jornais e, conseqüentemente, ter-se obtido um maior lucro. Essa estimativa é igualmente útil para evitar produzir em excesso e mitigar custos desnecessários.

Desta forma, analisando os gráficos, compreendemos que a maioria apresenta uma grande percentagem de jornais não vendidos, com maior ênfase no Diário de Notícias (DN) e no OJogo (JG), onde cerca de 72% e 46%, respetivamente, das suas tiragens não foram vendidas, representado na Figura 10.

Posteriormente, numa nova reunião com a GMG, foi abordada a diferença percentual entre os exemplares vendidos e as sobras, tendo estes desvios sido justificados como uma escolha empresarial e estratégica, de forma que a tiragem permita abranger o máximo de pontos de vendas para manter presença e visibilidade em todas as regiões de Portugal.

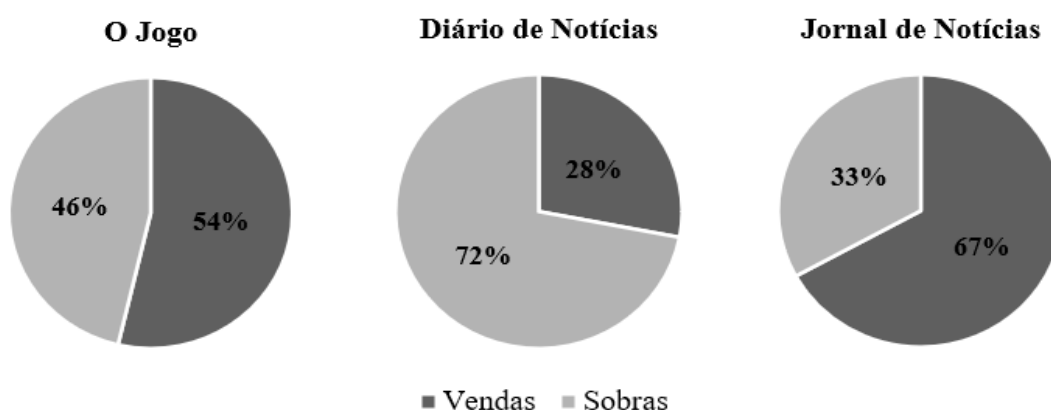


Figura 10. Divisão das tiragens em percentagem

Após concluir-se que as vendas são bastante influenciadas pelo fim de semana, foi fundamental investigar, através de uma análise referente às notícias mais relevantes que ocorreram ao longo do ano 2022, e analisar a forma como esses acontecimentos específicos, afetaram as vendas de jornais na semana em que ocorreram, comparativamente à semana homóloga.

Desta forma, a identificação de possíveis padrões do comportamento das vendas perante determinados acontecimentos permite ajustar tanto as tiragens de jornais para maximizar a procura por parte dos leitores, como ajustar possíveis ações de marketing específicas.

Analisando em específico o primeiro acontecimento de maior relevância no ano, com o início de uma guerra a 24 de fevereiro de 2022, através da invasão russa ao território ucraniano, denota-se que houve um aumento substancial de vendas quer no dia anterior, quer nos dias subsequentes, originando um crescimento de 2,2% comparativamente às vendas da semana homóloga, representado na Figura 11.

De notar que nesta análise apenas foram consideradas as vendas referentes ao Jornal de Notícias (JN), não tendo sido considerado JG dado tratar-se de um jornal desportivo que não sofre influência deste tipo de notícias.

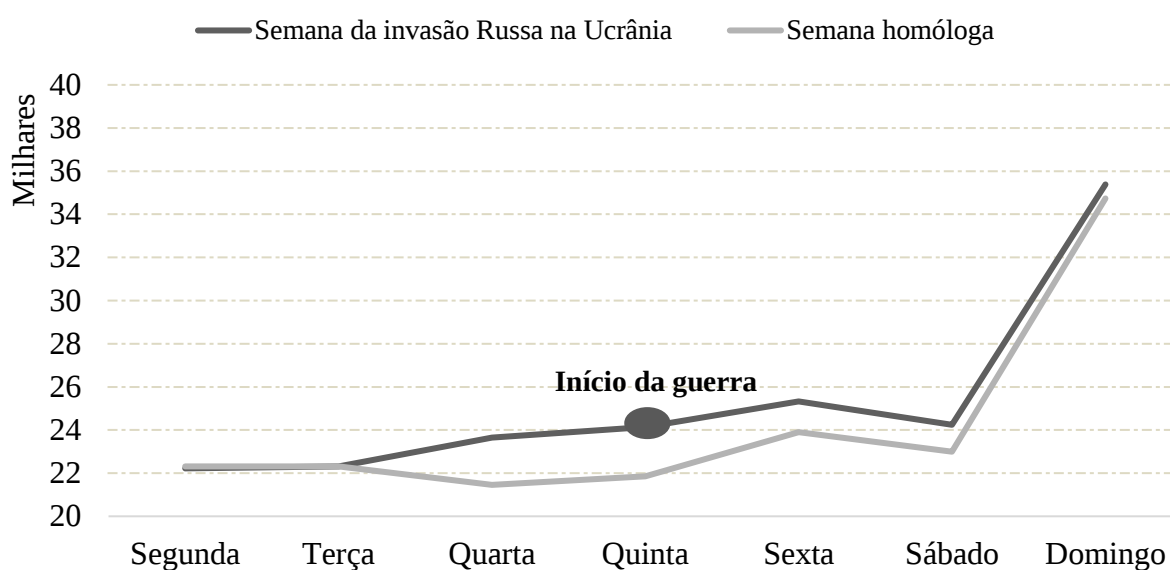


Figura 11. Invasão Russa na Ucrânia

Outro evento que provocou um aumento de vendas no JN e DN, a rondar os 2,6% face aos resultados da semana homóloga, foi a morte da Rainha Isabel II, ocorrida no dia 8 de setembro de 2022. Como se pode observar na Figura 12, os dias anteriores à data da sua morte influenciaram os leitores através da necessidade de se manterem informados do seu estado de saúde, bem como nos dias seguintes para compreender como iriam decorrer as cerimónias fúnebres e a sucessão do seu filho Rei Carlos III ao trono.

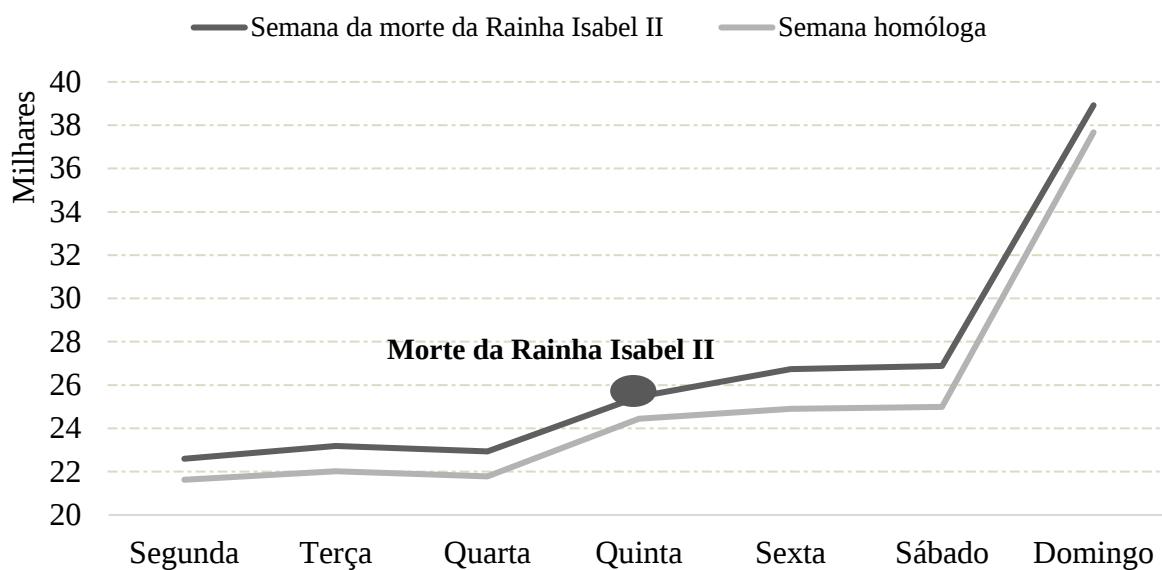


Figura 12. Morte da Rainha Isabel II

Por fim, na Figura 13, foi analisada a forma como grandes acontecimentos desportivos podem influenciar a procura do JG, em específico a conquista do campeonato nacional de futebol por parte do FC Porto no dia 7 de maio de 2022. Em geral, originou um aumento de 16% comparativamente à semana homóloga, tendo surtido um maior efeito nos dias próximos da conquista, devido às reportagens especiais contendo fotos das celebrações, entrevistas aos jogadores e treinadores, e, por vezes, ofertas de cartazes.

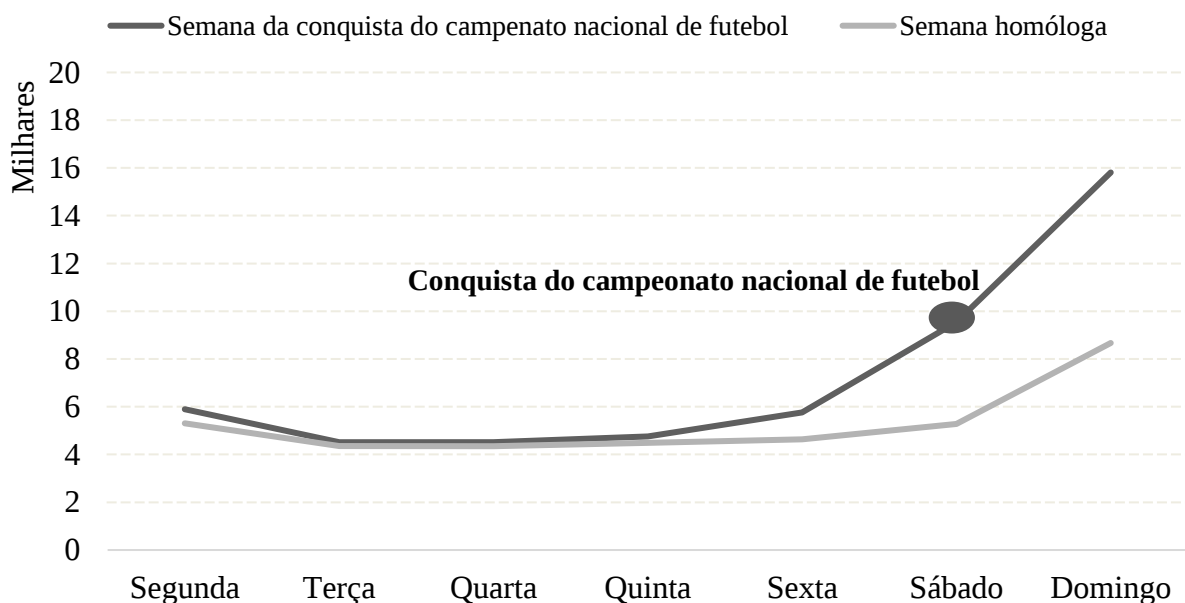


Figura 13. Conquista do campeonato nacional de futebol

3.4. Prophet – Aplicação

Nesta fase do estudo, foi implementado o algoritmo Prophet, software de código aberto desenvolvido pela equipa de ciência de dados do Facebook, para, numa primeira fase, prever os valores referentes às vendas diárias, e, numa segunda fase, prever as tiragens para as publicações em questão.

A escolha deste algoritmo [27] para resolução do problema deveu-se ao facto de ser bastante preciso e eficaz no tratamento de dados referentes a séries temporais. Acrescente-se, ainda, a sua capacidade para ter em consideração elementos como a sazonalidade, as tendências, os padrões e feriados, o que resulta numa previsão muito mais fiável.

Uma segunda razão para a sua escolha prende-se com o facto de ser consideravelmente robusto em relação à presença de *outliers*, bem como a sua adaptabilidade de parâmetros para maximizar a precisão das previsões e desta forma fornecer informações de apoio à tomada de decisão.

No que concerne à abordagem para criação de um modelo que efetivamente permita fazer previsões com o maior nível de precisão possível, foram exploradas duas abordagens, univariável e multivariável, para entender qual das duas permitiria obter os melhores resultados.

Foram utilizados os dados desde 1 de janeiro de 2021 até 31 novembro de 2022 como conjunto de treino do modelo para ambas as abordagens. O intuito desta escolha prendeu-se com a necessidade de comparar os dados reais de dezembro de 2022 com as previsões originadas pelo modelo selecionado de forma a compreender a sua precisão.

Na procura das variáveis corretas para aplicar no *Prophet*, foi utilizada a matriz de correlações, que permite identificar as principais correlações entre as variáveis de um conjunto de dados.

Conclui-se que as variáveis que melhor se correlacionam, através da observação da Figura 14, se encontram entre a variável Vnd Apurada e a variável Expedição, por apresentar quase uma correlação positiva perfeita com cerca de 0,98, seguida da elevada correlação entre a variável Expedição e Sobras com 0,84, bem como uma forte correlação entre a variável Vnd Apurada e a variável Sobras, apesar de inferior com as outras variáveis, rondando os 0,73. Por fim, a variável Feriado não se correlaciona positivamente com nenhuma das restantes variáveis.

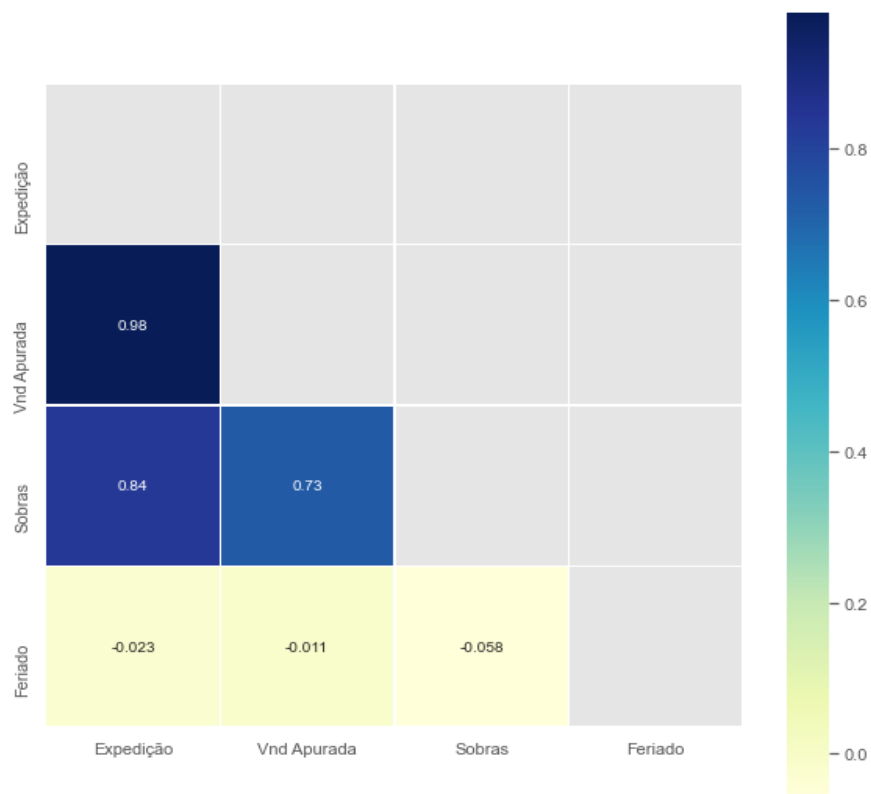


Figura 14. Matriz de correlação

3.4.1. Modelo Univariável

Numa primeira fase, foi criado simultaneamente dois conjuntos de dados, um referente à previsão de vendas e o outro referente à previsão de tiragens, uma vez que o uso de outras variáveis foi excluído nesta fase.

Desta forma, foram renomeadas as variáveis Vnd Apurada e Expedição para “y” em cada um dos conjuntos de dados, bem como a variável Data para “ds” para cumprir os requisitos de aplicação do *Prophet*. De forma a utilizar o seu histórico para prever e comparar os resultados referentes ao mês de dezembro para cada um dos jornais em estudo, com o intuito de compreender a precisão dos modelos desenvolvidos equiparando os resultados criados face aos resultados reais, foi originada a coluna “yhat” que representa os valores previstos.

No DN, os resultados originados pelos modelos, tanto para as vendas como para as tiragens, apesar de se encontrarem na mesma ordem de valores e apresentarem padrões consistentes, representados através da Figura 15 e 16, não são, de facto, os esperados face aos resultados reais.

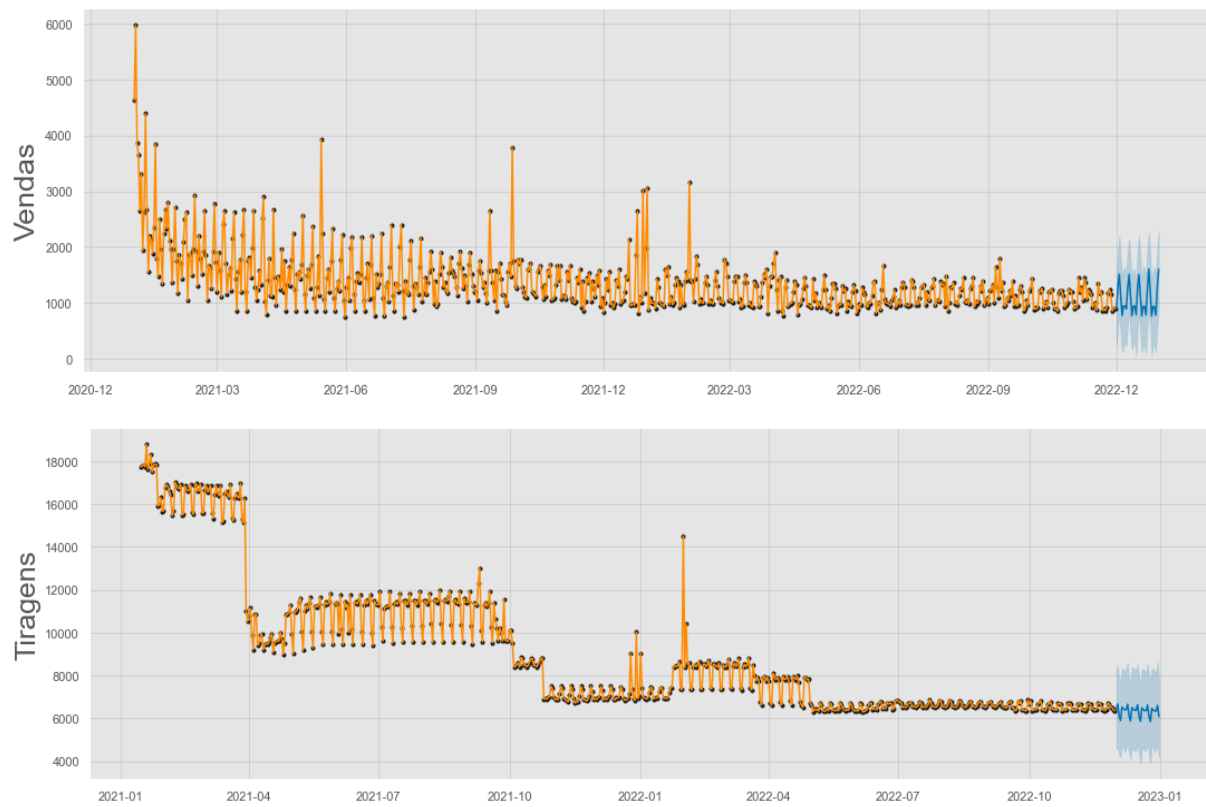


Figura 15. Séries temporais de vendas e tiragens do DN

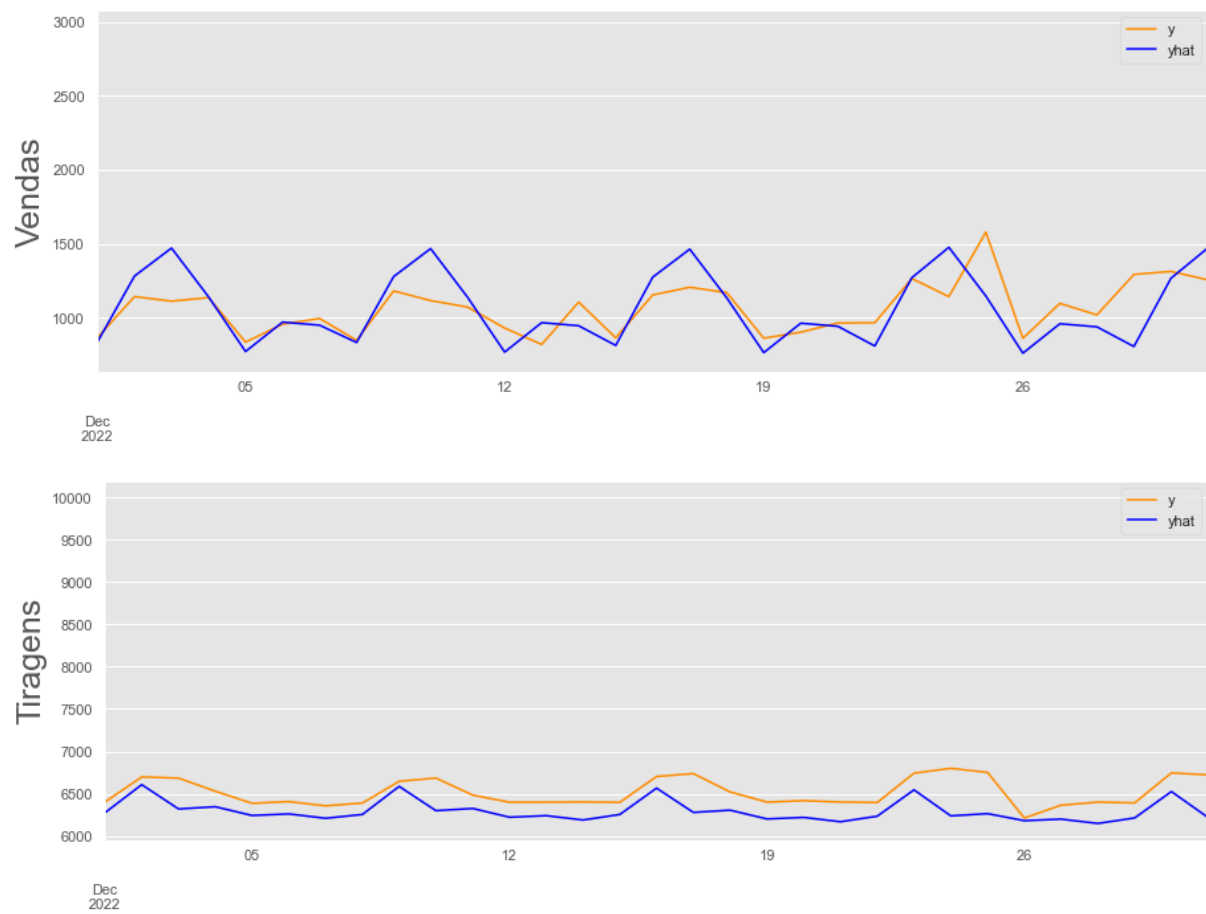


Figura 16. Real vs Previsões DN

Quanto ao JN, os resultados foram um pouco mais satisfatórios, sendo que as previsões vão ao encontro dos padrões e tendências apresentando uma grande similaridade, ilustre na Figura 17, apesar de se notar algumas disparidades face aos resultados reais, de maior notoriedade nos dias adjacentes às festas natalícias, como se pode observar na Figura 18.

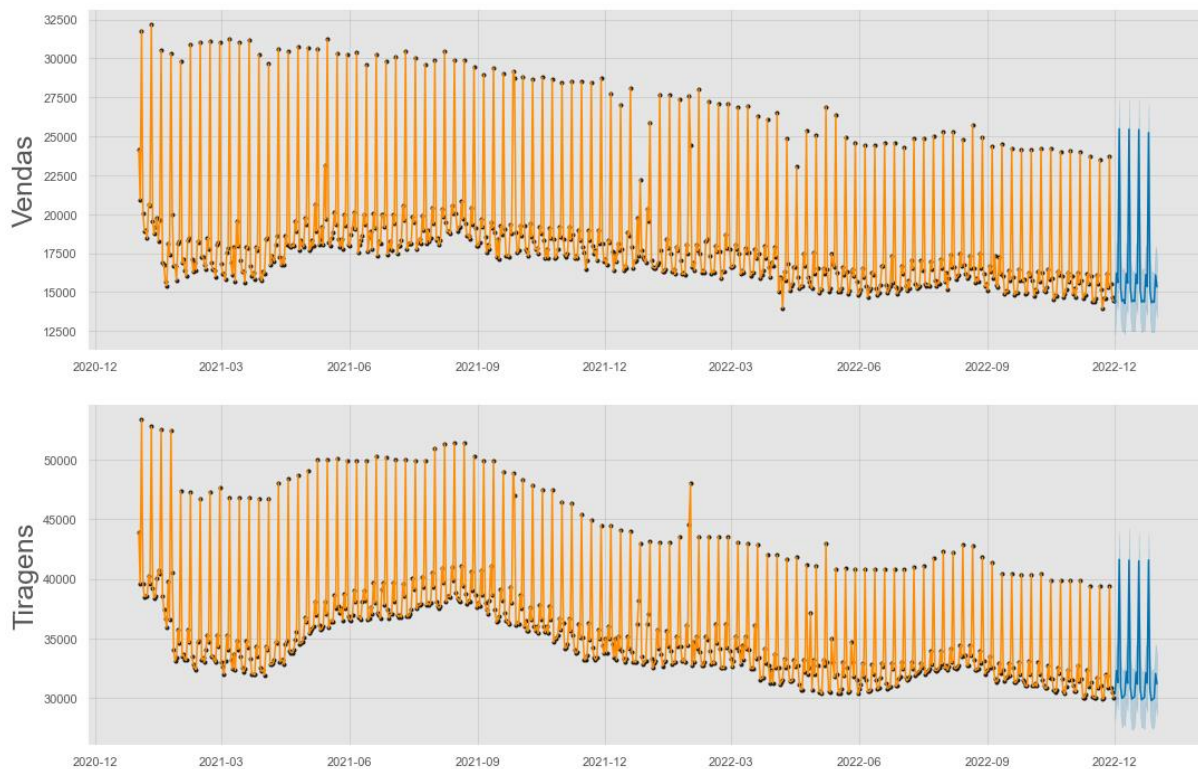


Figura 17. Séries temporais de vendas e tiragens do JN

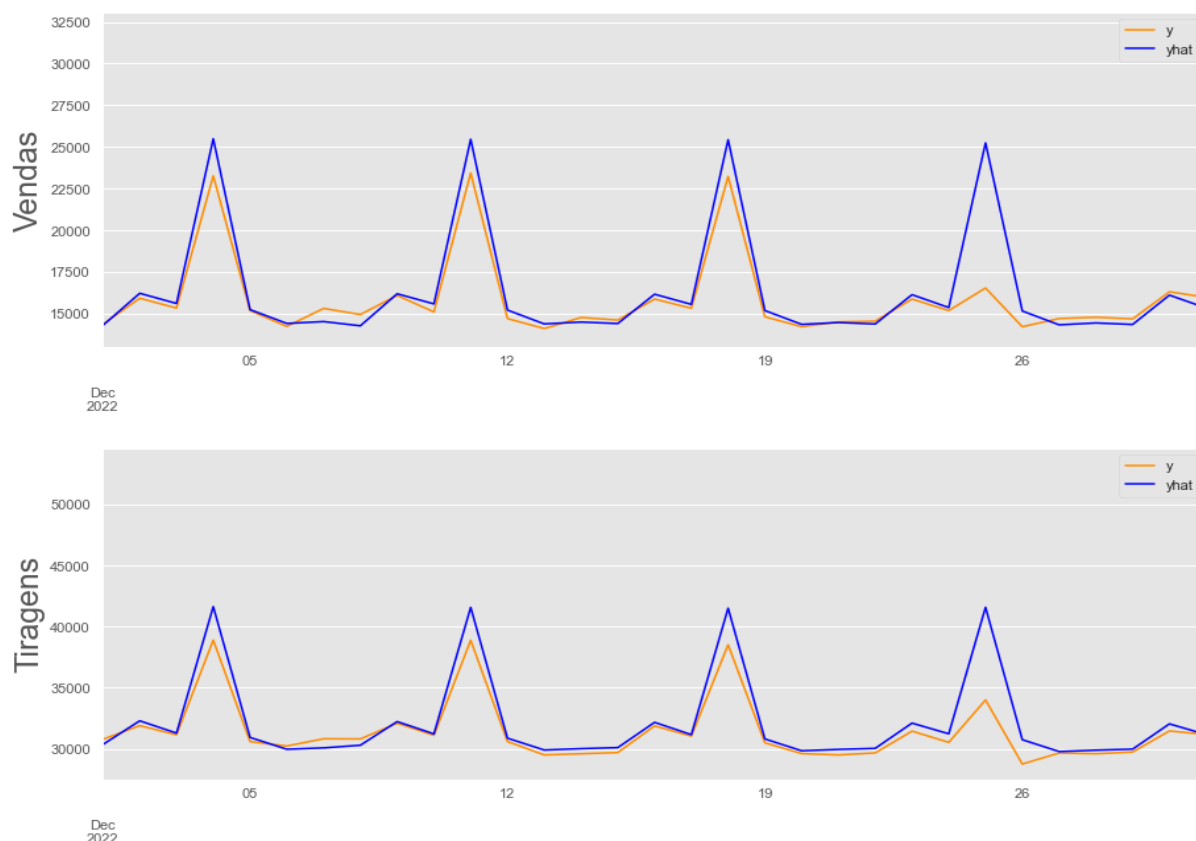


Figura 18. Real vs Previsões JN

Por fim, o JG foi o mais difícil de prever, devido não só à imprevisibilidade intrínseca à sua natureza, mas também porque sendo um jornal desportivo afeto, em específico, a um clube de futebol, está bastante dependente dos resultados que o mesmo alcança, bem como de acontecimentos relevantes no desporto. Outro dos motivos será uma sazonalidade anteriormente identificada nos meses de verão, que poderá, de certa forma, estar a influenciar em demasia os resultados de previsão obtidos, identificável na Figura 19.

Na Figura 20 observamos os resultados tanto para as vendas como para as tiragens, apesar de se enquadrarem na mesma ordem de valores, apresentaram padrões inconsistentes bem como previsões bastante díspares, ficando aquém face aos resultados reais.

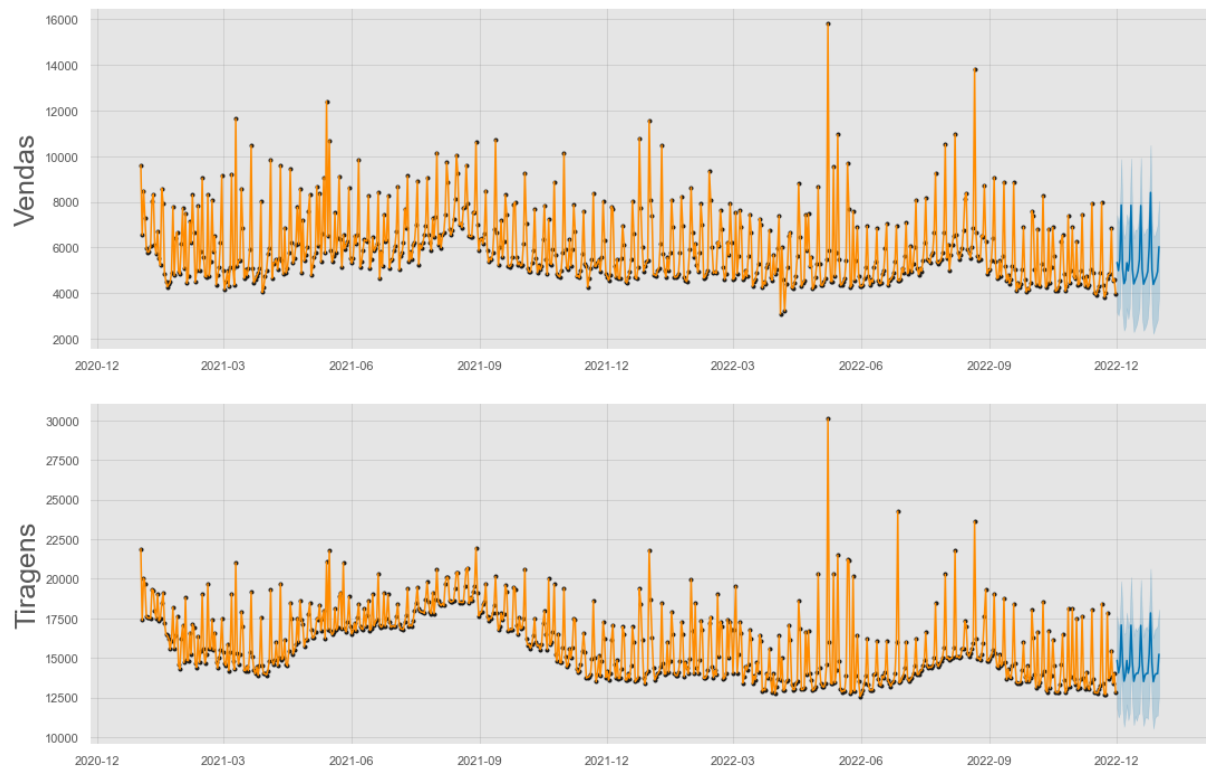


Figura 19. Séries temporais de vendas e tiragens do JG

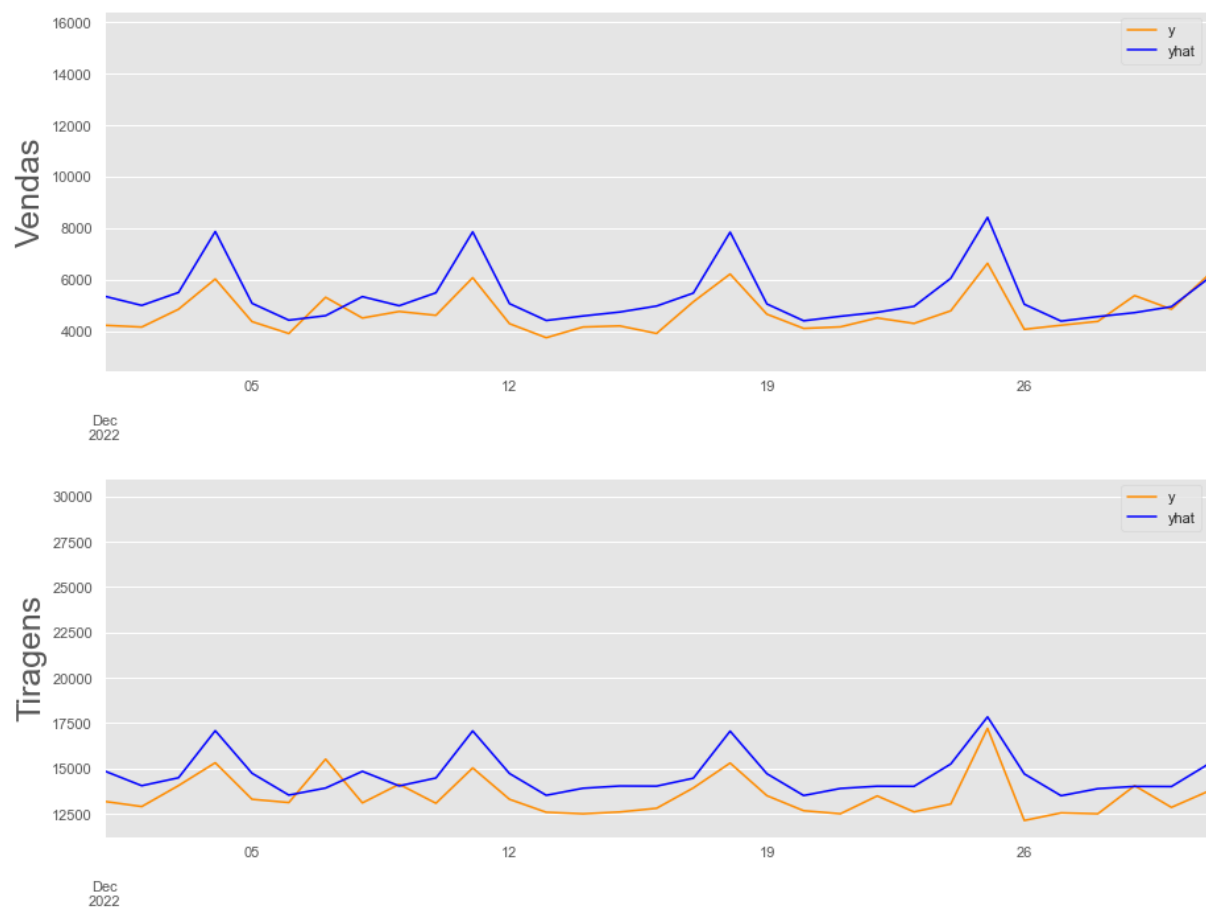


Figura 20. Real vs Previsões JG

3.4.2. Modelo Multivariável

Foi necessário experimentar a utilização de regressores, e avaliar se a metodologia adotada faria os modelos atingirem melhores resultados. De facto, seria possível que trouxessem ineficiências ao algoritmo, ou, pelo contrário, que o melhorassem.

Desta forma, aplicando dois conjuntos diferentes de regressores, nomeadamente, a escolha das variáveis Expedição e Sobras que melhores correlações apresentavam com a variável Vnd Apurada, permitiu uma possível melhoria dos modelos de previsão de vendas, bem como a utilização do conjunto de variáveis Vnd Apurada e Sobras para prever as tiragens.

No DN, os resultados sofreram melhorias significativas face aos modelos univariados, permitindo, através do modelo multivariado, identificar valores considerados habituais, falhando apenas em pontuais *outliers*, mas permitindo ter uma previsão adequada, muito próxima dos valores e padrões reais de vendas, observável na Figura 21. Quanto às tiragens, o facto de terem uma flutuação de valores mais regular permite obter melhores resultados, sendo as tiragens provisionadas quase idênticas às tiragens reais.

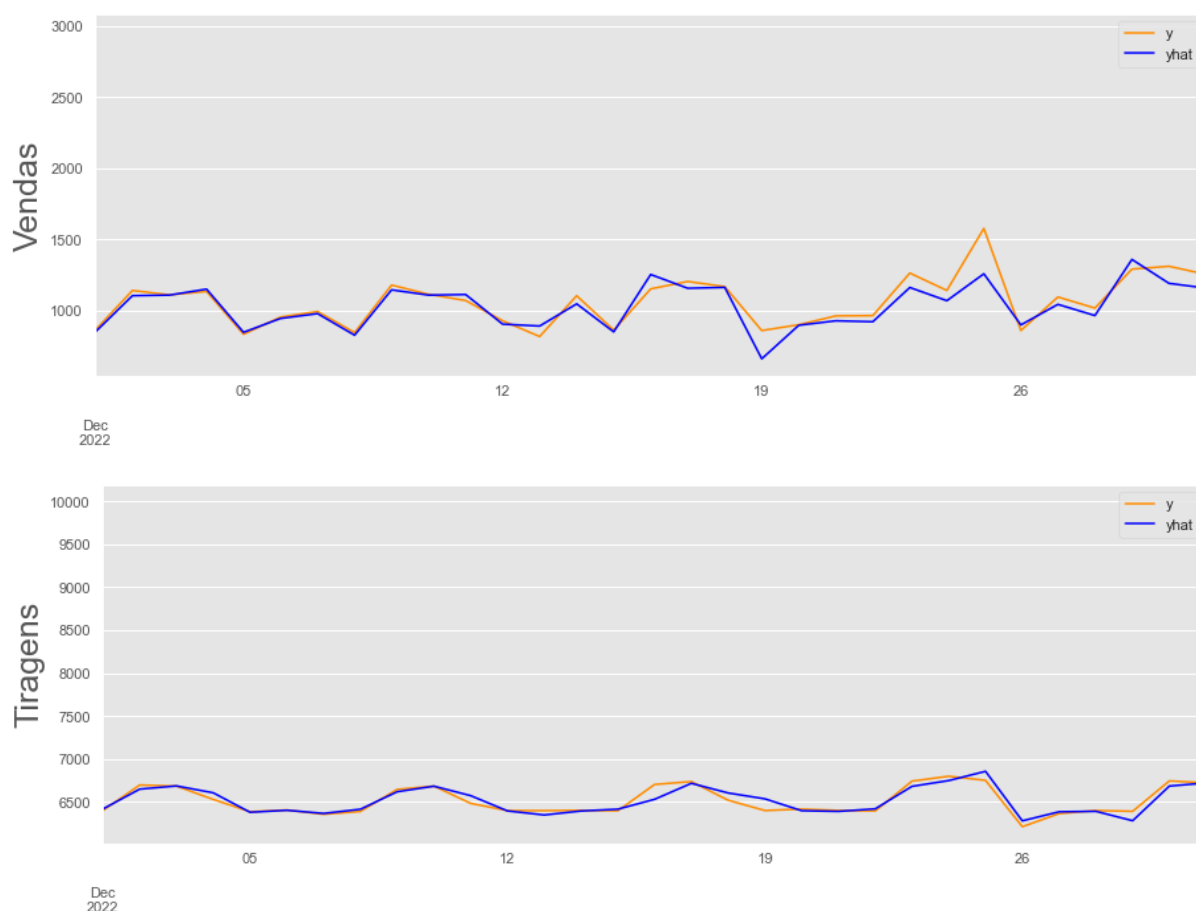


Figura 21. Real vs Previsões DN

No que se refere ao JN, os resultados da Figura 22 foram bastante satisfatórios, superando novamente o modelo univariado, e apresentando resultados com padrões e tendências quase idênticos aos resultados reais, exibindo uma grande precisão e robustez do modelo face a *outliers*.

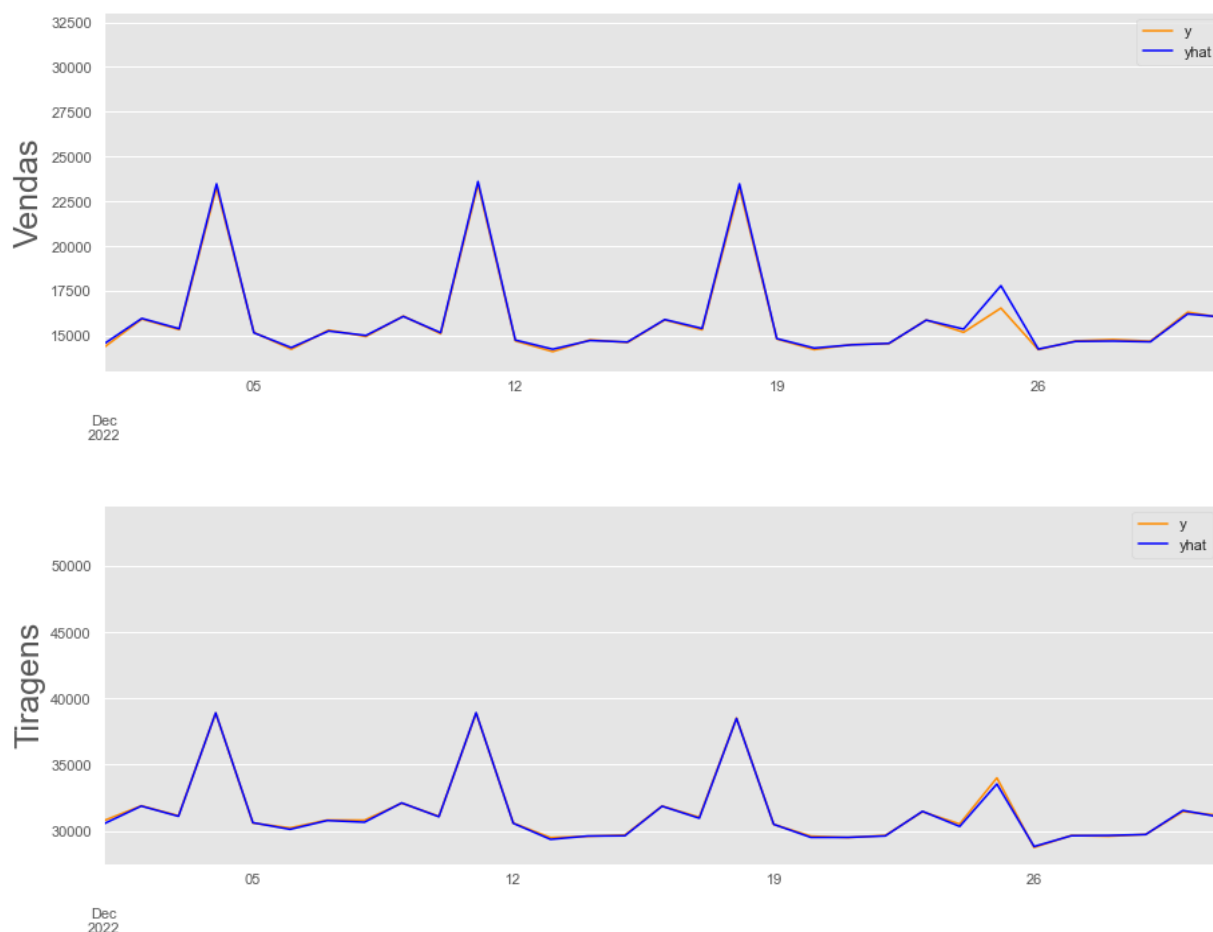


Figura 22. Real vs Previsões JN

Quanto ao JG, observamos na Figura 23 também uma grande melhoria, tendo obtido valores bastante semelhantes aos reais. De notar que estes resultados são bastante positivos, devido à natureza do jornal ser por si bastante imprevisível dependente que é de acontecimentos e resultados desportivos, como referido anteriormente.

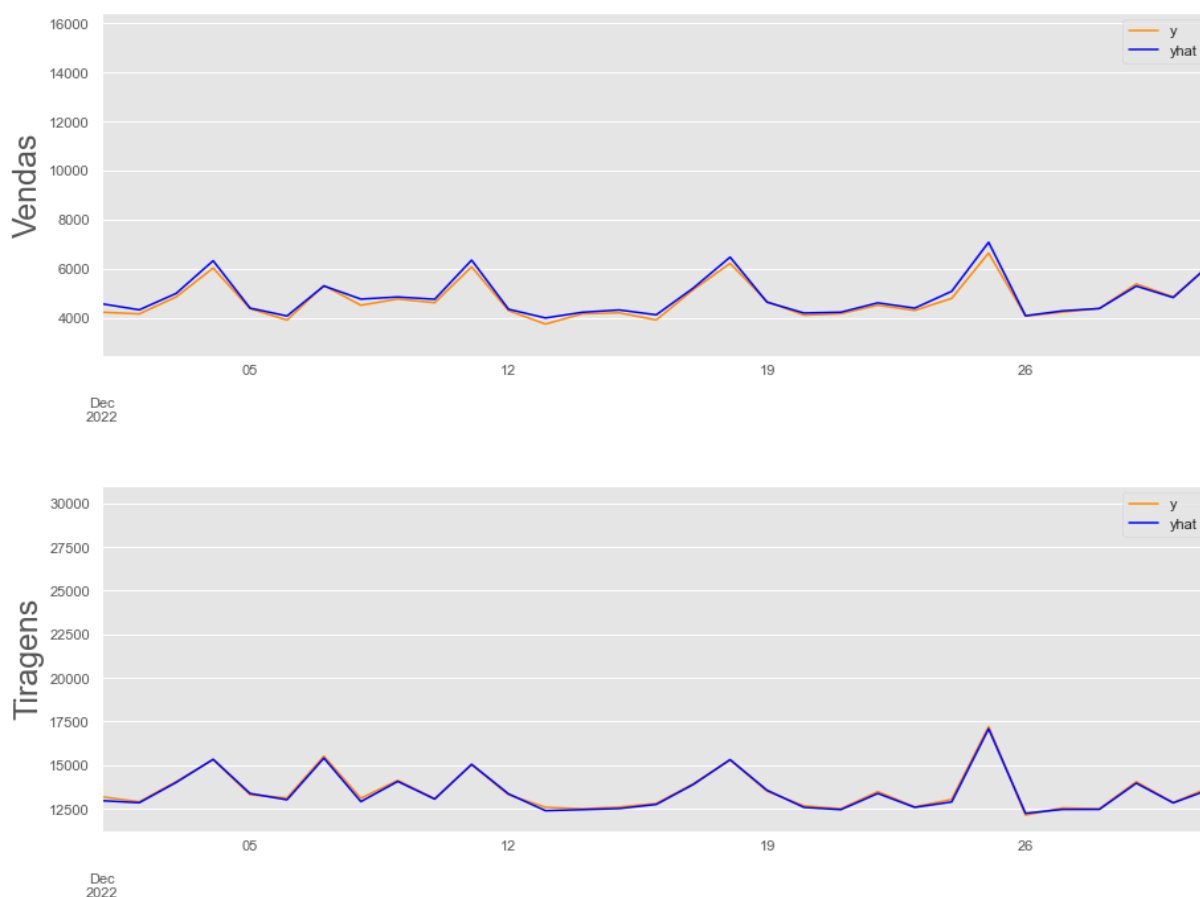


Figura 23. Real vs Previsões JG

Nesta fase procedeu-se à extração dos componentes dos dados previstos para entender de que modo o seu padrão semanal e mensal, bem como o facto de um determinado dia ser ou não feriado, contribui, significativamente, para influenciar os resultados dos modelos de previsão.

Nas Figura 24, 25 e 26, observa-se que os modelos apresentaram tendências negativas semelhantes para os três jornais ao longo do mês de dezembro, bem como gráficos equivalentes referentes à identificação de feriados, apresentando somente diferenças no que dizia respeito aos seus padrões semanais, especificamente nas variações entre os dias de semana e os fins de semana, mais notáveis no JN e no JG.

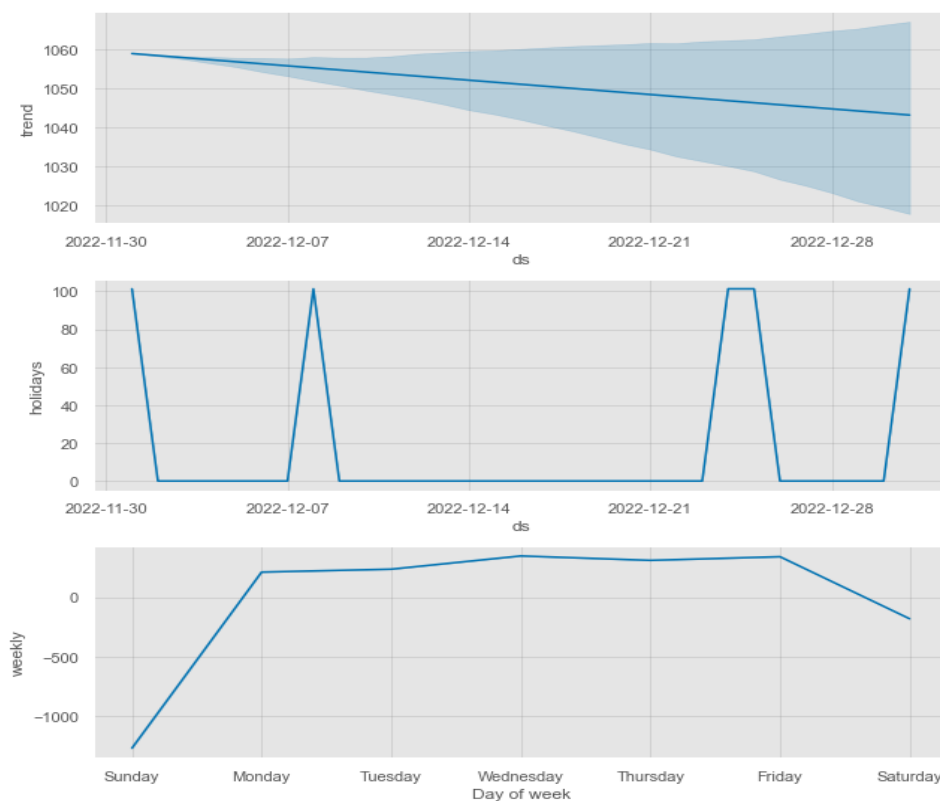


Figura 24. Gráficos mensais e semanais com previsão de dezembro para o JN

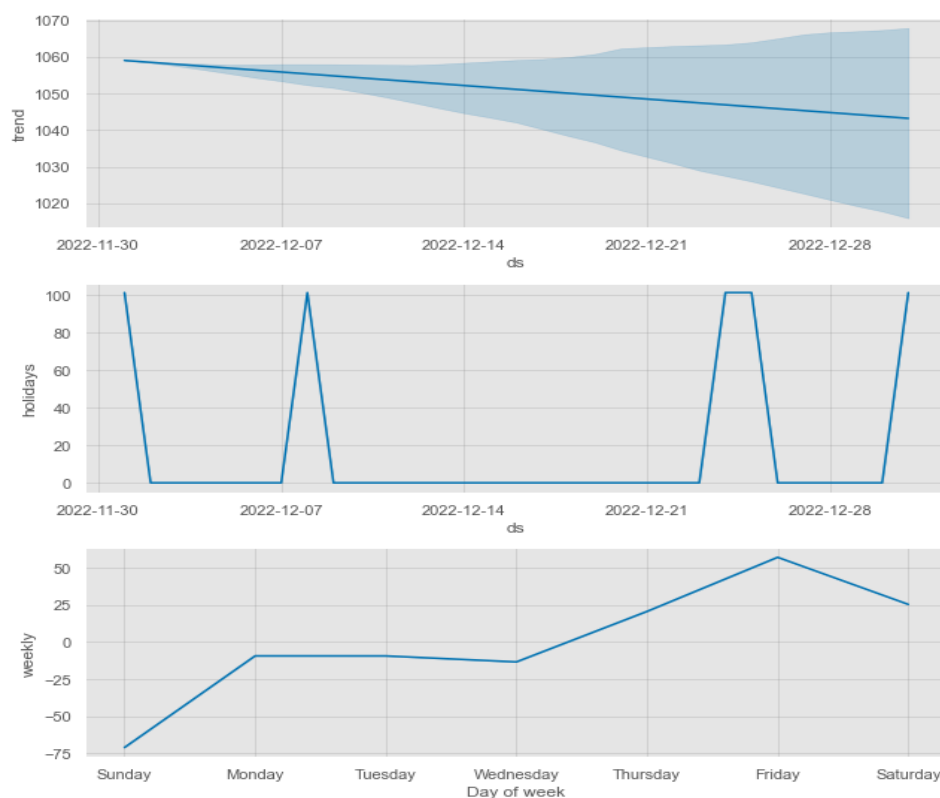


Figura 25. Gráficos mensais e semanais com previsão de dezembro para o DN

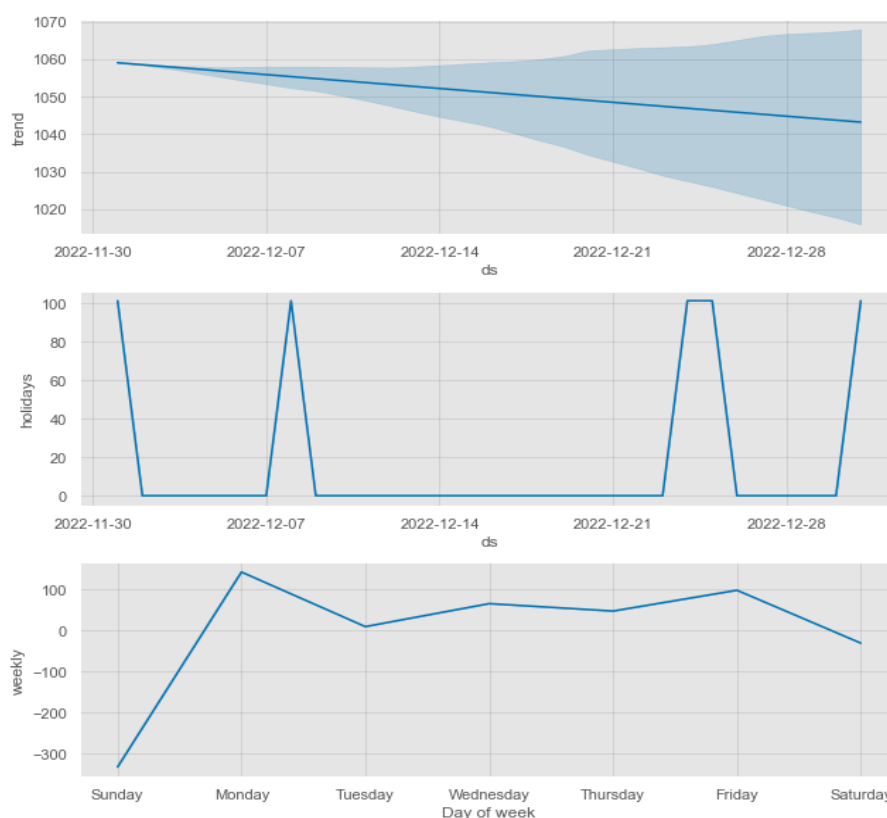


Figura 26. Gráficos mensais e semanais com previsão de dezembro para o JG

3.4.3. Comparação Modelo Multivariável vs Univariável

Com os resultados obtidos através da busca do melhor modelo de previsão, foi necessário estabelecer uma comparação entre os modelos univariados e os multivariados, de forma a avaliar o desempenho de ambos uma vez que as previsões previamente realizadas se assemelham bastante ao real, tornando essencial uma sobreposição dos gráficos.

Referentemente aos modelos de previsão de vendas representados na Figura 27, foi possível perceber a sua precisão e eficácia, tendo-se conseguido, de uma forma geral, prever resultados com uma grande confiança para cada um dos jornais em estudo.

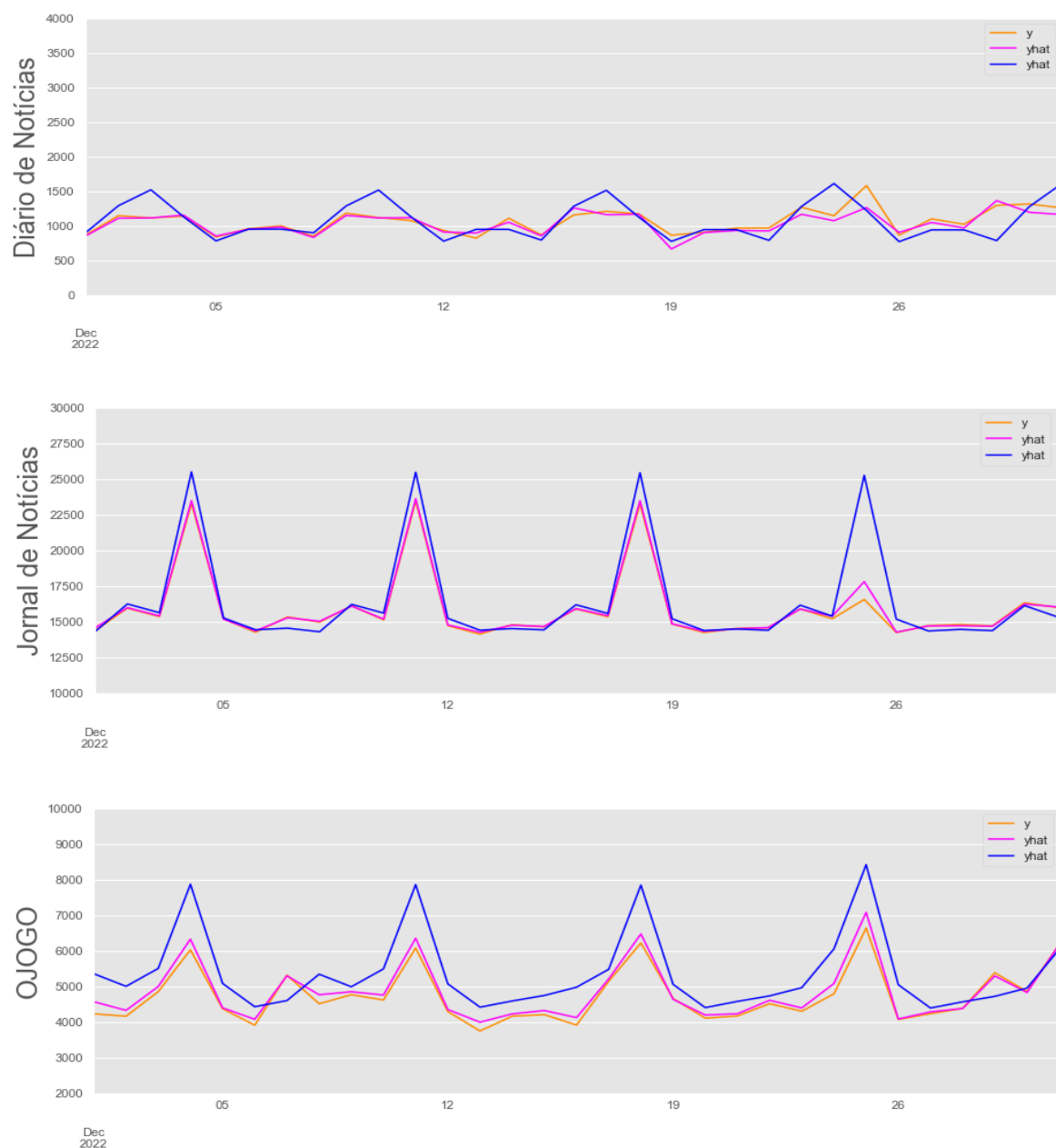


Figura 27. Comparação dos modelos de vendas

Por fim, fazendo uma sobreposição dos gráficos da Figura 28 para os modelos de tiragens referentes a cada jornal, obtivemos os seguintes outputs, que traduzem, novamente, excelentes resultados.

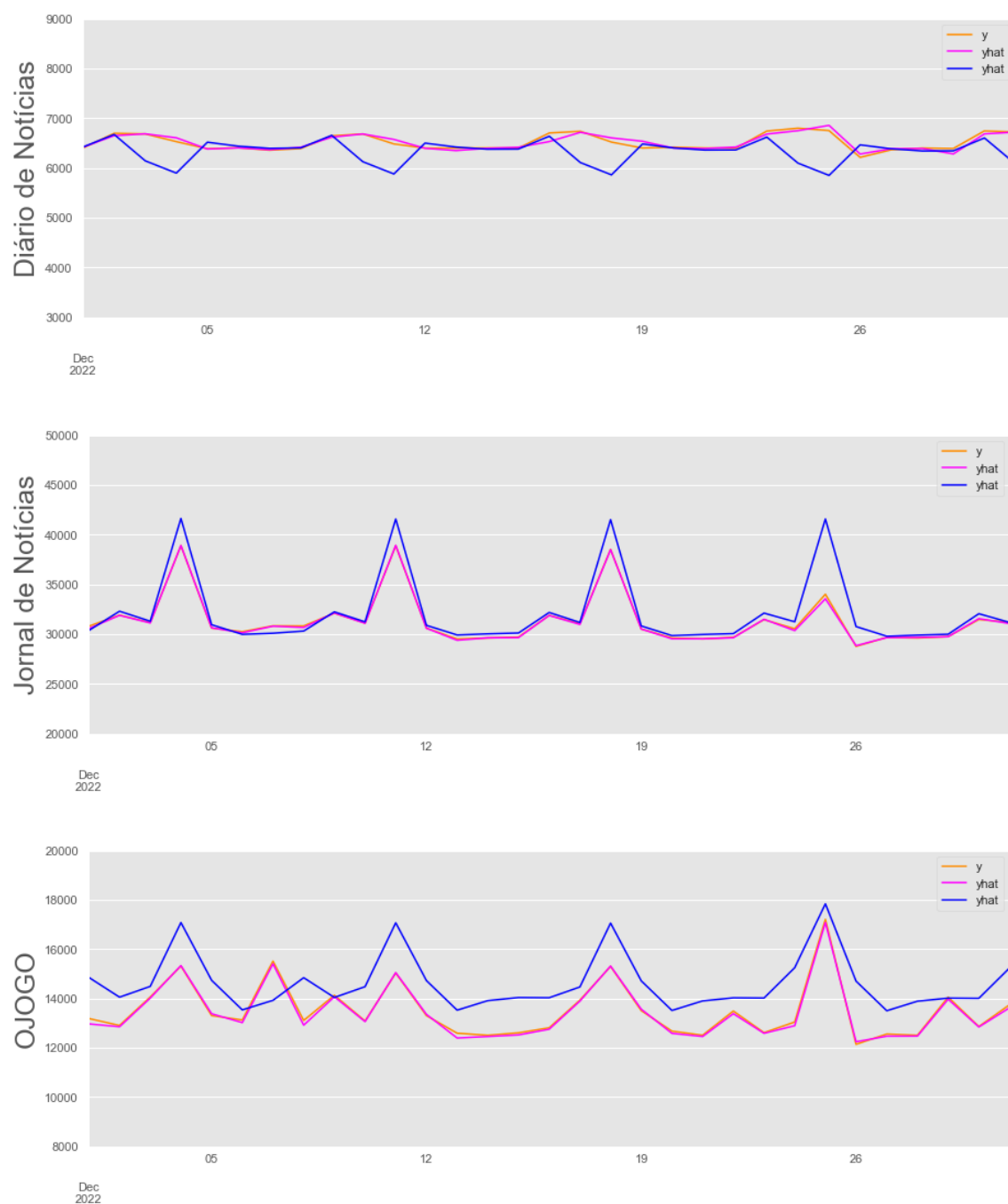


Figura 28. Comparação dos modelos de tiragens

Com base nestes resultados, conclui-se que a aplicação do *Prophet* com regressores se revela de grande eficácia, podendo evidenciar a clara superioridade dos modelos multivariados em questões de performance, no que toca ao seu rigor mais aproximado, onde em alguns casos se sobrepõe de tal maneira que igualam os dados reais.

Após a visualização através de gráficos, foram representadas nas Tabelas 4 e 5 as métricas para cada jornal e modelo, de forma a verificar a precisão e qualidade de cada um dos modelos criados. A eficácia do algoritmo pode ser avaliada pelo coeficiente de determinação obtido, que se verificou mais ajustado nos modelos multivariados, tanto para as vendas como para as tiragens.

Tabela 4. Métricas de análise para modelos de previsão - vendas

	DN		JN		JG	
	Univariável	Multivariável	Univariável	Multivariável	Univariável	Multivariável
Coefficiente de Determinação	-0,45	0,76	0,52	0,99	-0,34	0,94
Erro Médio Quadrado	43942,13	7184,26	3027110,71	59587,38	786266,03	31070,24
Erro Médio Absoluto	151,60	55,80	770,59	112,22	736,28	135,53

Tabela 5. Métricas de análise para modelos de previsão - tiragens

	DN		JN		JG	
	Univariável	Multivariável	Univariável	Multivariável	Univariável	Multivariável
Coefficiente de Determinação	-4,05	0,85	0,57	0,99	-0,48	0,99
Erro Médio Quadrado	131435,12	3749,70	2872065,09	13473,05	1862024,76	9343,45
Erro Médio Absoluto	231,87	43,30	868,85	73,06	1233,39	80,51

3.5. Avaliação de resultados

De forma a avaliar o trabalho realizado, foi realizado um simples questionário e enviado aos representantes da GMG, nomeadamente no que respeita aos modelos de previsão criados, representado na Tabela 6.

Tanto as questões colocadas como os resultados obtidos observados permitem desta forma concluir que de uma forma geral os resultados foram bastante consistentes, cumprindo os requisitos propostos pela empresa para atingir os seus objetivos. Assim, foram definidos quatro níveis de avaliação que permitem classificar o trabalho desenvolvido:

- Não alcançado (NA) - [0-9];
- Parcialmente alcançado (PA) -]10-14];
- Alcançado (A) -]15-17];
- Completamente alcançado (CA) -]18-20];

Tabela 6. Questionário Modelos de Previsão

Questões	Avaliador 1
Auxilia a empresa no processo de avaliação e tomada de decisões empresariais?	CA
Fornece resultados concretos e ajustados à realidade?	CA
Apresenta resultados detalhados?	A
A abordagem adotada em relação aos dados disponíveis parece-lhe a mais correta?	A

Implementação do *Dashboard*

4.1. Dados e conhecimento do negócio

Nesta etapa realizou-se novamente uma reunião em conjunto com a GMG onde se procurou compreender os objetivos referentes ao desenvolvimento de um dashboard de apoio à tomada de decisão que permitisse analisar tendências de evolução temporal e geográfica de vendas de diferentes jornais e revistas, o que permitiria avaliar e tomar decisões através da visualização de dados, bem com das necessidades empresariais e acompanhado os objetivos orçamentados da empresa.

É importante notar que foi necessário compreender a ferramenta utilizada para construir uma visualização completa que permitisse uma maior compreensão tanto dos dados como de aspetos específicos e fundamentais do negócio, cumprindo-se, assim, os objetivos delineados e acordados com a empresa.

Foi fornecido, por parte do departamento de circulação da GMG, um ficheiro em *Excel* totalmente organizado que apresentava uma grande variedade de dados e, consequentemente, permitia uma grande extração de conhecimentos dos mesmos. O documento contemplava três folhas com a seguinte estrutura: Ano corrente (Tabela 7) com 19 colunas e 85 linhas; Ano homólogo (Tabela 8) com 13 colunas e 85 linhas; Vendas distrito (Tabela 9) com 5 linhas e 1597 linhas.

Tabela 7. Ano corrente

Variável	Descrição
Meses	Categoriza os meses do ano
Média banca	Representa a venda média em banca para um determinado mês
% sobras	Representa a % de sobras de jornais em banca para um determinado mês
VD	Representa a média de vendas diretas para um determinado mês
Assinaturas papel	Representa a média diária de assinaturas em papel para um determinado mês
Assinaturas digitais	Representa a média diária de assinaturas digitais para um determinado mês
Total assinaturas	Representa a soma das assinaturas em papel e em digital para um determinado mês
Circulação Paga	Representa a soma da venda média em banca com a totalidade de assinaturas para um determinado mês

Ofertas	Representa a média diária de ofertas para um determinado mês
Circulação total	Representa a soma da circulação paga com as ofertas para um determinado mês
Orçamento vendas banca	Representa a venda média em banca orçamentada para um determinado mês
Orçamento Sobras	Representa a % de sobras de jornais em banca orçamentada para um determinado mês
Orçamento VD	Representa a média de vendas diretas orçamentada para um determinado mês
Orçamento Papel	Representa a média diária de assinaturas em papel orçamentada para um determinado mês
Orçamento Digital	Representa a média diária de assinaturas digitais orçamentada para um determinado mês
Orçamento Circ. Paga	Representa a média diária de circulação paga orçamentada para um determinado mês
Publicação	Categoriza a publicação jornalística

Tabela 8. Ano homólogo

Variável	Descrição
Meses	Categoriza os meses do ano
Média banca	Representa a venda média em banca para um determinado mês
% sobr.	Representa a % de sobras de jornais em banca para um determinado mês
VD	Representa a média de vendas diretas para um determinado mês
Assinaturas papel	Representa a média diária de assinaturas em papel para um determinado mês
Assinaturas digitais	Representa a média diária de assinaturas digitais para um determinado mês
Total assinaturas	Representa a soma das assinaturas em papel e digital para um determinado mês
Circulação Paga	Representa a soma da venda média em banca com a totalidade de assinaturas para um determinado mês
Ofertas	Representa a média diária de ofertas para um determinado mês
Circulação total	Representa a soma da circulação paga com as ofertas para um determinado mês
Publicação	Categoriza a publicação jornalística

Tabela 9. Vendas por distrito

Variável	Descrição
Meses	Categoriza os meses do ano
Distritos	Categoriza os distritos de Portugal
Vendas	Representa as vendas em banca num determinado distrito e mês
Publicação	Categoriza a publicação jornalística

4.2. Dashboard

Tendo em conta os objetivos delineados com a GMG, foi solicitado que o *dashboard* fosse adaptado às necessidades empresariais e que cumprisse determinados parâmetros, uma vez que o mesmo seria utilizado para a gestão de topo tomar decisões.

Como tal, foi solicitado que o mesmo fosse construído de forma a permitir uma visão geral de todos os aspetos relevantes da circulação da empresa, e que, através de uma fácil interpretação, permitisse compreender os principais indicadores de desempenho de jornais e revistas, não só através de gráficos, como também de outros elementos visuais que indicassem métricas importantes, tais como a média do ano atual, a média do ano homólogo, a média orçamental, a percentagem de desvio entre o ano atual e o ano homólogo, e a percentagem de desvio entre o ano atual e o orçamento.

Como resultado, e em resumo, seria necessário para responder às questões propostas e fornecer informações aos futuros utilizadores através da criação de páginas no *dashboard* que abordassem os seguintes temas:

1) Circulação total paga – A circulação paga desempenha um papel crucial para ajudar a garantir a sustentabilidade financeira da empresa, sendo que representa a totalidade das vendas em banca, bem como as vendas de assinaturas.

Este *dashboard* inicial, representado na Figura 29, permite não só ter uma visão global ao acompanhar o desempenho das diferentes publicações, relativamente às suas quantidades, através do gráfico de linhas onde estão representados o ano atual, o ano homólogo e o orçamento em termos médios mensais, bem como permite a escolha individual de uma publicação através dos seus respetivos logótipos onde seja possível identificar diferentes tendências e padrões.

Simultaneamente, na tabela acima do gráfico de linhas, é possível apurar os desvios percentuais face ao ano homólogo e ao orçamento.

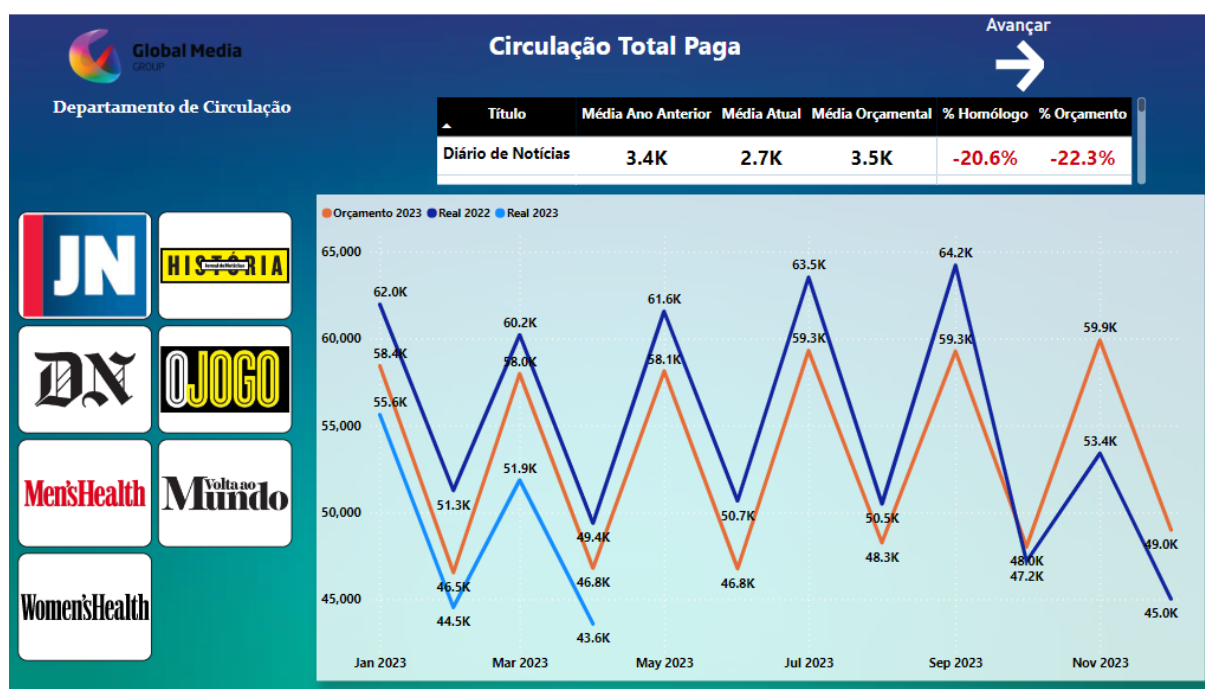


Figura 29. Circulação Total Paga

2) Vendas em banca – Apesar de terem vindo a diminuir ao longo dos anos devido a uma constante e inexorável transformação digital, as vendas em banca ainda desempenham um papel significativo no que diz respeito a receitas, especialmente para os jornais diários e as revistas impressas.

Como tal, era imperativo ter uma visão abrangente das métricas relacionadas com as vendas em banca, como o volume de vendas, a localização das bancas, o desempenho por região, a evolução ao longo do tempo e a comparação entre diferentes publicações.

Desta forma, foram desenvolvidas duas páginas, a primeira representada através da figura 30, onde foi recriado, uma construção e visualização bastante similar à Figura 29, com a pequena diferença de se tratar de um gráfico onde é possível observar a percentagem de vendas nas zonas norte e sul de Portugal.

A segunda página, representada na Figura 31, fornece dados úteis tanto para uma análise de mercado, como para uma segmentação do público, permitindo, através de informações sobre o desempenho de vendas em diferentes regiões geográficas ao longo do tempo, perceber quais os distritos mais representativos em termos de vendas em banca. Simultaneamente, serve para auxiliar na compreensão dos interesses dos leitores em diversas zonas e desenvolver estratégias de marketing mais eficazes.



Figura 30. Vendas em Banca

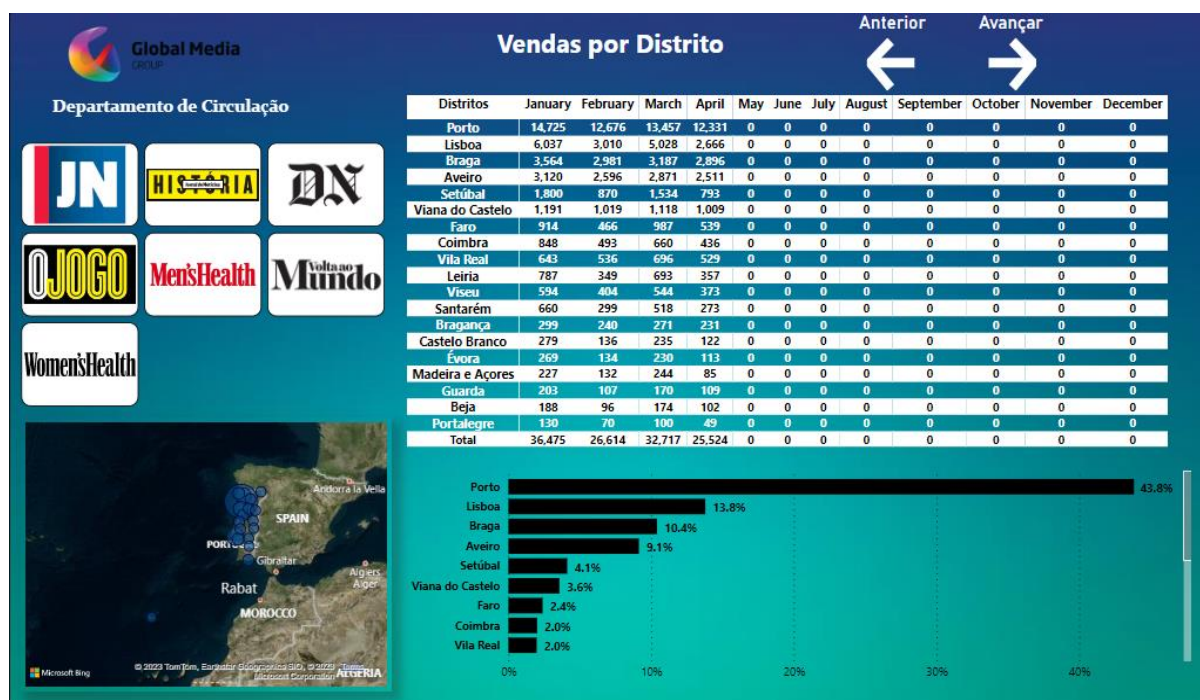


Figura 31. Vendas por Zona Geográfica

3) Evolução de assinaturas – Fazendo parte da circulação paga, é importante analisar como evoluem as assinaturas em papel, assim como as assinaturas digitais que têm vindo a evoluir e a ganhar mais ênfase com a transformação digital dos *media*.

Nesta última página, representada na Figura 32, estabelece-se uma comparação entre as assinaturas em papel e as digitais, com o intuito de não só perceber a sua evolução temporal como também perceber a percentagem de representação de cada uma face ao total de assinaturas.

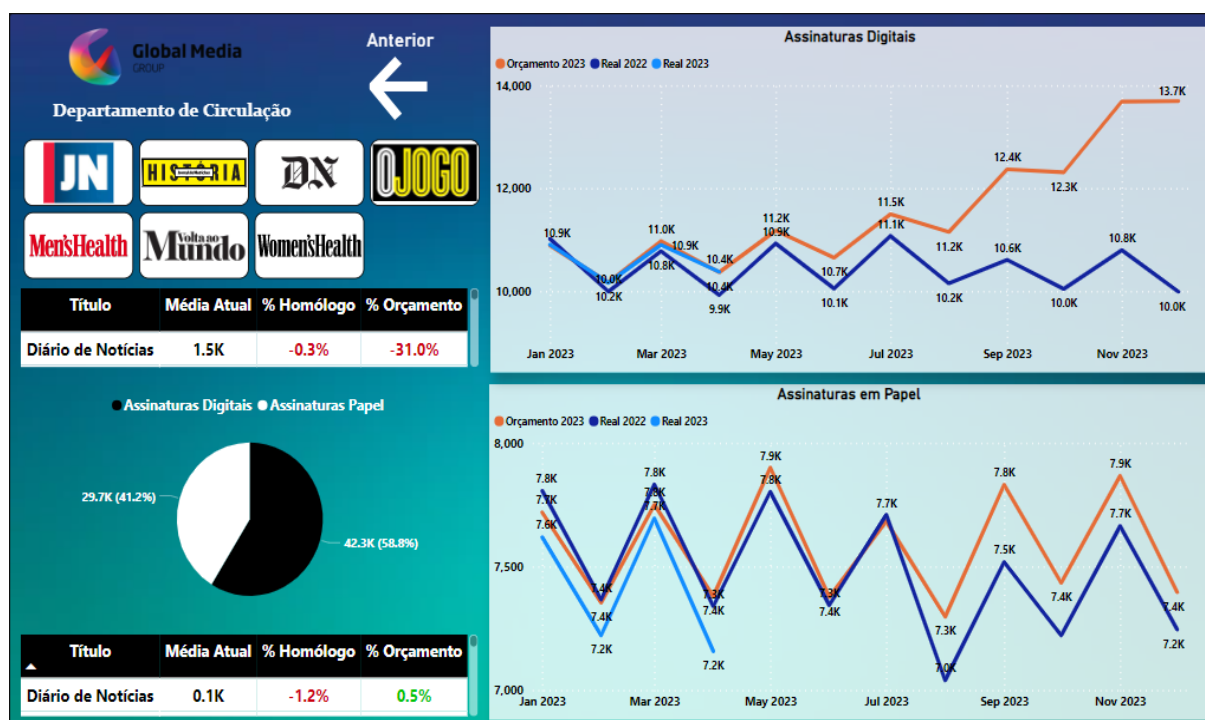


Figura 32. Evolução de assinaturas

4.3. Avaliação de resultados

De forma a avaliar o trabalho realizado, foi novamente realizado um simples questionário e enviado aos representantes da GMG, nomeadamente no que respeita à criação e implementação de um *dashboard* para apoio à tomada de decisão, representado na Tabela 10.

Tanto as questões colocadas como os resultados obtidos observados permitem desta forma concluir que de uma forma geral os resultados foram bastante consistentes, cumprindo os requisitos propostos pela empresa para atingir os seus objetivos. Assim, foram definidos quatro níveis de avaliação que permitem classificar o trabalho desenvolvido:

- Não alcançado (NA) - [0-9];
- Parcialmente alcançado (PA) -]10-14];
- Alcançado (A) -]15-17];
- Completamente alcançado (CA) -]18-20];

Tabela 10. Questionário Dashboard

Questões	Avaliador 2
Auxilia a empresa no processo de avaliação e tomada de decisões empresariais?	CA
Permite uma maior compreensão tanto dos dados como de aspetos específicos e fundamentais do negócio?	A
Permite analisar tendências de evolução temporal e geográfica?	CA
A abordagem adoptada em relação aos dados disponíveis parece-lhe a mais correcta?	A
Os resultados apresentados revelaram-se inovadores, relevantes ou apresentados de uma perspetiva mais interessante do que o habitual?	A

Conclusões

5.1. Discussão de resultados

A título conclusivo, é importante referir que o trabalho de análise que se encetou e do qual se apresentam agora os resultados atingiu os objetivos propostos no capítulo 1. Por conseguinte, importa sublinhar a sua importância e os vários benefícios que apresentou para a empresa.

A investigação iniciou-se, num primeiro momento, com a exploração de padrões, tendências e sazonalidades observadas entre as variáveis em questão, referentes a três dos principais jornais de circulação nacional, tanto no que respeita a tiragens como no que concerne a vendas. Esta visualização permitiu determinar quais os dias, ao longo da semana, em que ocorre um maior volume de vendas, vendas estas passíveis de serem verificadas pelos diversos picos de consumo bastante similares, bem como a evidente sazonalidade de vendas em períodos coincidentes com férias.

Permitiu, igualmente, proceder a uma análise mais específica referente às notícias mais relevantes que ocorreram ao longo do ano 2022. A análise assim conduzida permitiu identificar possíveis padrões quanto ao modo como as vendas se comportam perante determinados acontecimentos e ajustar futuras tiragens. Paralelamente, e comparando os dados relativos ao final de 2022, concluiu-se do claro declínio de vendas e tiragens desde o início do ano de 2021.

Partindo para os modelos criados no presente trabalho, concluiu-se que a escolha do *Prophet* se revelou totalmente acertada, sendo possível constatar o seu excelente desempenho assim como a sua adaptabilidade tanto a variações sazonais como a *outliers*. No que toca à sua precisão, observa-se uma clara superioridade dos modelos multivariados com duas variáveis, que, em alguns casos, se sobrepõem de tal maneira que igualam os dados reais.

Assim, é possível afirmar que a busca de um melhor modelo de predição possível foi bastante bem-sucedida, permitindo à gestão de topo ter uma previsão bastante confiável dos resultados esperados, possibilitando, ainda, adicionar uma ferramenta de trabalho e análise bastante eficaz para o processo de tomada de decisão.

O referido modelo de predição revelou-se ainda bastante importante na predição de quantidades vendidas, permitindo, desta forma, à empresa ter um maior controlo sobre as quantidades que são, de facto, produzidas para serem expedidas e assim, diminuir os seus custos com matérias-primas, para além da otimização de vendas esperadas, o que, mais uma vez, traz vantagens para o planeamento estratégico da empresa a médio e a longo prazo.

De salientar ainda o *dashboard* desenvolvido de apoio à tomada de decisão que permite não só analisar tendências de evolução temporal e geográfica de vendas para os diferentes jornais e revistas, como também avaliar e tomar decisões baseadas na visualização de dados, através das necessidades empresariais e acompanhar os objetivos orçamentais da empresa. Para além do mais, o mesmo cumpriu os objetivos solicitados pela empresa.

Em resumo e em conclusão, este trabalho atingiu os objetivos a que se propôs, primeiramente na exploração de padrões, tendências e sazonalidades, evoluindo, posteriormente, para a criação de modelos sólidos para a predição de resultados fiáveis e concretos, tanto para as vendas como para as tiragens dos jornais em questão. Numa outra fase com a criação e implementação de um *dashboard* para apoio à tomada de decisão, cumprindo os requisitos propostos pela GMG.

5.2. Limitações e trabalho futuro do estudo

No que respeita às limitações mais notórias encontradas ao longo da dissertação, estas prenderam-se com a matriz estrutural existente para cada um dos jornais em estudo, apresentando bastantes limitações ao nível da organização e em variedade de dados, bem como a falta de mais variáveis relevantes que pudessem influenciar as vendas e tiragens de jornais. Além disso, a informação encontrava-se bastante dispersa e não apresentava uma base de dados para apoio à tomada de decisão através de dados.

Por fim, para trabalho futuro, propõe-se a investigação e o melhoramento referente aos modelos de previsão para o jornal DN, que apresentaram resultados inferiores aos restantes modelos, o que permitiria alcançar um panorama mais próximo do real, bem como uma melhor previsão de resultados.

Outro ponto para trabalho a realizar futuramente seria a integração dos modelos de previsão em um *dashboard* interativo que permitisse qualquer pessoa, sem experiência prévia, uma utilização simplificada e rápida.

Referências Bibliográficas

- [1] T. L. Quintanilha, «2029 - The end of print newspapers in Portugal? A longitudinal study on the main performance indicators in Portugal's traditional print-newspaper industry | 2029 - o fim dos jornais em papel em Portugal? Um estudo longitudinal sobre os principais indicado», *Observatorio*, vol. 12, n. 3, pp. 138–155, 2018.
- [2] «APCT». <https://www.apct.pt/> (acedido 21 de Fevereiro de 2023).
- [3] B. Akbaş, D. Karahoca, A. Karahoca, e A. Güngör, «Predicting newspaper sales by using data mining techniques», em *ACM International Conference Proceeding Series*, 2014, pp. 158–165. doi: 10.1145/2659532.2659630.
- [4] «Global Media Group -Global Media Group». <https://www.globalmediagroup.pt/> (acedido 21 de Fevereiro de 2023).
- [5] «Prophet | Forecasting at scale.» <https://facebook.github.io/prophet/> (acedido 21 de Fevereiro de 2023).
- [6] «Welcome to Python.org». <https://www.python.org/> (acedido 21 de Fevereiro de 2023).
- [7] «Visualização de Dados | Microsoft Power BI». <https://powerbi.microsoft.com/pt-pt/> (acedido 21 de Fevereiro de 2023).
- [8] «CRISP-DM - A Framework For Data Mining And Analysis». <https://thinkinsights.net/data/crisp-dm/> (acedido 21 de Fevereiro de 2023).
- [9] (Ncr e J. Clinton, «CRISP-DM 1.0 Step-by-step data mining guide», 1999.
- [10] F. Martínez-Plumed *et al.*, «CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories», doi: 10.1109/TKDE.2019.2962680.
- [11] O. Dogan e O. F. Gurcan, *Realising Newspaper Sales by Using Statistic Methods*, vol. 1197 AISC. 2021. doi: 10.1007/978-3-030-51156-2_101.
- [12] G. Incesu, B. Asikgil, e M. Tez, «Sales forecasting system for newspaper distribution companies in Turkey», *Pakistan Journal of Statistics and Operation Research*, vol. 8, n. 3, pp. 685–699, 2012, doi: 10.18187/pjsor.v8i3.539.
- [13] D. Kovačević, A. Poša, N. Pribačić, R. AntoniĆ, e A. Kovačević, «Modeling human operator and process in different environments», em *Proceedings of the International Conference on Information Technology Interfaces, ITI*, 2009, pp. 299–304. doi: 10.1109/ITI.2009.5196098.
- [14] A. S. Fakharudin, M. A. M. Hamza, e M. U. Johan, «Newspaper vendor sales prediction using artificial neural networks», em *2009 International Conference on Education Technology and Computer, ICETC 2009*, 2009, pp. 339–343. doi: 10.1109/ICETC.2009.67.
- [15] B. Bakker e T. Heskes, «Task clustering and gating for bayesian multitask learning», *Journal of Machine Learning Research*, vol. 4, n. 1, pp. 83–99, 2004, doi: 10.1162/153244304322765658.

- [16] T. Heskes, J.-J. Spanjers, B. Bakker, e W. Wiegerinck, «Optimising newspaper sales using neural-Bayesian technology», *Neural Comput Appl*, vol. 12, n. 3–4, pp. 212–219, 2003, doi: 10.1007/s00521-003-0384-x.
- [17] T. Ragg, *Building committees by clustering models based on pairwise similarity values*, vol. 2167. 2001. doi: 10.1007/3-540-44795-4_35.
- [18] R. Akbar, Wahyudi, J. Rahmadoni, e H. A. Lubis, «The Implementation of Business Intelligence on Visualisation of Transaction Data Analysis using Dashboard System Case Study: XYZ Convenience Store», em *Proceeding - 2022 International Symposium on Information Technology and Digital Innovation: Technology Innovation During Pandemic, ISITDI 2022*, 2022, pp. 42–49. doi: 10.1109/ISITDI55734.2022.9944494.
- [19] E. H. Sharaf Addin, N. Admodisastro, S. N. S. Mohd Ashri, A. Kamaruddin, e Y. C. Chong, «Customer Mobile Behavioral Segmentation and Analysis in Telecom Using Machine Learning», *Applied Artificial Intelligence*, vol. 36, n. 1, 2022, doi: 10.1080/08839514.2021.2009223.
- [20] P. P. Ramadhani, S. Hadi, e R. Rosadi, «Implementation of Data Warehouse in Making Business Intelligence Dashboard Development Using PostgreSQL Database and Kimball Lifecycle Method», em *2021 International Conference on Artificial Intelligence and Big Data Analytics, ICAIBDA 2021*, 2021, pp. 88–92. doi: 10.1109/ICAIBDA53487.2021.9689697.
- [21] R. Akbar, M. Silvana, M. H. Hersyah, e M. Jannah, «Implementation of Business Intelligence for Sales Data Management Using Interactive Dashboard Visualization in XYZ Stores», em *2020 International Conference on Information Technology Systems and Innovation, ICITSI 2020 - Proceedings*, 2020, pp. 242–249. doi: 10.1109/ICITSI50517.2020.9264984.
- [22] A. M. Purnamasari, C. E. A. Pah, M. D. I. Yoga, A. S. Girsang, e S. M. Isa, «Business Intelligent in an E-Commerce Industry», em *IOP Conference Series: Materials Science and Engineering*, 2019. doi: 10.1088/1757-899X/598/1/012085.
- [23] M. J. Page *et al.*, «The PRISMA 2020 statement: an updated guideline for reporting systematic reviews», doi: 10.1136/bmj.n71.
- [24] «NumPy». <https://numpy.org/> (acedido 30 de Junho de 2023).
- [25] «pandas - Python Data Analysis Library». <https://pandas.pydata.org/> (acedido 30 de Junho de 2023).
- [26] «seaborn: statistical data visualization — seaborn 0.12.2 documentation». <https://seaborn.pydata.org/> (acedido 30 de Junho de 2023).
- [27] S. J. Taylor e B. Letham, «Forecasting at Scale», <https://doi.org/10.1080/00031305.2017.1380080>, vol. 72, n. 1, pp. 37–45, Jan. 2018, doi: 10.1080/00031305.2017.1380080.