Lisbon School of Economics & Management
Universidade de Lisboa

Mestrado em

Métodos Quantitativos para a Decisão Económica e Empresarial

Projeto

Previsão das Vendas de um Seguro de Saúde: Comparação entre Modelos Estatísticos e de *Machine Learning*

Ana Margarida Rodrigues de Jesus

Orientação: Professor Doutor Jorge Caiado

DOCUMENTO ESPECIALMENTE ELABORADO PARA OBTENÇÃO DO GRAU DE MESTRE

JULHO 2025

AGRADECIMENTOS

Gostaria de agradecer ao meu orientador, Professor Doutor Jorge Caiado pela ajuda, sugestões e boa orientação dada ao longo do processo da escrita do trabalho final de mestrado.

À minha excelente equipa na companhia Generali Tranquilidade por todo o apoio, compreensão e ajuda, em particular ao Jorge Rosa, Bruno Santos e Cristina Matos, que me permitiram desenvolver este projeto na empresa e me deram liberdade para escolher fazer aquilo que fosse o melhor para o meu projeto académico.

À minha família e amigos, por todo o apoio.

Ao meu irmão André, por estar sempre comigo.

GLOSSÁRIO

PIB - Produto Interno Bruto

IPC - Índice de Preços do Consumidor

SNS - Serviço Nacional de Saúde

ASF - Autoridade de Supervisão de Seguros e Fundos de Pensões

ARIMA - *Autoregressive Integrated Moving Average*

ARIMAX - *Autoregressive Integrated Moving Average Exogenous*

XGBOOST - *eXtreme Gradient Boosting*

EAM - Erro Absoluto Médio

EPAM - Erro Percentual Absoluto Médio

EQM - Erro Quadrático Médio

REQM - Raiz Erro Quadrático Médio

FAC - Função de autocorrelação

FACP - Função de autocorrelação parcial

ARMA - *Autoregressive and Moving Average*

ADF - *Augmented* Dickey-Fuller

HWA - Holt-Winters Aditivo

HWM - Holt-Winters Multiplicativo

AIC - Critério de Informação de Akaike

BIC - Critério de Informação Bayesiano

HQIC - Critério de Informação Hannan–Quinn

LASSO - Least Absolute Shrinkage and Selection Operator

ML - *Machine Learning*

RESUMO

O presente estudo teve como objetivo principal avaliar diferentes metodologias para a previsão das vendas de um seguro de saúde, comparando modelos clássicos de séries temporais, como o Holt-Winters e o ARIMA, com abordagens baseadas em aprendizagem automática (*Machine Learning*), como o XGBoost e o Random Forest. Para enriquecer a capacidade explicativa dos modelos, foram incorporadas variáveis exógenas económicas e relacionadas com o setor da saúde.

O tratamento de observações anómalas (*outliers*) constituiu uma etapa metodológica central. Os valores atípicos foram identificados e corrigidos com base em critérios estatísticos, sendo avaliado o impacto dessa correção no desempenho preditivo dos modelos, comparando os valores das séries original e prevista. Nos modelos de *Machine Learning*, como o XGBoost, criaram-se desfasamentos da série temporal em estudo de modo a incorporar a dependência temporal nos algoritmos não sequenciais. A validação considerou a separação temporal entre as amostras de treino e teste, complementada por validação cruzada temporal sempre que aplicável.

Os resultados demonstraram que o modelo com melhor desempenho foi o modelo ARIMAX, com as variáveis exógenas "IPC total - taxa de variação homóloga" e "Nº de pessoas desempregadas", aplicado à série com *outliers* corrigidos. Este modelo beneficiou da integração de fatores macroeconómicos e do tratamento prévio de valores atípicos para melhorar a capacidade preditiva.

Os modelos de *Machine Learning*, como o Random Forest e o XGBoost, apresentaram desempenho competitivo, mas inferior ao ARIMAX, especialmente quando não houve correção de *outliers*. O modelo Holt-Winters teve o desempenho mais fraco, refletindo limitações em séries com dinâmicas mais complexas.

Estes resultados reforçam a importância do tratamento adequado dos dados e da combinação de diferentes técnicas para obter previsões mais robustas e confiáveis.

PALAVRAS-CHAVE:

Machine Learning, Séries Temporais; Vendas; Seguro de Saúde; Previsão

ABSTRACT

The main objective of this study was to evaluate different methodologies for forecasting health insurance sales, comparing classical time series models, such as Holt-Winters and ARIMA, with approaches based on Machine Learning, such as XGBoost and Random Forest. To enrich the explanatory power of the models, exogenous economic and health-related variables were incorporated.

The treatment of *outliers* constituted a central methodological step. Atypical points were identified and smoothed based on statistical criteria, and the impact of this correction on the performance of the models was assessed by comparing the original and forecasted series. In the Machine Learning models, *lagged* variables were created to incorporate time dependence in non-sequential algorithms, such as XGBoost. Validation considered the temporal separation between training and testing, complemented by temporal cross-validation whenever applicable.

The results showed that the best performing model was the ARIMAX model, with the exogenous variables "Total CPI – annual rate of change" and "Nº of unemployed people", applied to the series with corrected *outliers*. This model highlighted the relevance of integrating macroeconomic factors and pre-treating *outliers* to improve predictive accuracy.

Machine learning models, such as Random Forest and XGBoost, performed competitively but inferiorly to ARIMAX, especially when there was no outlier correction. The Holt-Winters model performed the weakest, reflecting limitations in series with more complex dynamics.

These results reinforce the importance of adequate data processing and combining different techniques to obtain more robust and reliable forecasts.

KEYWORDS:

Machine Learning; Time Series; Sales; Health Insurance; Forecast

Indíce

LISTA DE TABELAS

LISTA DE FIGURAS

1. INTRODUÇÃO

1.1. Enquadramento

Este projeto surge no âmbito do trabalho final do mestrado em Métodos Quantitativos para a Decisão Económica e Empresarial, realizado no Instituto Superior de Economia e Gestão (ISEG) da Universidade de Lisboa. O desenvolvimento do projeto decorreu numa empresa do setor segurador, a Generali Tranquilidade.

A Generali Tranquilidade é uma das principais seguradoras em Portugal, com uma longa trajetória que remonta a 1871, quando foi fundada como Companhia de Seguros Tranquilidade Portuense. Ao longo do tempo, a empresa passou por várias mudanças de propriedade e processos de fusão. Em 2020, foi adquirida pelo grupo italiano Assicurazioni Generali e, desde 2024, opera sob a designação Generali Tranquilidade. A companhia oferece uma ampla variedade de produtos, incluindo seguros automóvel, saúde, habitação, vida e acidentes de trabalho, destacando-se pela aposta na inovação e na digitalização dos seus serviços. Com uma quota de mercado aproximada de 14%, a Generali Tranquilidade é atualmente uma das maiores seguradoras em Portugal, evidenciando-se particularmente no ramo automóvel.

1.2. Objetivos do Trabalho

O presente trabalho tem como principal objetivo avaliar diferentes metodologias para a previsão das vendas - isto é, do número de novas apólices emitidas - de um produto estratégico para a empresa: o seguro de saúde. Um seguro de saúde é um tipo de seguro que se destina a cobrir os custos médicos incorridos por indivíduos ou grupos. Este tipo de seguro permite cobrir uma variedade de despesas médicas, incluindo estadias hospitalares, medicamentos prescritos e consultas médicas.

Atualmente, na Generali Tranquilidade, realiza-se todos os anos uma projeção das vendas, que é essencial para definir metas e incentivos para as equipas comerciais, para o planeamento financeiro da empresa, para monitorizar o crescimento e para validar as estratégias implementadas. Esta projeção é atualmente realizada através de Excel, utilizando fórmulas complexas e detalhadas que exigem revisões periódicas por parte dos

colaboradores, consumindo tempo e recursos. Nesse sentido, surge a necessidade de desenvolver um modelo preditivo mais eficiente e automatizado, baseado em técnicas estatísticas e de análise preditiva, que permita simplificar o processo e aumentar a robustez das previsões.

Neste estudo foi feita uma comparação entre três modelos clássicos de previsão de séries temporais (Holt-Winters, ARIMA e ARIMAX) e dois modelos de *Machine Learning* supervisionado (XGBoost e Random Forest), para avaliar a sua eficácia na previsão das vendas no setor de seguros, considerando diferentes combinações de variáveis e estruturas dos dados. Para a implementação dos modelos e análise de dados, foi utilizada a linguagem de programação Python, recorrendo a bibliotecas como o pandas, o numpy, o statsmodels, o scikit-learn, o xgboost, o matplotlib, seaborn, entre outras.

1.3. Estrutura

Este trabalho está organizado em cinco capítulos. No primeiro, é feita a introdução ao tema, são definidos os objetivos do estudo e as motivações que justificam a escolha do mesmo, assim como a estrutura do trabalho. O segundo capítulo consiste na revisão da literatura, onde se fundamenta teoricamente as metodologias aplicadas e se destaca estudos relevantes que demonstram a eficácia destas abordagens na área seguradora. No terceiro, descreve-se a metodologia adotada, detalhando os dados utilizados, o seu tratamento e as técnicas aplicadas. Além disso, são explicados os critérios de avaliação e validação dos modelos, que medem a qualidade das previsões. O quarto capítulo apresenta a aplicação prática dos modelos, analisando os resultados obtidos e comparando o desempenho das diferentes abordagens. Discute-se a precisão das previsões e a influência das variáveis consideradas. O quinto capítulo sumariza as conclusões do estudo, evidenciando o modelo mais adequado para a previsão das vendas.

Na conclusão, apontam-se as limitações e são sugeridas linhas para trabalhos futuros, com vista à melhoria contínua das técnicas de previsão neste contexto. Por fim, são apresentadas as referências bibliográficas que sustentam o desenvolvimento do projeto.

2. REVISÃO DE LITERATURA

A previsão de vendas no setor de seguros de saúde, é uma área de investigação fundamental para as seguradoras, dada a sua relevância para a gestão de riscos, definição de políticas comerciais e sustentabilidade financeira.

Os modelos de *Deep Learning* emergiram como ferramentas poderosas para a previsão de séries temporais complexas, como as vendas no setor de seguros. Entre estes, os modelos de redes neuronais recorrentes, como o Long Short-Term Memory (LSTM), destacam-se pela sua capacidade de capturar dependências temporais de longo prazo nas sequências de dados e captar não linearidades. Por exemplo, Diao e Wang (2019) demonstraram que um modelo LSTM aplicado à previsão da receita mensal de prémios entre 1999 e 2019 superou os modelos de redes neuronais tradicionais como as redes *backpropagation*, comprovando a sua eficácia neste domínio.

Dash et al. (2024) apresentam uma análise comparativa de diferentes técnicas de *Machine Learning* para prever prémios de seguros de saúde, utilizando um conjunto de dados com variáveis como idade, género, IMC, hábitos tabágicos e localização geográfica. Os autores aplicaram vários modelos de regressão, incluindo o *Gradient Boosting*, o *Random Forest*, o *Decision Tree*, o *XGBoost* e a Regressão Linear, destacando-se o *Gradient Boosting* como o mais preciso, com uma precisão de 88%. Os resultados indicam que a idade, o IMC e os hábitos tabágicos são os fatores mais influentes na determinação dos prémios. O estudo demonstra a utilidade destes modelos para seguradoras e clientes, permitindo preços mais justos e decisões informadas, com erros médios absolutos inferiores a 5%.

Sijie et al. (2024) propõe um método de *ensemble learning* baseado em *stacking* para prever prémios de seguros de saúde, combinando Redes Neurais Artificiais (ANN) com outros modelos como o XGBoost, o Decision Tree, o Random Forest e Support Vector Machine (SVM). Utilizando o conjunto de dados *US Health Insurance Dataset* (1338 registos com variáveis como a idade, o IMC, os hábitos tabágicos e a região), os autores demonstraram que os modelos *ensemble* superam os modelos individuais, destacando-se o modelo híbrido ANN+Random Forest com os melhores resultados em termos de qualidade preditiva. O estudo conclui que a abordagem de *stacking* melhora a precisão ao

aproveitar as vantagens complementares dos modelos base, sendo promissora para aplicações no setor de seguros.

Orji e Ukwandu (2024) propõem uma abordagem de *Machine Learning* para prever custos de seguros de saúde, utilizando três modelos de *ensemble learning*: XGBoost, Gradient Boosting Machine e Random Forest. O estudo emprega técnicas de Explainable AI (XAI), como SHAP (SHapley Additive exPlanations) e ICE (Individual Conditional Expectation), para identificar e interpretar os fatores determinantes dos preços dos prémios. O conjunto de dados utilizado, disponível no repositório Kaggle, contém 986 registos com variáveis como a idade, o IMC, o histórico médico e os hábitos de saúde. Os resultados mostram que todos os modelos alcançaram um bom desempenho, com o XGBoost a destacar-se, mas consumindo mais recursos computacionais. O Random Forest obteve menor erro absoluto médio e foi o mais eficiente em termos computacionais. As análises SHAP e ICE revelaram que a idade, o IMC, as doenças crónicas e o histórico familiar de cancro foram os fatores mais influentes nos custos, enquanto os diabetes e as alergias conhecidas tiveram impacto mínimo. Além disso, um número elevado de cirurgias principais paradoxalmente reduziu os preços dos prémios.

Os modelos tradicionais de séries temporais, como o Holt-Winters e o ARIMA, continuam a ser amplamente utilizados na previsão de vendas devido à sua robustez, simplicidade e interpretabilidade. Veiga et al. (2024) aplicam o modelo ARIMA para prever cinco séries temporais (PIB, IPCA, taxa de desemprego, número total de beneficiários de planos de saúde e número de beneficiários de planos individuais) no contexto das crises económicas brasileiras entre 2000 e 2020. O estudo mostra a eficácia dos modelos de previsão empregues, com precisão superior a 95% em diversas séries, e destaca o seu valor estratégico para os gestores da saúde pública e privada, especialmente em períodos de recessão, como a crise de desemprego de 2019. O artigo demonstra que o modelo ARIMA, apesar de ser um modelo linear e relativamente simples, oferece previsões robustas e aplicáveis a contextos reais, sendo útil para o planeamento do orçamento e a tomada de decisões em contextos de flutuações económicas.

Masih et al. (2024) propõem uma análise preditiva dos prémios de seguros de saúde na Índia entre 2005 e 2023 utilizando modelos ARIMA. Os modelos ARIMA(0,2,0) e

ARIMA(0,2,1) foram os que melhor se ajustaram aos dados da série em estudo. Os resultados obtidos revelaram que estes modelos demonstraram elevada precisão e fiabilidade na previsão dos prémios, com valores de EPAM abaixo de 10%.

Ulyah et al. (2019) investigaram a precisão de diferentes abordagens de modelação na previsão das provisões técnicas associadas a sinistros em seguros de educação. O principal objetivo foi comparar o desempenho de um modelo de séries temporais com variáveis exógenas e sazonalidade com um método de regressão não paramétrica mais flexível. Para tal, os autores compararam dois modelos de previsão: um modelo SARIMAX e uma regressão não paramétrica baseada em séries de Fourier. A análise foi aplicada a dados mensais, tendo a variável exógena ('t') sido utilizada no modelo SARIMAX para captar a tendência crescente observada na série. Os resultados revelaram que o modelo SARIMAX superou o modelo não paramétrico em termos de precisão, com um EPAM de 4.03%. O estudo sublinha a capacidade dos modelos ARIMAX para integrar fatores sazonais e tendências externas de forma eficaz, sendo uma ferramenta relevante para a gestão financeira e orçamental em seguros com padrões temporais complexos.

Neel (2025), realizou uma análise comparativa aprofundada para prever prémios de seguro de vida. O principal objetivo deste trabalho foi avaliar o desempenho de diferentes modelos de séries temporais na previsão deste tipo de receitas no setor de seguros. Para tal, foram comparados os seguintes modelos: ETS (Exponential Smoothing State Space Model), Holt-Winters, NNETAR (Neural Network Time Series Forecast) e TBATS (Trigonometric, Box-Cox, ARMA Errors, Trend, and Seasonal components). Os resultados obtidos indicaram que, entre as abordagens testadas, o método multiplicativo de Holt-Winters demonstrou ser o mais preciso, alcançando uma impressionante taxa de exatidão de 97,56%.

Embora grande parte da literatura referida se foque na previsão de prémios, e não diretamente nas vendas, a sua inclusão nesta revisão justifica-se por dois motivos. Em primeiro lugar, todos os estudos apresentados se inserem no contexto do setor segurador e demonstram a aplicabilidade e eficácia de diferentes modelos de séries temporais e algoritmos de *Machine Learning* em problemas preditivos com elevada complexidade. Em segundo lugar, a escassez de investigações especificamente dedicadas à previsão de

vendas no setor dos seguros - nomeadamente ao número de novas apólices emitidas - justifica a incorporação de estudos adjacentes que, embora com objetivos distintos, contribuem para a validação da abordagem metodológica adotada neste trabalho.

Com base na literatura analisada, que demonstra a eficácia dos modelos de séries temporais tradicionais como o Holt-Winters e o ARIMA, bem como a crescente aplicação de modelos com variáveis exógenas (como o ARIMAX) e de algoritmos de Machine Learning como o XGBoost e o Random Forest, optou-se por aplicar e comparar estes cinco modelos no presente estudo.

3. METODOLOGIA

3.1 Decomposição clássica das forças componentes

A análise de séries temporais é uma ferramenta estatística fundamental para compreender, modelar e prever dados que evoluem ao longo do tempo. Um passo essencial nesta análise é a decomposição da série em componentes estruturais que facilitam a interpretação e previsão dos dados. A literatura destaca dois modelos clássicos para esta decomposição: o modelo aditivo e o modelo multiplicativo, como referido por Caiado (2022). No modelo aditivo, assume-se que a série temporal Y_t é composta pela soma linear de três componentes: tendência (T_t), sazonalidade (S_t) e ruído ou resíduo (R_t), formalmente expressa como:

$$Y_t = T_t + S_t + R_t$$

Este modelo é apropriado quando a variabilidade da série é constante ao longo do tempo, ou seja, quando a amplitude da sazonalidade não depende do nível da série. É indicado para séries onde os efeitos sazonais mantêm uma magnitude aproximadamente constante independentemente do valor da tendência, conforme descrito por Hyndman e Athanasopoulos (2021) e por Caiado (2022). Por outro lado, o modelo multiplicativo considera que as componentes interagem de forma multiplicativa, de acordo com a seguinte especificação:

$$Y_t = T_t \times S_t \times R_t$$

Este modelo é preferível quando a amplitude da sazonalidade varia em função do nível da série, ou seja, o efeito sazonal é maior/menor quando a série está em valores mais elevados/baixos. Em muitos contextos económicos e financeiros, onde a volatilidade da série aumenta com o valor da tendência, o modelo multiplicativo é mais adequado.

A tendência (T_t) reflete o comportamento de longo prazo da série, capturando padrões de crescimento, declínio ou estabilidade ao longo do tempo. Estimá-la corretamente é crucial para identificar mudanças estruturais e orientar previsões. Os métodos comuns para estimar a tendência incluem as médias móveis, o alisamento exponencial e a decomposição da série, usando, por exemplo, o método STL (Seasonal-

Trend decomposition using Loess) — uma técnica robusta e flexível, proposta por Cleveland et al. (1990).

A sazonalidade (S_t) corresponde a padrões que se repetem em intervalos regulares, como ciclos mensais, trimestrais ou anuais. A correta identificação e extração desta componente é vital para evitar enviesamentos em modelos de previsão e permitir que as decisões estratégicas contemplem variações sazonais.

Por fim, o ruído ou resíduo (R_t) representa a parte da série que não é explicada pelos componentes de tendência e sazonalidade. Idealmente, deve comportar-se como um processo estacionário, com média zero e variância constante. A análise dos resíduos é essencial para validar o modelo, identificar *outliers*, e ajustar o modelo conforme necessário.

Esta estruturação detalhada dos dados é fundamental para o desenvolvimento de modelos preditivos eficazes, pois permite compreender e modelar melhor os comportamentos e padrões temporais intrínsecos, contribuindo para a melhoria da precisão das previsões.

3.2 Identificação e Tratamento de Outliers

A presença de *outliers* em séries temporais pode comprometer significativamente a qualidade da análise estatística, afetando a estimação dos parâmetros, a identificação de padrões e, sobretudo, a capacidade preditiva dos modelos. Por este motivo, a deteção e o tratamento adequado de valores atípicos constitui uma etapa crítica no pré-processamento de dados temporais.

Neste estudo, os *outliers* foram identificados com base na análise dos resíduos da série, após a decomposição nas suas componentes estruturais — tendência, sazonalidade e ruído. Adotou-se como critério para a deteção de um *outlier* os valores situados fora do intervalo definido por dois desvios padrão acima ou abaixo da média dos resíduos. Uma vez identificados, os valores atípicos foram corrigidos por substituição pelos valores reconstruídos a partir da soma da tendência e da sazonalidade estimadas para o instante temporal correspondente, como sugerido por Caiado (2022). Esta abordagem visa

preservar a estrutura determinística da série - removendo apenas o ruído anómalo - e minimizar o impacto dos *outliers* na modelação subsequente. O procedimento foi aplicado de forma seletiva, exclusivamente nos pontos identificados como anómalos, assegurando assim a integridade global da série.

3.3 Amostra de treino e amostra de teste

Para garantir uma avaliação rigorosa da capacidade preditiva dos modelos de séries temporais desenvolvidos, a base de dados foi segmentada em dois subconjuntos cronologicamente ordenados: a amostra de treino e a amostra de teste.

No contexto de séries temporais, é essencial preservar a ordem temporal dos dados ao realizar qualquer processo de validação. Técnicas tradicionais de validação cruzada, como o *k-folds cross-validation*, que envolvem aleatorização das observações, são inadequadas, pois violam a dependência sequencial inerente à série. Como destacado por Hyndman e Athanasopoulos (2021), a técnica recomendada de validação é a *time series cross-validation*, onde cada conjunto de teste contém uma única observação posterior a todo o conjunto de treino, que inclui apenas observações anteriores (ou seja, sem "ver o futuro"). Esta abordagem evita que o modelo utilize informações futuras durante o treino, prevenindo o fenómeno conhecido como *data leakage*, que provoca avaliações excessivamente otimistas e compromete a validade das previsões.

Assim, optou-se por uma divisão temporal estrita, em que os modelos foram treinados exclusivamente com dados históricos mais antigos e testados apenas em observações conhecidas mais recentes. Neste estudo, foram utilizadas 5 *folds* quando a técnica *time series cross-validation* foi utilizada.

A amostra de treino compreendeu as primeiras 96 observações da série, correspondentes ao período de janeiro de 2016 a dezembro de 2023, servindo de base para a estimação dos parâmetros e ajustamento dos modelos. A amostra de teste, por sua vez, incluiu as 12 observações seguintes — de janeiro a dezembro de 2024 — sendo utilizada para avaliar a capacidade de generalização dos modelos em previsões *out-of-sample*. A escolha de um horizonte de 12 meses para o teste justifica-se pela prática comum de previsão anual no contexto empresarial e financeiro, além de permitir captar padrões sazonais completos.

3.4 Erros de previsão

A avaliação da qualidade das previsões geradas por modelos de séries temporais exige a utilização de métricas específicas que quantifiquem a discrepância entre os valores observados e os valores previstos. Entre as métricas mais utilizadas encontram-se o Erro Absoluto Médio (EAM), o Erro Quadrático Médio (*EQM*), a Raiz do Erro Quadrático Médio (*REQM*) e o Erro Percentual Absoluto Médio (*EPAM*).

O EAM representa a média dos erros absolutos e tem a vantagem de ser intuitivo, pois mantém a unidade da variável. As medidas do EQM e do REQM, por sua vez, penalizam de forma mais severa os erros grandes, uma vez que elevam ao quadrado as diferenças, o que as tornam mais sensíveis a *outliers*. Neste estudo, optou-se por utilizar o EPAM como principal métrica de avaliação. O EPAM mede o erro absoluto médio em termos relativos, expressando-o como uma percentagem dos valores observados. A fórmula do EPAM é dada por:

$$EPAM(\%) = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100,$$

em que y_t representa o valor observado no instante t, $\hat{y}_t$ o valor previsto e n o número total de previsões avaliadas. O valor resultante expressa, em média, o quão longe as previsões estão dos valores reais, em termos percentuais.

Apesar de o EPAM apresentar limitações quando os valores observados se aproximam de zero, esse problema não se verifica neste estudo, dado que todas os valores assumem valores estritamente positivos ao longo da série. Assim, a escolha do EPAM justifica-se pela sua clareza interpretativa, simplicidade computacional e adequação prática. Estas medidas de erro referidas, e outras que também poderiam ter sido utlizadas, estão explicadas no artigo de Plevris et al. (2022), onde se apresentam algumas vantagens e desvantagens de cada.

3.5 Método de Holt-Winters

De acordo com Ostertagová e Ostertag (2013), os métodos de alisamento exponencial constituem uma família de técnicas de previsão amplamente utilizadas na

modelação de séries temporais. A principal característica destes métodos reside no facto de atribuírem pesos decrescentes exponencialmente aos valores passados, permitindo que as observações mais recentes tenham maior influência na previsão. Esta abordagem torna os modelos particularmente responsivos a alterações recentes nos padrões da série, mantendo, simultaneamente, uma estrutura computacional simples.

Existem várias formas de alisamento exponencial, desde o método simples (*Single Exponential Smoothing*), adequado para séries sem tendência e sem sazonalidade, até métodos mais complexos como o método de Holt, que incorpora uma componente de tendência linear, e ainda o método de Holt-Winters, que estende esta lógica para séries com comportamento sazonal (Caiado, 2022). A escolha do método mais adequado depende da estrutura da série temporal em análise.

Neste estudo, serão exploradas duas variantes do método Holt-Winters: o modelo aditivo e o modelo multiplicativo. Ambos os métodos são adequados para séries temporais que apresentam tendência e sazonalidade, mas diferem na forma como tratam a variação da amplitude sazonal.

O método de Holt-Winters aditivo assume que os componentes da série (tendência, sazonalidade e ruído) se combinam de forma linear. Este modelo é particularmente apropriado quando a amplitude da componente sazonal é constante ao longo do tempo, ou seja, quando a magnitude das variações sazonais não depende do nível da série.

O método de Holt-Winters aditivo é definido pelas seguintes equações de atualização:

$$a_t = \alpha(Y_t - S_{t-s}) + (1-\alpha)(a_{t-1} + b_{t-1}), \qquad 0 < \alpha < 1$$
$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}, \qquad 0 < \beta < 1$$
$$S_t = \gamma(Y_t - a_t) + (1-\gamma)S_{t-s}, \qquad 0 < \gamma < 1$$

Nestas equações, a_t expressa o nível da série no momento t; b_t expressa a tendência (ou declive) no momento t; S_t refere-se à componente sazonal da série e s denota o período sazonal (Caiado, 2022).

As previsões a h passos à frente são obtidas através da seguinte expressão:

$$P_{t+h} = a_t + hb_t + S_{t+h-s}, h = 1,2,3,\dots$$

Por outro lado, o método de Holt-Winters multiplicativo assume que as componentes interagem de forma proporcional. Neste caso, a sazonalidade varia proporcionalmente ao

nível da série, tornando este método mais adequado quando se observa que os picos e vales sazonais aumentam ou diminuem com o valor médio da série. Na forma multiplicativa, o método de Holt-Winters é definido pelas seguintes equações:

$$a_t = \alpha(Y_t/S_{t-s}) + (1-\alpha)(a_{t-1} + b_{t-1}), \qquad 0 < \alpha < 1$$

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}, \qquad 0 < \beta < 1$$

$$S_t = \gamma(Y_t/a_t) + (1-\gamma)S_{t-s}, \qquad 0 < \gamma < 1$$

As previsões a h passos à frente são obtidas através da seguinte expressão:

$$P_{t+h} = (a_t + hb_t)S_{t+h-s}, h = 1,2,3, \ldots$$

A estimação dos parâmetros dos métodos de alisamento de Holt-Winters (aditivo e multiplicativo), incluindo os fatores de alisamento do nível (α), da tendência (β) e da sazonalidade (γ), é realizada geralmente por métodos iterativos de otimização, com o objetivo de minimizar uma função de erro preditivo, como o erro quadrático médio ou o erro percentual absoluto médio.

A escolha entre o modelo aditivo e o multiplicativo deve basear-se numa análise prévia da série temporal, em particular na variabilidade da componente sazonal ao longo do tempo. Esta análise pode ser feita através da decomposição da série, comparação visual de gráficos e avaliação da estabilidade da amplitude sazonal.

3.6 Modelos ARIMA

A modelação com modelos ARIMA constitui uma abordagem amplamente utilizada na análise e previsão de séries temporais, por permitir captar a dependência estocástica entre observações ao longo do tempo. Segundo Kaur et al. (2023, p. 19618), *"o modelo ARIMA é o mais amplamente aceite para séries temporais"*, sendo aplicado em diversos domínios devido à sua elevada precisão matemática, natureza flexível e resultados fiáveis.

O modelo ARIMA combina três componentes essenciais: um termo autorregressivo (AR), um termo de média móvel (MA) e uma componente de integração (I), associada à diferenciação da série temporal. Este modelo univariado tem como base os modelos ARMA, os quais pressupõem a estacionariedade da série temporal para uma

aplicação válida. Uma série estacionária é uma série cujas propriedades estatísticas são independentes do tempo, onde se verifica uma média e variância constantes e a covariância é independente do tempo, como descrito por Stancu et al. (2017).

Dado que muitas séries temporais observadas, na prática, apresentam características de não estacionariedade, a introdução de um operador de diferenciação permite transformar a série original numa série estacionária. Esta extensão dá origem aos modelos ARIMA, que se revelam mais robustos e aplicáveis a um leque mais amplo de séries. A estrutura geral dos modelos ARIMA é definida pela seguinte expressão:

$$\phi(B)(1-B)^d y_t = \theta(B)\varepsilon_t,$$

onde: y_t é valor da série temporal no tempo t; B é o operador atraso ou de desfasamento (*lag*), tal que $B^k y_t = y_{t-k}$; d é a ordem da diferenciação necessária para tornar a série estacionária; $\phi(B)$ é polinómio autoregressivo (AR) de ordem p, da forma $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$; $\theta(B)$ é o polinómio de médias móveis (MA) de ordem q, da forma $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$; e ε_t é termo de erro aleatório (ruído branco).

O termo $(1-B)^d y_t$ aplica uma diferenciação de ordem d à série temporal original, sendo responsável por eliminar tendências e tornar a série estacionária. A componente autoregressiva, representada pelo polinómio $\phi(B)$, modela a dependência linear entre o valor atual da série e os seus valores passados. Por outro lado, a componente de médias móveis, descrita por $\theta(B)$, capta a dependência dos valores atuais em relação aos erros (ou choques) passados. O termo de erro ε_t assume-se como ruído branco, ou seja, uma sequência de variáveis aleatórias independentes e identicamente distribuídas, com média zero e variância constante. A conjugação destas três componentes permite que os modelos ARIMA sejam suficientemente flexíveis para modelar e prever séries temporais não estacionárias, após a devida transformação por diferenciação.

A metodologia utilizada para a modelação ARIMA teve como base o estudo de Hyndman e Athanasopoulos (2021, capítulo 9.7). A primeira etapa da modelação envolveu a verificação da estacionariedade da série, condição essencial para a aplicação de modelos ARIMA. Para tal, utilizou-se o teste ADF, proposto por Dickey e Fuller (1979), cuja hipótese nula (H_0) assume a presença de uma raiz unitária (indicando não-

estacionariedade), sendo rejeitada em favor da alternativa (H_1) apenas se o valor-p for inferior ao nível de significância – neste estudo considerado 5%. Quando necessário, procedeu-se à diferenciação da série até alcançar estacionariedade. A diferenciação aplicada pode assumir duas formas: simples ou sazonal, dependendo da natureza da tendência observada na série. A diferenciação simples consiste em subtrair cada valor da série pelo valor imediatamente anterior (ordem 1), sendo apropriada quando a série apresenta uma tendência linear persistente ao longo do tempo. Por outro lado, quando a série evidencia padrões cíclicos ou flutuações sistemáticas que se repetem em intervalos regulares - como variações mensais, trimestrais ou anuais - recorre-se à diferenciação sazonal, que subtrai cada valor pelo valor correspondente no mesmo período do ciclo anterior (por exemplo, $Y_t - Y_{t-12}$ no caso de dados mensais com sazonalidade anual). Em alguns casos, pode ser necessário aplicar ambas as transformações (simples e sazonal) para eliminar completamente a não-estacionariedade da série.

Após a diferenciação, identificaram-se os valores apropriados dos parâmetros p e q com base na inspeção visual das funções de autocorrelação (FAC) e autocorrelação parcial (FACP). A FAC mede o grau de correlação linear entre os valores atuais de uma série temporal e os seus respetivos valores com desafasamento. Essa relação é quantificada pela seguinte expressão:

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_y-\bar{Y})(Y_{y-k}-\bar{Y})}{\sum_{t=1}^{n}(Y_y-\bar{Y})^2}, K = 1,2,\dots,n-1,$$

onde n representa o número total de observações da amostra e k a ordem do desfasamento.

Por outro lado, PACF avalia a correlação entre os valores atuais da série e os valores em determinado instante passado, removendo o efeito intermediário dos *lags* situados entre esses dois pontos. A sua estimativa é obtida de forma recursiva, segundo a equação abaixo, onde $p_{kk} = r_1$ (inicialização) e p_{kj} representa o coeficiente parcial na k-ésima regressão e para o j-ésimo *lag*.

$$p_{kk} = \frac{r_k - \sum_{j=1}^{k-1} p_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} p_{k-1,j} r_j}, k = 2,3,\dots$$

A função FAC é útil para identificar a ordem do termo MA, observando o ponto em que a autocorrelação deixa de ser estatisticamente significativa. A FACP, por sua vez,

auxilia na identificação da ordem do termo AR, pela observação de onde ocorre o corte brusco nos valores parciais da autocorrelação. A análise conjunta dos gráficos da FAC e da FACP é essencial para a identificação dos parâmetros ideais p e q no modelo ARIMA, respetivamente associados aos componentes autoregressivos e de média móvel.

Considerando que a série temporal analisada é mensal e abrange um período de nove anos (108 observações), foram utilizados 36 *lags* nos gráficos das funções FAC e FACP. Esta escolha permite identificar não apenas padrões de autocorrelação de curto prazo, mas também potenciais padrões sazonais anuais (*lag* 12), bienais (*lag* 24) e trienais (*lag* 36). A aplicação de 36 *lags* revelou-se particularmente útil na análise dos resíduos dos modelos ARIMA ajustados, contribuindo para detetar a presença de autocorrelação sazonal residual e avaliar se a estrutura da série foi devidamente captada pelo modelo.

Após a análise gráfica, procedeu-se à estimativa de múltiplos modelos candidatos, com diferentes combinações dos parâmetros, cuja qualidade foi comparada com base em critérios de informação. Os critérios de informação mais comuns são o AIC, o BIC e o HQIC. Todos estes critérios procuram um equilíbrio entre a qualidade do ajustamento (verossimilhança) e a penalização pela complexidade do modelo (número de parâmetros estimados). Embora o BIC seja amplamente utilizado por penalizar mais fortemente a complexidade do modelo, tal característica pode ser excessiva em amostras pequenas ou moderadas, conduzindo à seleção de modelos demasiado parcimoniosos, que não capturam adequadamente a estrutura temporal dos dados. Por outro lado, o AIC tende a favorecer modelos mais complexos, podendo resultar em sobreajustamento (*overfitting*). Neste estudo, optou-se por adotar o HQIC como critério de seleção de modelos, uma vez que fornece uma penalização intermédia entre o AIC e o BIC. Segundo Hannan e Quinn (1979), o HQIC é particularmente adequado quando se procura consistência na seleção de modelos em amostras de tamanho limitado, como é o caso da presente aplicação. A expressão do HQIC é dada por:

$$HQIC = -2ln(\hat{L}) + 2kln(ln(n)),$$

onde $\hat{L}$ é o logaritmo de verossimilhança do modelo ajustado; k é o número de parâmetros estimados no modelo e n é o número de observações na amostra.

Uma vez selecionado o modelo com base no HQIC, o passo seguinte foi a estimação dos parâmetros e a validação estatística do modelo. Aqui, a análise de resíduos

desempenhou um papel central. Foi aplicado o teste de Ljung–Box para verificar a ausência de autocorrelação nos resíduos, sendo a estatística do teste definida por:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k}{n-k},$$

onde n é o número de observações, $\hat{\rho}_k$ é a autocorrelação dos resíduos no *lag* k, e h é o número de desfasamentos considerados.

Um valor-p no teste de Ljung-Box superior ao nível de significância (de 5% neste estudo) indica que não há evidências estatísticas de autocorrelação significativa nos resíduos, ou seja, não se rejeita a hipótese nula de que os resíduos se comportam como ruído branco. Isto sugere que o modelo ajustado é adequado para capturar a estrutura temporal da série. Valores-p próximos de um, são ainda mais indicativos da adequação do modelo, pois mostram que a estatística Q está muito abaixo do limiar crítico, confirmando que não há autocorrelação remanescente nos resíduos.

Adicionalmente, foram analisados os gráficos das funções FAC e FACP dos resíduos. A ausência de picos significativos reforça que os resíduos se comportam como ruído branco, um requisito essencial para a validade do modelo no contexto preditivo.

De forma complementar, foi avaliado o algoritmo auto.arima() da biblioteca pmdarima em Pyhton, que automatiza o processo de seleção de p e q com base na minimização do AIC.

Por fim, a avaliação da performance preditiva dos modelos foi realizada com base no EPAM, como referido anteriormente. Este conjunto de procedimentos metodológicos permite construir modelos ARIMA robustos e estatisticamente válidos.

3.7 Modelos ARIMAX

O modelo ARIMAX é uma generalização do modelo ARIMA que incorpora variáveis exógenas, permitindo captar efeitos externos que influenciam a variável dependente, potencialmente melhorando a capacidade preditiva e explicativa. A expressão geral do modelo ARIMAX(*p, d, q*) é dada por:

$$\phi(B)(1-B)^d y_t = \theta(B)\varepsilon_t + \beta(B)X_t$$

Nesta equação, X_t é o vetor de variáveis exógenas (observáveis no tempo t) e $\beta(B)$ é o vetor de coeficientes associado às variáveis exógenas.

Para a seleção das variáveis exógenas que contribuem para a melhoria do modelo ARIMAX, foram aplicadas três abordagens metodológicas complementares:

a) Método Stepwise com Critério de Informação HQIC

O procedimento Stepwise é um algoritmo iterativo que visa encontrar o subconjunto ótimo de variáveis explicativas entre todas as variáveis candidatas. A sua lógica assenta na avaliação sequencial de modelos, adicionando ou removendo variáveis com base na melhoria ou deterioração de um critério de informação. Desboulets (2018, p. 5) refere que alguns dos indicadores que podem ser usados são o R^2, o AIC o BIC, o HQIC, o erro de previsão e *leave-one-out cross validation.* Neste estudo foi utilizado o HQIC, de forma a encontrar um modelo parcimonioso que evite a inclusão de variáveis supérfluas. De acordo com Badshah e Bulut (2020), este método pode ser aplicado em três variantes principais: seleção *forward*, *backward* e o método combinado (bidirecional).

Na abordagem *forward*, o algoritmo começa por estimar o modelo ARIMA puro, sem qualquer variável exógena. A cada iteração, testa-se a inclusão de cada uma das variáveis ainda não presentes no modelo, calculando-se o HQIC resultante de cada inclusão. A variável que conduzir à maior redução do critério é então adicionada ao modelo. Este processo repete-se até que nenhuma nova variável consiga melhorar o HQIC, momento em que se dá por concluída a seleção.

Na abordagem *backward*, o ponto de partida é o modelo completo, que inclui todas as variáveis exógenas candidatas. Em cada iteração, elimina-se uma variável de cada vez e avalia-se o impacto da sua remoção no HQIC. A variável cuja exclusão resultar na maior melhoria do critério é eliminada. O processo continua até que a remoção de qualquer variável cause um aumento do HQIC.

O procedimento Stepwise Bidirecional combina etapas de inclusão (*forward selection*) e exclusão (*backward elimination*) de variáveis de forma iterativa. A cada passo, são estimados modelos com novas variáveis candidatas e, em seguida, os regressores já incluídos são reavaliados para possível exclusão, de modo a garantir um modelo parcimonioso e estatisticamente significativo. No método combinado, utilizado

neste estudo, o algoritmo iniciou-se com um modelo vazio, e em cada etapa avaliou-se tanto a possibilidade de incluir uma nova variável como a de eliminar uma já presente. Esta abordagem bidirecional permite maior flexibilidade, corrigindo decisões tomadas em iterações anteriores e maximizando as hipóteses de encontrar um modelo mais robusto.

b) Regressão LASSO

A regressão LASSO é uma técnica de regularização que adiciona uma penalização L_1 à função objetivo da regressão, forçando alguns coeficientes a serem exatamente zero e, assim, promovendo a seleção automática de variáveis, conforme descrito por Tibshirani (1996). Esta abordagem é particularmente útil quando existe um grande conjunto de variáveis candidatas e quando se pretende evitar modelos excessivamente complexos.

Especificamente, o modelo LASSO resolve o seguinte problema de minimização:

$$min_{(\beta)} \left\{ \sum_{t=1}^{n} \left(y_t - \beta^0 - \sum_{j=1}^{p} \beta_j x_{tj} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

onde y_t representa o valor da variável dependente no instante t, x_{tj} são os valores das variáveis independentes (ou candidatas a variáveis exógenas), β_j são os coeficientes a estimar e λ é o parâmetro de penalização que controla a magnitude da regularização.

O termo de penalização $\lambda \sum_{j=1}^{p} |\beta_j|$ promove o encolhimento (*shrinkage*) dos coeficientes, podendo forçar alguns a tornarem-se exatamente zero.

No presente estudo, a regressão LASSO foi aplicada utilizando validação temporal por meio do método *TimeSeriesSplit*, que divide a série temporal em várias janelas para treino e validação, respeitando a ordem cronológica dos dados.

O parâmetro de penalização λ foi ajustado de forma otimizada, minimizando o erro preditivo sobre os dados de validação. As variáveis com coeficientes $\beta_j = 0$ foram excluídas, enquanto as com $\beta_j \neq 0$ foram consideradas relevantes e incluídas na formulação dos modelos ARIMAX.

c) Avaliação Individual de Variáveis Exógenas com Modelos ARIMAX

Para complementar os métodos anteriores, procedeu-se à avaliação individual do impacto de cada variável exógena no desempenho do modelo. Para cada variável

candidata, ajustou-se um modelo ARIMAX contendo apenas essa variável exógena, mantendo a estrutura ARIMA previamente selecionada para a série endógena.

Cada modelo foi avaliado usando o HQIC e o EPAM obtido na amostra de teste. As variáveis que proporcionaram uma redução significativa destes indicadores em relação ao modelo ARIMA puro (sem variáveis exógenas) foram consideradas potencialmente informativas e merecedoras de inclusão no modelo final.

A combinação destas três abordagens visa garantir que as variáveis exógenas selecionadas apresentam robustez estatística, poder preditivo e não adicionam complexidade desnecessária ao modelo. Posteriormente, o modelo final ARIMAX foi avaliado quanto à qualidade do ajuste e à validade dos resíduos, incluindo a análise da autocorrelação dos resíduos para assegurar a adequação do modelo.

3.8 XGBoost

Machine Learning (aprendizagem automática) refere-se a um conjunto de métodos capazes de detetar automaticamente padrões nos dados e usá-los para prever resultados futuros ou tomar decisões em contextos de incerteza. Em contraste com os modelos estatísticos tradicionais, que normalmente partem de pressupostos explícitos sobre a estrutura dos dados, os algoritmos de *Machine Learning* são capazes de modelar relações complexas e não lineares, adaptando-se automaticamente ao comportamento observado.

No âmbito da previsão de séries temporais, os modelos de *Machine Learning* apresentam vantagens na capacidade de capturar interações não lineares e efeitos complexos, embora não incorporem, de forma nativa, a estrutura sequencial dos dados temporais. Para tal, é necessário o recurso a técnicas de engenharia de variáveis, como a criação de *lags*, que traduzem a dependência temporal em variáveis explicativas explícitas (Bontempi et al., 2013).

O modelo XGBoost é um algoritmo de aprendizagem supervisionada baseado em árvores de decisão que implementa a técnica de *boosting*, onde várias árvores fracas são combinadas sequencialmente para construir um modelo robusto, conforme descrito por Ferreira e Figueiredo (2012). O XGBoost destaca-se pela sua eficiência computacional,

capacidade de lidar com dados heterogéneos e regularização integrada para controlo do sobreajustamento, beneficiando de otimizações como computação paralela e distribuída, aprendizagem adaptada a dados esparsos e técnicas *out-of-core*, que permitem escalar o modelo para centenas de milhões de observações mesmo em máquinas com recursos limitados (Chen e Guestrin, 2016). O algoritmo funciona iterativamente, ajustando cada nova árvore aos resíduos (erros) do modelo anterior, minimizando uma função objetivo composta pelo erro de previsão e um termo de penalização que promove a simplicidade do modelo. A combinação das árvores permite capturar relações complexas entre as variáveis explicativas e a variável dependente, incluindo interações e efeitos não lineares.

Apesar do seu desempenho competitivo em múltiplas tarefas preditivas, o XGBoost partilha limitações intrínsecas da estrutura de *Gradient Boosting Machines* (GBM), conforme originalmente formulado por Friedman (2001). Em particular, o algoritmo está sujeito ao risco de *overfitting*, sobretudo quando se utiliza um número excessivo de iterações, árvores de elevada profundidade ou taxas de aprendizagem inadequadas. Embora o XGBoost mitigue parcialmente esta vulnerabilidade através de técnicas de regularização explícita (L1 e L2), *shrinkage* e *subsampling*, a calibragem dos hiperparâmetros continua a ser crítica para assegurar a generalização do modelo (Chen e Guestrin, 2016). Adicionalmente, o custo computacional associado ao treino sequencial de múltiplos *weak learners* pode tornar o método menos eficiente em ambientes com grandes volumes de dados ou restrições de tempo. Do ponto de vista interpretativo, a complexidade do *ensemble* dificulta a extração de inferência direta sobre a contribuição individual das variáveis preditoras. Finalmente, em cenários com elevada presença de ruído ou *outliers*, o modelo pode incorrer em sobreajustamento, uma vez que a estratégia de minimização do erro em cada iteração não distingue entre ruído aleatório e informação relevante. (Friedman, 2001; Chen e Guestrin, 2016).

No presente estudo, a utilização do XGBoost implicou a transformação da série temporal num conjunto de dados tabular, através da criação de variáveis de *lag* da variável dependente, para representar a dependência temporal.

A afinação dos hiperparâmetros do XGBoost foi realizada utilizando *TimeSeriesSplit* combinada com pesquisa em grelha (*GridSearchCV*). Os principais

hiperparâmetros otimizados incluíram o número de árvores (*n_estimators*), a profundidade máxima das árvores (*max_depth*), a taxa de aprendizagem (*learning_rate*), a proporção de amostras utilizadas em cada árvore (*subsample*) e a proporção de variáveis consideradas para cada divisão (*colsample_bytree*).

Posteriormente, foi efetuada uma análise da importância relativa das variáveis preditoras, com base na frequência com que cada variável foi utilizada para realizar divisões nas árvores do modelo, refletindo a sua contribuição relativa para a construção do modelo.

3.9 Random Forest

O Random Forest é um algoritmo de *Machine Learning* do tipo *ensemble*, introduzido por Breiman (2001), que consiste na construção de múltiplas árvores de decisão independentes, cuja combinação melhora a precisão e robustez da previsão.

O método baseia-se em dois princípios fundamentais: o uso de amostragem aleatória dos dados por meio de *bootstrap* para criar subconjuntos de treino diferentes para cada árvore, e a seleção aleatória de um subconjunto limitado de variáveis explicativas para determinar a divisão de cada nó na árvore. Estas duas fontes de aleatoriedade geram árvores menos correlacionadas, o que reduz o sobreajustamento e aumenta a capacidade de generalização do modelo.

Cada árvore é construída até um determinado grau de complexidade e, para problemas de regressão, a previsão do Random Forest é calculada como a média das previsões fornecidas por todas as árvores (Hastie et al., 2009, p. 588). Esta abordagem permite capturar relações não lineares e interações complexas entre variáveis, sem pressupor um modelo paramétrico específico, tornando o método especialmente útil para séries temporais com dinâmica complexa.

Como preditores, foram incluídas variáveis com desfasamento da variável dependente, para capturar a dependência temporal e o efeito de persistência dos valores passados sobre o futuro, e as variáveis exógenas. A construção dessas variáveis com desfasamento implicou a eliminação das primeiras observações com valores ausentes, decorrentes da criação dos *lags*.

Para validar o modelo, utilizou-se o *TimeSeriesSplit*. A seleção dos melhores hiperparâmetros do Random Forest foi realizada através de *Grid Search*, otimizando parâmetros essenciais como o número de árvores na floresta, a profundidade máxima das árvores, o número mínimo de amostras para realizar uma divisão interna e o número mínimo de amostras necessárias para formar uma folha terminal. A calibração adequada destes parâmetros é fundamental para equilibrar a complexidade do modelo, evitando tanto o *underfitting* quanto o *overfitting*. Posteriormente, foi efetuada uma análise da importância relativa das variáveis preditoras, baseada na contribuição de cada uma para a redução da impureza ao longo das árvores. Esta análise possibilitou identificar quais os desfasamentos da série temporal e quais as variáveis exógenas que têm maior peso na previsão, fornecendo conhecimento valioso sobre a dinâmica temporal da série e a importância das influências externas no comportamento da série.

De acordo com Salman et al. (2024), o algoritmo Random Forest apresenta um conjunto de características que explicam a sua ampla utilização em tarefas de classificação e regressão. Uma das principais vantagens é a sua versatilidade, sendo aplicável a diferentes tipos de problemas sem necessidade de alterações estruturais. Além disso, o modelo é capaz de produzir bons resultados mesmo sem ajustamentos extensivos de hiperparâmetros, o que o torna atrativo em contextos com recursos limitados. Os autores também destacam a sua capacidade de lidar com o problema do sobreajuste (*overfitting*), uma vez que o modelo combina diversas árvores de decisão treinadas sobre subconjuntos diferentes dos dados originais. Esta abordagem, conhecida como *bagging* (*bootstrap aggregating*), contribui para a redução do enviesamento e da variância, resultando num modelo mais estável e com melhor desempenho em dados não observados. Outro aspeto vantajoso apontado por Salman et al. (2024) é o facto de o algoritmo permitir alguma compreensão do processo de decisão, nomeadamente através da avaliação da importância das variáveis.

Contudo, os autores referem também algumas desvantagens. Uma das principais limitações é o tempo de treino elevado, sobretudo quando se constrói um número muito elevado de árvores. Embora o treino possa ser relativamente eficiente, a previsão torna-se significativamente mais lenta, o que compromete a sua aplicação em contextos que exigem respostas em tempo real. Outra desvantagem importante mencionada é a redução da interpretabilidade do modelo à medida que se aumenta a sua complexidade. Apesar de

cada árvore individual ser interpretável, a combinação de centenas ou milhares de árvores transforma o modelo numa espécie de "caixa negra", dificultando a compreensão das relações entre variáveis e decisões tomadas.

4. CASO DE ESTUDO

4.1 Apresentação dos dados

A base de dados utilizada no estudo do seguro de saúde é composta por 108 observações. Trata-se de uma série temporal com periocidade mensal, sendo os dados compreendidos no período entre janeiro de 2016 e dezembro de 2024. Na série, a variável dependente é a variável Vendas – que representa o número de novas apólices emitidas do seguro de saúde -, cujos valores históricos foram extraídos da base de dados da empresa. Na figura seguinte é apresentada a série original, com os seus valores expressos em unidades.

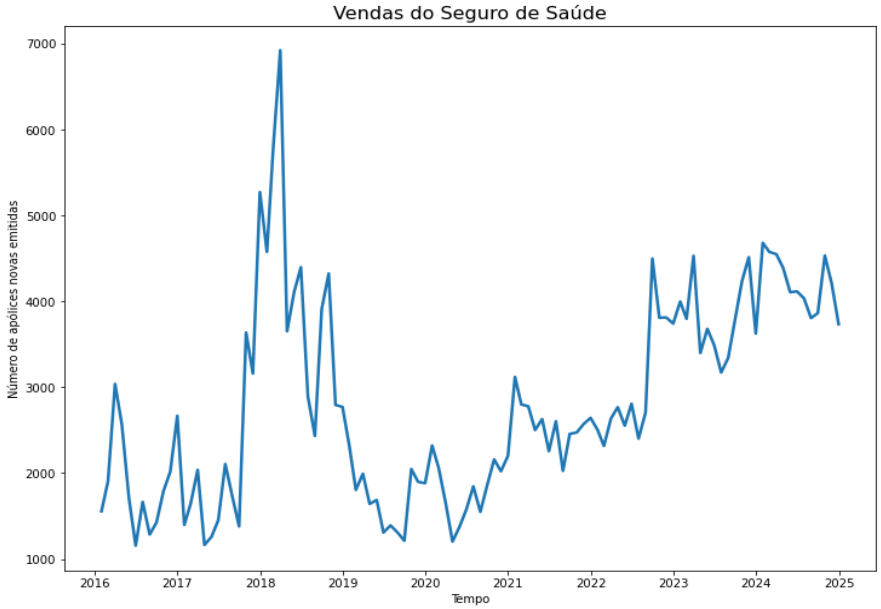


Figura 1: Vendas do seguro de saúde entre 2016 e 2024

Ao analisar a série temporal, observa-se um pico acentuado em 2018, seguido por um período de queda e posterior recuperação, com um crescimento significativo a partir de 2023. O pico registado em 2018 é explicado pela migração dos dados de uma empresa, resultantes de uma fusão. A recuperação das vendas a partir de 2023 pode estar associada ao impacto da pandemia da COVID-19 - que poderá ter levado a um reforço da perceção da importância dos seguros de saúde - ou ao aumento do custo dos serviços de saúde privados e da inflação, levando mais indivíduos e empresas a procurarem proteção financeira. Observando o gráfico, verifica-se a existência de *outliers*. Foi aplicada a metodologia referida no capítulo anterior e foram encontrados cinco valores com *outliers*: 1379 registado a 30/09/2017; 5271 a 31/12/2017; 5792 a 28/02/2018; 6922 a 31/03/2018

e 4497 a 30/09/2002. Verificando-se a existência de *outliers*, optou-se por criar uma cópia da série temporal corrigida dos mesmos, cuja representação segue na figura abaixo.

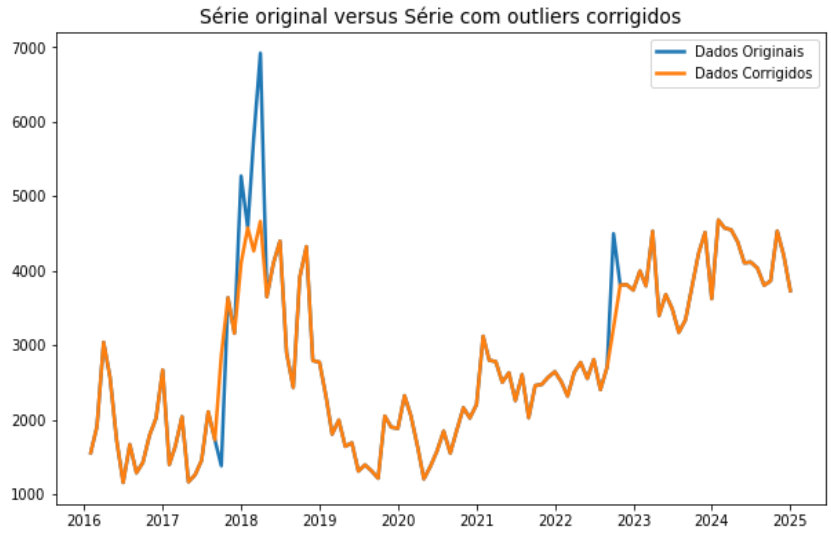


Figura 2: Série Original e Série Corrigida - Vendas do Seguro de Saúde

É possível observar que o tratamento dos *outliers* da série diminui a magnitude de alguns picos que se registavam antes. Assim, nos modelos de previsão que irão ser testados nos tópicos seguintes, serão comparados os erros tanto da série corrigida de *outliers* como da série original. Dado que a intenção do trabalho é prever as vendas de um ano, as duas séries foram divididas em amostra de treino e amostra de teste, sendo o período de teste os últimos 12 meses disponíveis, correspondente ao ano de 2024.

Relativamente às variáveis exógenas, foram selecionadas tanto variáveis internas como externas. A variável interna considerada foi a taxa de conversão de simulações, definida como a proporção de simulações de seguro que resultaram efetivamente na subscrição de uma apólice, ou seja, o número de simulações convertidas em contratos dividido pelo total de simulações realizadas. Esta métrica permite captar a eficácia do processo comercial e o grau de interesse real dos clientes, funcionando como um indicador direto da propensão à compra.

Para além disso, foram incluídas variáveis externas relacionadas com o crescimento económico, uma vez que este está geralmente associado a uma maior disponibilidade financeira por parte das famílias e, consequentemente, a uma maior predisposição para contratar seguros de saúde, sobretudo quando existe essa necessidade ou desejo. Nesse sentido, selecionaram-se como variáveis de crescimento económico o

número de pessoas desempregadas em cada mês e a taxa de variação homóloga do IPC (índice de preços do consumidor), sendo que estes dados foram extraídos da base de dados do Banco de Portugal (conhecida por BPstat). Em alternativa, teria sido interessante utilizar variáveis mais conhecidas como a taxa de desemprego mensal e o PIB, contudo, dado que estes dados não estão disponíveis com a periodicidade mensal necessária para o modelo, optou-se por utilizar estas variáveis substitutas de elevada correlação com os indicadores referidos.

O número de pessoas desempregadas corresponde ao total de indivíduos em idade ativa que, no final de cada mês, se encontram registados nos centros de emprego como estando à procura de trabalho. Trata-se de um indicador do mercado de trabalho que reflete o dinamismo económico e o nível de empregabilidade da população. Uma diminuição do número de desempregados tende a sinalizar um aumento da confiança dos consumidores, maior rendimento disponível e, portanto, maior capacidade para adquirir produtos de seguro. Já o IPC, mede a evolução média dos preços de um cabaz fixo de bens e serviços representativo dos padrões de consumo das famílias. Um IPC mais elevado pode indicar pressão inflacionária e diminuição do poder de compra, o que afeta negativamente a capacidade das famílias para manter ou adquirir novos seguros. Por outro lado, um IPC estável está frequentemente associado a um ambiente económico mais previsível, o que favorece o planeamento financeiro e a aquisição de seguros. Estas variáveis permitem assim capturar, de forma indireta, as condições macroeconómicas que influenciam a procura por seguros de saúde.

As outras variáveis exógenas foram escolhidas tendo em conta a informação disponibilizada no Observatório dos Seguros de Saúde, um website desenvolvido pela NOVA Information Management School em parceria com a ASF. De acordo com a informação disponível neste espaço dedicado a fornecer informação sobre seguros de saúde, as principais razões para adquirir este seguro são, em primeiro lugar, a dificuldade de acesso a serviços do SNS; em segundo, a redução da espera – marcação de consultas/ exames / tratamentos - e em terceiro, a maior qualidade dos serviços prestados no privado.

Assim, as razões prendem-se com a facilidade de acesso, a maior eficiência e a melhor qualidade. Nesse sentido, foi feita uma pesquisa por variáveis relacionadas com estes três conceitos. O SNS possui uma base de dados pública presente no seu website e

foi deste que se extraíram os dados possíveis, de acordo com cada um destes temas. Na tabela abaixo apresentam-se descritas as variáveis retiradas e a categoria com a qual estão relacionadas.

Tabela 1: Variáveis exógenas

Variável	*Descrição*	*Categoria*
% Cir. Ambulatório p/ Procedimentos Ambul.	Percentagem de cirurgias realizadas em ambulatório no total de cirurgias programadas para procedimentos ambulatorizáveis.	Qualidade
Nº de atendimentos em urgência	Número de atendimentos em urgência.	Acesso
Nº de consultas médicas hospitalares	Número de consultas médicas hospitalares.	Acesso
Nº de consultas em telemedicina	Número de consultas realizadas com recurso à utilização de comunicações interativas e audiovisuais e de dados recolhidos na presença do doente.	Acesso
Nº intervenções cirúrgicas	Número de intervenções cirúrgicas nos cuidados hospitalares.	Acesso
% primeiras consultas realizadas em tempo adequado	Proporção de utentes referenciados para a primeira consulta externa, com consulta externa prestada dentro do tempo máximo de resposta garantido, no total de primeiras consultas externas prestadas no período em análise.	Acesso
Taxa de Ocupação Hospitalar	Relação percentual entre o total de dias de internamento no ano e a capacidade do estabelecimento.	Eficiência

Relativamente a estes dados, não foi necessário realizar limpezas ou transformações. Apenas foi preciso garantir que os dados estavam disponíveis na mesma periodicidade da variável dependente. Todos os dados recolhidos foram compilados numa folha de cálculo no Excel.

4.2 Decomposição clássica das forças componentes

Na figura três é apresentada a decomposição das forças componentes da série - tendência, sazonalidade e resíduos. Uma vez que a magnitude das flutuações sazonais não varia muito ao longo do tempo, optou-se por utilizar a descomposição aditiva.

A análise da tendência revela um crescimento acentuado até 2018, seguido de uma queda significativa entre 2019 e 2020. A partir de 2021, observa-se uma recuperação progressiva, refletindo uma possível estabilização do mercado.

A componente sazonal evidencia padrões recorrentes ao longo do tempo, sugerindo que as vendas seguem um comportamento previsível em determinados períodos do ano. A sazonalidade parece ser mais evidente no final e início do ano e em outubro e novembro. Estes períodos podem estar associados a renovação de contratos ou a campanhas de final de ano.

Por fim, a análise dos resíduos indica uma maior variabilidade entre 2017 e 2019, período em que ocorreu a migração da informação resultante da fusão de uma empresa. Nos anos mais recentes, os resíduos apresentam menor dispersão, sugerindo que a série segue um padrão mais estável, com menor influência de fatores imprevisíveis.

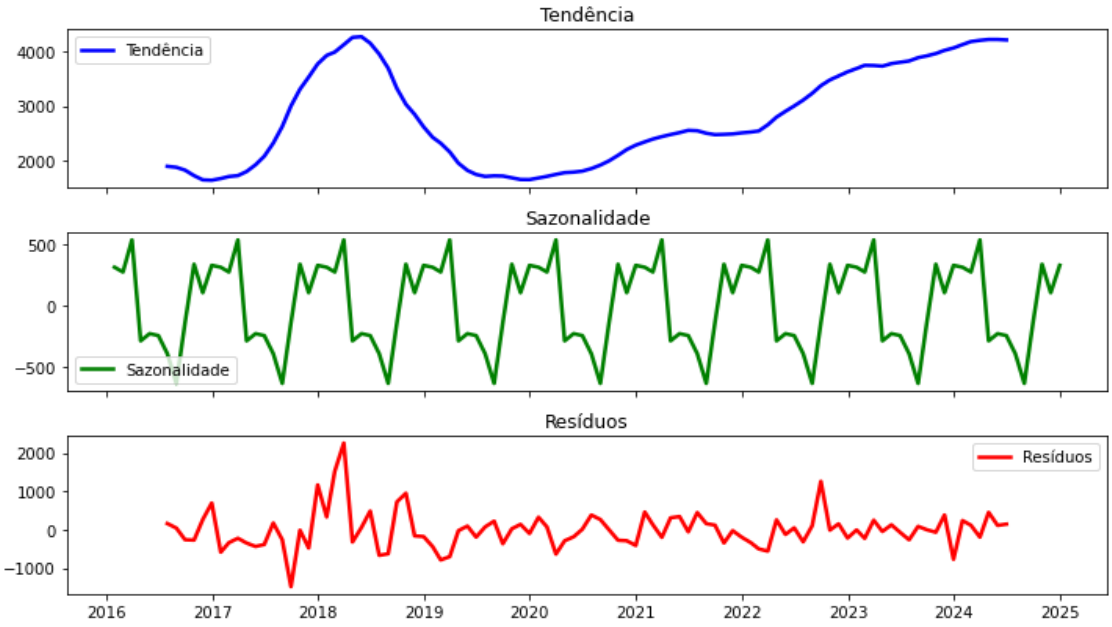


Figura 3: Decomposição das forças componentes

4.3 Método de Holt-Winters

Como referido anteriormente, serão estudados dois métodos de Holt-Winters: o aditivo e o multiplicativo, sendo o multiplicativo apropriado para séries cuja amplitude da sazonalidade aumenta com a tendência, e o aditivo em séries onde tal não se verifica. Optou-se por testar o EPAM em ambos, embora o modelo aditivo seja o que à partida

mais se adequa, dado que não se verifica uma variação da amplitude da sazonalidade com o nível da série. Os resultados do EPAM avaliado na série original e corrigida de *outliers* encontram-se na tabela seguinte.

Tabela 2: Erros de Previsão dos modelos Holt-Winters

EPAM	***Série original***	***Série corrigida***
HWM	25,26%	15,29%
HWA	13,61%	14,28%

Verifica-se assim que a previsão efetuada na série original com o método HMA é a que apresenta o EPAM mais baixo.

4.4 Modelos ARIMA (p,d,q)

De forma a aplicar modelos ARMA, é necessário garantir que a variável dependente é estacionária, isto é, não contém uma raiz unitária. Olhando para a série original, conseguimos prever que à partida não será estacionária dada a variabilidade verificada ao longo dos anos. Além disso, observando os gráficos das funções FAC e FACP, representados na figura seguinte, também temos esses indícios uma vez que a FAC decai lentamente para zero, enquanto a FACP decai bruscamente do *lag* 1 para o *lag* 2.

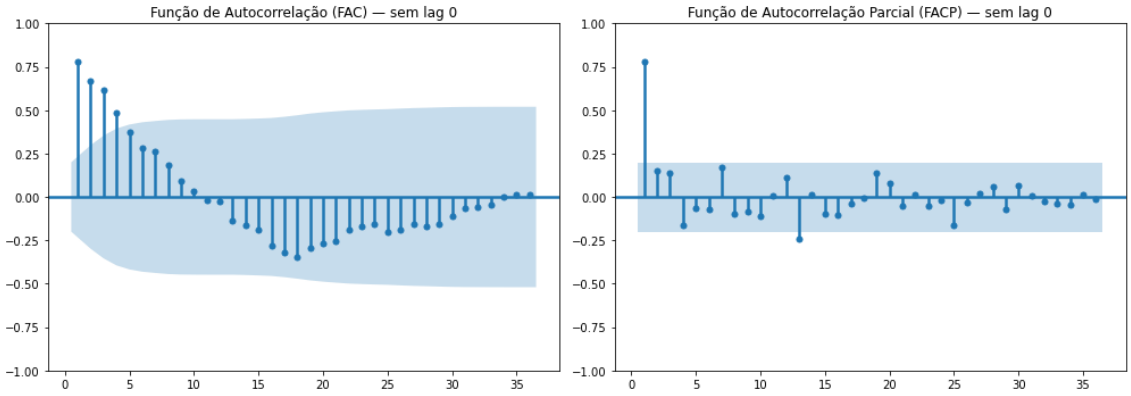


Figura 4: FAC e FACP da série original

Para testar a estacionariedade, foi aplicado o teste ADF tanto na série original como na série corrigida de *outliers*. O valor-p obtido na série original foi de 0,17 e na série com os *outliers* corrigidos foi de 0,36. Assim, com as hipóteses definidas pelo teste ADF (H_0: Existe uma raiz unitária e H_1: Não existe raiz unitária) pode concluir-se que, a um nível de significância de 5%, não se rejeita a hipótese nula em ambas as séries, pelo que é necessário tornar as séries estacionárias.

Relativamente à série original, inicialmente, foi aplicada uma diferenciação sazonal com o objetivo de eliminar a componente sazonal da série, uma vez que esta apresentava padrões repetitivos ao longo do tempo. No entanto, esta transformação revelou-se insuficiente, pois a série permaneceu não estacionária, de acordo com os resultados do teste de ADF (valor-p = 0,27). Em seguida, foi aplicada uma diferenciação simples à série já diferenciada sazonalmente, o que resultou numa série estacionária. Contudo, ao aplicar apenas uma diferenciação simples diretamente à série original, esta tornou-se também estacionária, sem necessidade da diferenciação sazonal. Tendo em conta o princípio da parcimónia, que privilegia o uso do menor número possível de transformações para evitar a sobrediferenciação da série, resultando em perda de informação e complexidade desnecessária no modelo, optou-se por considerar apenas uma diferenciação simples ($d = 1$) no processo de modelação.

Quanto à série corrigida, verificou-se exatamente o mesmo caso da série original pelo que se considerou apenas uma diferenciação simples no processo de modelação ($d = 1$).

Estes resultados indicam que, embora existam indícios de sazonalidade, estes não apresentam uma estrutura suficientemente persistente ou dominante que exija, por si só, uma transformação sazonal. A tendência da série parece ser o principal fator de não estacionariedade, sendo adequadamente tratada por uma única diferenciação simples.

Após tornar as séries estacionárias, avançou-se para a identificação de possíveis modelos ARIMA que melhor as descrevessem. Para isso, foram analisadas as funções FAC e FACP das séries com diferenciação simples. Os correlogramas apresentados nas figuras seguintes permitem observar o comportamento das autocorrelações ao longo dos diferentes *lags*, fornecendo indícios sobre a estrutura apropriada dos modelos -

nomeadamente, os valores mais adequados para os parâmetros autorregressivo (p) e de média móvel (q).

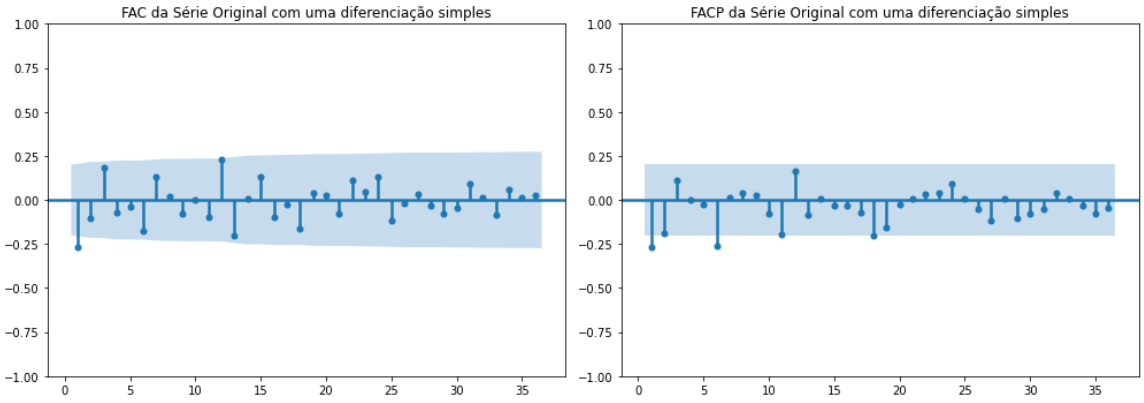


Figura 5: FAC e FACP da série original com uma diferenciação simples

Na série original estacionária, verifica-se que a FAC decai de uma forma mais gradual, ficando sempre dentro das bandas de significância após o *lag* 1, sugerindo uma componente MA ordem 1. Já a FACP decai ligeiramente do *lag* 1 para o *lag* 2 e mais acentuadamente do *lag* 2 para o *lag* 3, ficando dentro das bandas de significância nos *lags* seguintes, exceto no 6. Neste observa-se um valor ligeiramente fora das bandas de significância que poderá indicar uma correlação espúria ou um padrão sazonal semestral. Contudo, por uma questão de simplicidade do modelo e para evitar a sobreparametrização do modelo, optou-se por não considerar testar modelos com $p = 6$. Com base nesta informação, os modelos sugeridos para teste foram: ARIMA (1,1,0), ARIMA(0,1,1), ARIMA(1,1,1) e ARIMA(2,1,1), cujos resultados estão sumarizados na Tabela 3.

Na decomposição das forças componentes, observou-se que a série revelava sazonalidade, pelo que se optou por testar também modelos SARIMA$(p, 1, q)(P, 0, Q)_{12}$,. Pelos gráficos da Figura 5, observamos que na FAC regista-se um pico no *lag* 12 – no limite das bandas de significância – e depois decai no *lag* 24. Na FACP verifica-se também um pico no *lag* 12, contudo não é tão significativo. Assim, os modelos ARIMA(p,d,q) mencionados acima como os modelos sugeridos para teste, foram adicionalmente testados com a componente sazonal $(P, D, Q)_{12} = (0,0,1)_{12}$, sendo que foram ainda testadas outras combinações na componente sazonal que se revelaram menos favoráveis.

A Tabela 3 sumariza os resultados obtidos nos modelos ARIMA(*p,d,q*) testados, assim como o modelo $SARIMA(p,1,q)(P,0,Q)_{12}$, que se revelou adequado para a série.

Apesar de o valor do HQIC ser ligeiramente superior aos restantes modelos, o modelo SARIMA$(2,1,1)(0,0,1)_{12}$ apresenta o menor erro e uma boa qualidade de ajustamento.

Adicionalmente, foi testada ainda a função auto_arima com o objetivo de identificar automaticamente a melhor combinação de parâmetros (*p,d,q*) para o modelo SARIMA, com base em critérios de informação. No entanto, verificou-se que o modelo sugerido por esta abordagem – o SARIMA$(2,1,2)(2,0,0)_{12}$ - apesar de apresentar valores de HQIC mais baixos comparativamente aos restantes modelos testados, a sua performance preditiva avaliada pelo EPAM revelou-se inferior, com um erro de previsão de 12.13%.

A análise dos correlogramas da série com *outliers* corrigidos e com uma diferenciação simples revela que tanto a FAC como a FACP apresentam corte ao nível do *lag* 2, com os valores subsequentes a caírem dentro do intervalo de confiança.

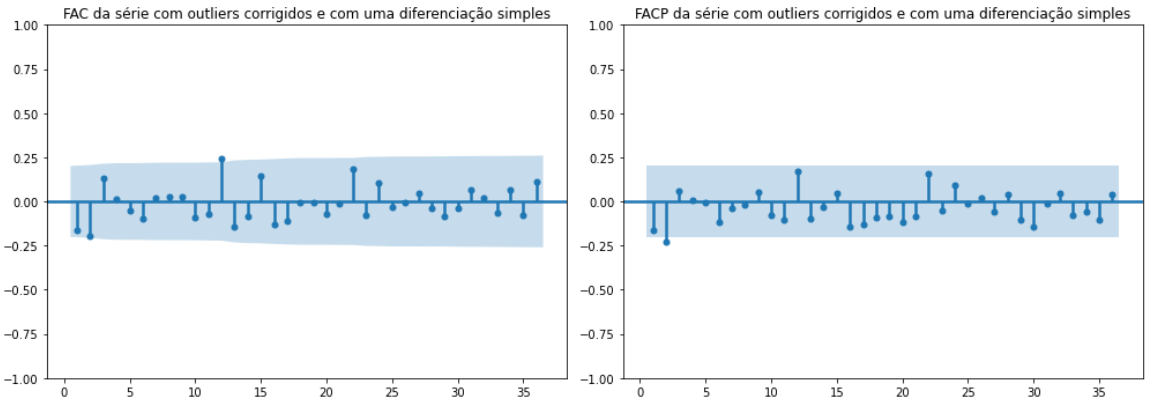


Figura 6: FAC e FACP da série corrigida de *outliers* com uma diferenciação simples

Consideraram-se como candidatos a análise os modelos ARIMA(2,1,0), ARIMA(0,1,2) e ARIMA(2,1,2). Relativamente à componente sazonal, à semelhança do que se observou na série original, a FAC regista um pico no *lag* 12 – aqui fora das bandas de significância – e depois decai no *lag* 24. Na FACP verifica-se também um pico no *lag* 12, contudo não é tão significativo. Assim, os modelos ARIMA(*p,d,q*) mencionados acima como os modelos sugeridos para teste, foram adicionalmente testados com a componente sazonal $(P,D,Q)_{12} = (0,0,1)_{12}$. Contudo, os testes efetuados revelaram que os modelos ARIMA com a componente sazonal demonstravam uma capacidade preditiva inferior aos modelos ARIMA, pelo que a componente sazonal não foi considerada.

Mais uma vez foi testada a função auto_arima, contudo, verificou-se que o modelo sugerido por esta abordagem, que foi o $SARIMA(2,1,1)(1,0,1)_{12}$ revelou um EPAM de 11.46%, superior a outros modelos testados.

A tabela abaixo sumariza os resultados obtidos nos modelos ARIMA tanto na série original como na série corrigida, assim como o modelo *SARIMA,* que se revelou adequado para a série.

Tabela 3: Sumário dos modelos ARIMA e SARIMA

Modelo	***ARIMA (1,1,0)***	***ARIMA (0,1,1)***	***ARIMA (1,1,1)***	***ARIMA (2,1,1)***	**SARIMA (2,1,1)(0,0,1)**	***ARIMA (2,1,0)***	***ARIMA (0,1,2)***	***ARIMA (2,1,2)***
Série	Original	Original	Original	Original	Original	Ouliers Corrigidos	Ouliers Corrigidos	Ouliers Corrigidos
HQIC	1520,93	1519,35	1522,36	1522,37	1523,29	1467,72	1468,89	1472,72
Significância do Coeficiente AR	valor-p=0,02 (Sig.)	N/A	valor-p=0,90 (Não Sig.)	valor-p (L1) =0,00 (Sig.) valor-p (L2) =0,00(Sig.)	valor-p (L1) =0,01 (Sig.) valor-p (L2) =0,00(Sig.)	valor-p (L1) =0,05 (Sig.) valor-p (L2) =0,01(Sig.)	N/A	valor-p (L1) =0,91 (Não Sig.) valor-p (L2) =0,00 (Sig.)
Significância do Coeficiente MA	N/A	valor-p=0,00 (Sig.)	valor-p=0,16 (Não Sig.)	valor-p=0,03 (Sig.)	valor-p (L1) =0,05 (Não Sig.) valor-p (S.L2) =0,40 (Não Sig.)	N/A	valor-p (L1) =0,11 (Não Sig) valor-p (L2) =0,07 (Não Sig.)	valor-p (L1) =0,41 (Não Sig.) valor-p (L2) =0,02 (Sig.)
Ljung-Box (Prob(Q) Resíduos)	0,59	0,98	0,95	0,91	0,95	0,93	0,80	0,95
FAC dos Resíduos	Sem *lags* significativos	Sem *lags* significativos	Sem *lags* significativos	Sem *lags* significativos	Sem *lags* significativos	Sem *lags* sig.	Sem *lags* sig.	Sem *lags* sig.
FACP dos resíduos	*Lag* 6 significativo	*Lag* 6 significativo	*Lag* 6 significativo	Sem *lags* sig.	*Lag* 18 significativo	Sem *lags* sig.	Sem *lags* sig.	Sem *lags* sig.
EPAM (%)	9,43%	8,80%	8,85%	8,76%	8,56%	9,06%	10,03%	8,57%

Pela análise da tabela é possível observar que a qualidade de ajustamento dos modelos melhorou com a correção de *outliers* (à luz do critério HQIC, que diminuiu), e a qualidade preditiva também melhorou ligeiramente em alguns modelos corrigidos, como foi o caso do ARIMA(2,1,2), que apresentou o menor EPAM na série com os *outliers* corrigidos. Verificou-se ainda que, de forma geral, os modelos não evidenciam autocorrelação

significativa nos resíduos, o que indica uma boa capacidade de captura da estrutura de dependência temporal da série. Este facto é corroborado pelos gráficos da FAC e FACP dos resíduos, onde, salvo em alguns casos pontuais de modelos aplicados à série original, não se observam *lags* significativos.

Apesar o ARIMA(2,1,2) e o $SARIMA(2,1,1)(0,0,1)_{12}$ terem um EPAM praticamente idêntico - 8,57% no ARIMA e 8,56% no SARIMA -, o ARIMA destaca-se por possuir um HQIC inferior (1472.72 contra 1523.29), o que indica uma estrutura mais parcimoniosa. Além disso, os resíduos do ARIMA não apresentam autocorrelação significativa, com um valor *p* do teste de Ljung-Box igual a 0,95, sugerindo um bom ajuste. Por outro lado, o SARIMA, apesar da leve vantagem no EPAM, revela autocorrelação significativa na FACP dos resíduos, o que compromete a qualidade do ajuste. Deste modo, conclui-se que o modelo ARIMA(2,1,2) oferece o melhor compromisso entre simplicidade, ajustamento estatístico e desempenho preditivo, justificando a sua escolha como modelo final.

A componente autoregressiva de ordem dois (AR(2)) significa que os valores atuais da série de vendas dependem significativamente dos seus valores nos dois períodos passados (*lags* um e dois). Isto sugere a presença de uma estrutura de dependência temporal de curto prazo, onde o histórico recente das vendas é um preditor importante para as vendas correntes. A componente de média móvel de ordem dois (MA(2)) indica que os valores atuais da série são também influenciados pelos erros de previsão (choques ou inovações) dos dois períodos passados. Isso implica que choques inesperados no passado (que não foram explicados pelo modelo) ainda exercem uma influência direta sobre o valor presente da série, sugerindo que a série reage e incorpora os efeitos de eventos não previstos por até dois períodos anteriores.

4.5 Modelo ARIMAX

Antes da aplicação dos métodos de seleção de variáveis e com o objetivo de compreender o impacto potencial das variáveis exógenas sobre a variável dependente, procedeu-se ao cálculo da matriz de correlação (anexo 1). Esta análise preliminar permitiu identificar associações lineares relevantes, úteis para orientar a seleção de variáveis no modelo. Para efeitos de interpretação, consideraram-se como referência as categorias de

força da correlação definidas por Dancey e Reidy (2020, p. 182), segundo as quais correlações entre 0,4 e 0,6 são consideradas moderadas, entre 0,1 e 0,3 fracas, e abaixo de 0,1 muito fracas ou inexistentes.

A variável que apresenta a correlação positiva mais elevada com as vendas é a "% Cir. Ambulatório p/ Procedimentos Ambul.", com um coeficiente de +0,48. Em contraste, a variável "% de primeiras consultas realizadas em tempo adequado" evidencia uma correlação negativa moderada (-0,53), o que poderá indicar que, em períodos de maior eficiência do SNS, existe menor procura por soluções privadas, refletindo-se em menores vendas. Destacam-se ainda, com correlações moderadas, o "IPC total – taxa de variação homóloga" (+0,35), "Taxa de Conversão de Simulações" (+0,34) e o "Nº de Pessoas Desempregadas" (-0,37). Por outro lado, as variáveis "Nº de atendimentos em urgência", "Nº de consultas médicas hospitalares", "Nº de consultas em telemedicina", "Nº de intervenções cirúrgicas" e a "Taxa de Ocupação Hospitalar" revelam correlações fracas, sugerindo fraca capacidade explicativa sobre as vendas.

A correlação negativa entre o número de pessoas desempregadas e as vendas mensais do seguro (-0,37) está em linha com a teoria económica, na medida em que o aumento do desemprego tende a reduzir o rendimento disponível das famílias e, por consequência, a sua propensão para contratar seguros de saúde privados. Relativamente ao IPC, a correlação observada foi positiva (+0,35), sugerindo que períodos de maior inflação podem estar associados a maior valorização ou procura por soluções privadas de saúde, eventualmente como resposta a dificuldades de resposta do setor público.

Após esta análise inicial, as três abordagens mencionadas na metodologia foram consideradas para selecionar as variáveis exógenas com o objetivo de avaliar se alguma das variáveis candidatas acresceria valor preditivo ao modelo ARIMA(2,1,2) previamente identificado como o mais adequado.

A primeira abordagem consistiu num procedimento híbrido de *stepwise*. Partindo de um modelo sem variáveis exógenas, foram sucessivamente testadas adições e remoções de cada variável, só validando aquelas que reduzissem efetivamente o HQIC. A variável exógena selecionada com base neste método foi a "% Cir. Ambulatório p/ Procedimentos Ambul.", obtendo-se um HQIC de 1467,59 – mais baixo que o ARIMA puro – e um erro de previsão de 11,12%, superior ao do ARIMA(2,1,2).

A segunda abordagem recorreu à regressão LASSO com penalização L_1, utilizando validação temporal para identificar automaticamente o parâmetro de regularização ótimo. As variáveis foram primeiro padronizadas e depois submetidas ao LassoCV, porém, todos os coeficientes foram forçados a zero, o que significou que nenhuma variável demonstrou relevância estatística para a previsão.

Por último, em cada uma das dez variáveis exógenas testou-se um ARIMAX(2,1,2), avaliando o HQIC e o EPAM na amostra de teste. O modelo puro obteve HQIC≈1472,72 e EPAM≈8,57 %. A inclusão isolada da taxa de variação homóloga do IPC reduziu o EPAM para 6,81% e aumentou o HQIC para 1477,47. A inclusão isolada da variável "Nº de pessoas desempregadas" reduziu o EPAM para 7,96%, levando a um aumento do HQIC para 1476,53. A combinação destas revelou a melhor capacidade preditiva, com um EPAM de 6,27%, apesar de uma qualidade do ajustamento ligeiramente inferior do que o ARIMA(2,1,2) puro. Na tabela abaixo observa-se os valores do HQIC, do EPAM e do teste à correlação dos resíduos, e a comparação entre os valores dos modelos ARIMAX com os do modelo ARIMA (2,1,2).

Tabela 4: Sumário dos modelos ARIMAX

Modelos	***EPAM (%)***	***HQIC***	***Prob(Q) Resíduos***
ARIMA(2,1,2)	8,57%	1472,72	0,95
ARIMA(2,1,2) + % Cir. p/ proced. Ambul. (*Stepwise*)	11,12%	1467,59	0,86
ARIMA(2,1,2) + IPC	6,81%	1477.,7	0,99
ARIMA(2,1,2) + Nº de pessoas desempregadas	7,96%	1476,53	0,92
ARIMA(2,1,2) + IPC+ Nº de pessoas desempregadas	**6,27%**	1477,55	0,92

Com base na avaliação comparativa, foi selecionado o modelo ARIMA(2,1,2) com as variáveis exógenas "Nº de Pessoas Desempregadas" e "IPC total – taxa de variação homóloga" como o mais adequado para a previsão das vendas do seguro de saúde. O teste de Ljung-Box, com valor-p de 0,92, indicou ausência de autocorrelação significativa nos resíduos, o que confirma a validade estatística das previsões.

Embora o modelo ARIMAX com as variáveis exógenas "IPC total – taxa de variação homóloga" e "Nº de pessoas desempregadas" tenha alcançado o menor EPAM, registou um valor de HQIC ligeiramente superior ao do modelo ARIMA puro. Este resultado evidencia um *trade-off* clássico entre precisão preditiva e parcimónia do modelo. Por um lado, a inclusão de variáveis exógenas permitiu captar efeitos contextuais relevantes, melhorando a capacidade de previsão. Por outro, o aumento do número de parâmetros levou a uma penalização no critério de informação, que favorece modelos mais simples. Esta tensão entre complexidade e desempenho deve ser gerida com base nos objetivos do estudo. Dado que o principal objetivo deste trabalho é a obtenção de previsões mais precisas, optou-se por privilegiar a acuidade preditiva em detrimento da parcimónia, justificando assim a escolha do modelo ARIMAX com duas variáveis exógenas como o mais adequado.

4.6 XGBoost

A abordagem adotada começou pela criação das variáveis de *lag* 1, 2 e 3 (curto prazo), bem como os *lags* 12, 24 e 36 (correspondentes à sazonalidade anual, bianual e trianual). Estas variáveis de *lag* foram combinadas com o conjunto de variáveis exógenas previamente definido. A criação dos *lags* implicou que as primeiras 36 observações da série se tornassem inválidas para modelação, uma vez que continham valores em falta resultantes da ausência de histórico suficiente. Por conseguinte, essas observações iniciais foram descartadas, garantindo que apenas dados completos fossem utilizados na modelação. Com o conjunto de dados preparado, foi realizada a divisão entre treino e teste, respeitando o mesmo corte temporal adotado nos outros modelos.

Com o modelo final treinado na amostra de treino, foram realizadas previsões com base na amostra de teste. A avaliação do desempenho foi feita com base no EPAM tendo-se obtido um valor de 9,64%.

A mesma metodologia foi aplicada à versão da série temporal onde os valores atípicos previamente identificados foram corrigidos. Este modelo, ao ser aplicado ao conjunto de teste, obteve um EPAM de 8.80%, inferior ao erro obtido na série original.

Adicionalmente, analisou-se a importância das variáveis, visível na figura abaixo. Os resultados revelaram que as variáveis exógenas assumiram um papel dominante nas previsões, com destaque para a "Taxa de Conversão de Simulações" e a "% de Cir. Ambulatório p/ Procedimentos Ambul". Entre os *lags*, destacaram-se os *lags* 36, 1 e 12, confirmando o valor preditivo tanto dos efeitos de curto prazo como dos ciclos sazonais.

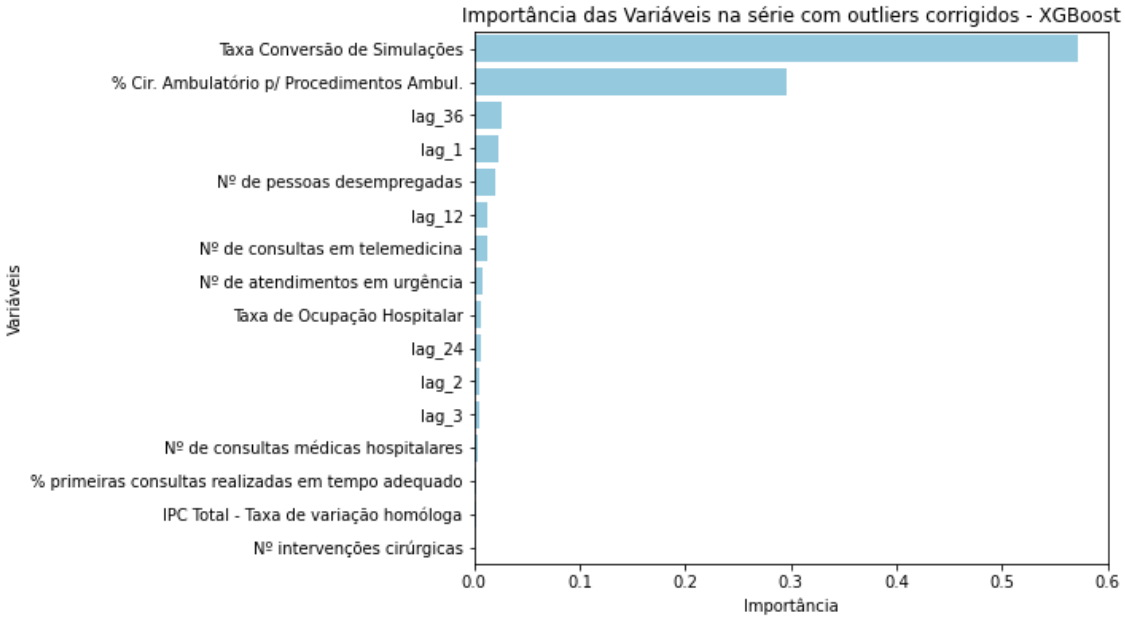


Figura 7: Importância das variáveis na série com *outliers* corrigidos – XGBoost

No conjunto da análise, o desempenho do XGBoost revelou-se competitivo: superou os modelos Holt-Winters em termos de erro percentual, mas não alcançou o desempenho obtido pelos modelos ARIMA e ARIMAX.

4.7 Random Forest

Para a construção do modelo Random Forest, inicialmente foram criadas variáveis com desfasamento (*lags*) da variável dependente correspondentes aos períodos de 1, 2, 3, 12, 24 e 36 meses. Após a criação dessas variáveis e a subsequente remoção das observações com valores ausentes gerados pelo deslocamento temporal, o modelo foi treinado. No cenário da série original, o modelo Random Forest obteve um EPAM de 8,02% na amostra de teste. Já no cenário com os *outliers* previamente corrigidos, o desempenho melhorou para um EPAM de 6,58%, evidenciando o impacto positivo da correção de *outliers*. Na imagem abaixo observa-se graficamente a importância de cada variável na série com os *outliers* corrigidos.

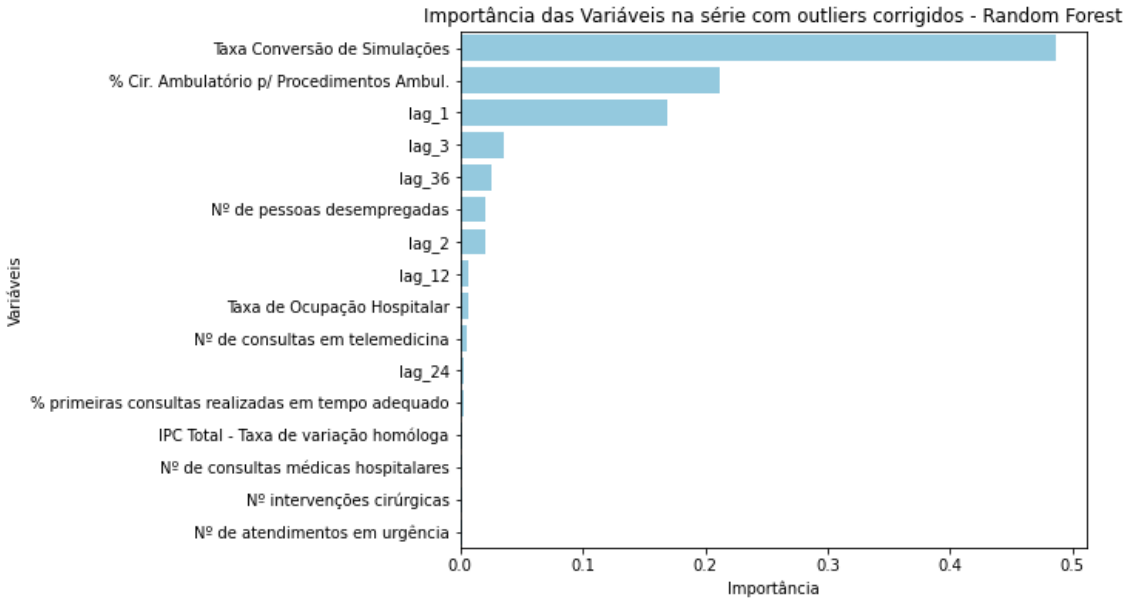


Figura 8: Importância das variáveis na série com *outliers* corrigidos - Random Forest

A análise da importância das variáveis revelou que a taxa de conversão de simulações foi a variável mais relevante para a previsão das vendas do seguro de saúde. Em seguida, destacou-se a percentagem de cirurgias ambulatórias para procedimentos ambulatorizáveis, que também apresentou elevada importância no modelo. Quanto às componentes autoregressivas da série temporal, o *lag* de 1 mês (lag_1) teve um papel importante, seguido pelos *lags* de 3 e 36 meses, embora com peso inferior. Os restantes *lags* e variáveis exógenas apresentaram importâncias relativamente reduzidas. Estes resultados coincidiram com os resultados obtidos no XGBoost e evidenciam que a

inclusão de variáveis exógenas é essencial para aprimorar a precisão preditiva do modelo, complementando a informação temporal fornecida pelas variáveis desfasadas.

4.8 Avaliação dos Erros de Previsão

A tabela seguinte resume o desempenho dos melhores modelos dentro de cada abordagem testada, permitindo uma comparação clara entre métodos tradicionais de séries temporais e técnicas de *Machine Learning*. Para cada método - Holt-Winters, ARIMA, ARIMAX, XGBoost e Random Forest - foi selecionado o modelo com menor EPAM, com identificação da respetiva série, de modo a avaliar a robustez dos modelos face a valores atípicos.

Tabela 5: Erros de Previsão dos melhores modelos em cada método

Modelo	*Série*	*EPAM (%)*
Holt-Winters	Original	13,61%
ARIMA	Com outliers corrigidos	8,57%
ARIMAX	Com outliers corrigidos	6,27%
XGBoost	Com outliers corrigidos	8,80%
Random Forest	Com outliers corrigidos	6,58%

Entre os modelos analisados, o HWA, aplicado à série original, apresentou o pior desempenho, com um EPAM de 13,61%. Embora tenha sido o melhor dentro do seu grupo, revelou-se significativamente inferior aos demais métodos.

O modelo ARIMA(2,1,2), aplicado à série com *outliers* corrigidos, mostrou uma melhoria substancial, alcançando um EPAM de 8,57%.

A incorporação de variáveis exógenas no modelo ARIMAX, nomeadamente com as variáveis "IPC total - taxa de variação homóloga" e "Nº de pessoas desempregadas", contribuiu para um ganho adicional de desempenho, fazendo com que o modelo atingisse o melhor resultado global, com um EPAM de 6,27%.

Os modelos baseados em *Machine Learning* também apresentaram desempenhos competitivos: o XGBoost, aplicado à série com *outliers* corrigidos, registou um EPAM de 8,80%, enquanto o Random Forest alcançou um erro de 6,58%. Apesar de não superar o ARIMAX, o Random Forest destacou-se pela sua capacidade de capturar interações

complexas entre variáveis sem pressupor uma estrutura paramétrica rígida, alcançando um erro muito próximo do modelo ARIMAX.

Em suma, o modelo ARIMAX revelou-se o mais eficaz para a previsão das vendas, superando tanto os métodos tradicionais de séries temporais como as técnicas de *Machine Learning* testadas.

5. DISCUSSÃO DOS RESULTADOS

A análise comparativa dos diferentes modelos de previsão permitiu identificar a abordagem com melhor desempenho, bem como compreender o impacto de decisões metodológicas importantes, como o tratamento de *outliers*, a escolha das técnicas de modelação e a inclusão de variáveis exógenas. Este capítulo apresenta e discute os principais resultados obtidos.

Um aspeto relevante a destacar é o efeito da correção dos *outliers* na performance dos modelos. A correção dos *outliers* revelou-se benéfica para todos os modelos avaliados, com exceção do Holt-Winters. Este resultado evidencia que a limpeza e pré-processamento cuidadoso dos dados pode contribuir para melhorar a precisão das previsões. Nos modelos ARIMA e ARIMAX, a remoção dos *outliers* conduziu a uma redução significativa do erro, com o ARIMA a registar um EPAM de 8,57% e o ARIMAX a destacar-se como o melhor modelo, com um EPAM de 6,27%. Este último integrou variáveis exógenas relevantes, nomeadamente a "IPC total – taxa de variação homóloga" e o "Nº de pessoas desempregadas", sublinhando a importância de incluir indicadores macroeconómicos para captar efeitos contextuais que influenciam a procura por seguros.

É de notar ainda que apesar de a inclusão de variáveis exógenas ter melhorado significativamente o desempenho do modelo ARIMAX, nem todas as variáveis com correlação aparente se traduziram em ganhos preditivos. Este resultado reforça a importância de uma abordagem empírica na seleção de variáveis preditivas: embora algumas variáveis apresentem correlações moderadas com a variável dependente, apenas a sua inclusão efetiva no modelo permite avaliar o seu verdadeiro contributo para a melhoria da previsão.

A análise da importância das variáveis nos modelos Random Forest e XGBoost revelou que as variáveis exógenas tiveram um peso preponderante na performance preditiva. A componente autorregressiva da variável dependente, manteve-se relevante, mas com menor influência relativamente às variáveis externas nestes modelos de *Machine Learning*. Este padrão sugere que, nestes algoritmos, a incorporação de múltiplas fontes de informação externa é fundamental para melhorar as previsões, complementando a informação temporal da série.

Os resultados obtidos neste estudo corroboram as evidências da literatura, que reconhece tanto a eficácia dos modelos tradicionais de séries temporais, como o ARIMA e o ARIMAX, quanto a crescente aplicabilidade de algoritmos de *Machine Learning*, como o Random Forest e o XGBoost, em modelos de previsão no setor segurador.

De forma geral, os resultados demonstram que não existe um modelo universalmente superior: o desempenho depende da natureza da série, do pré-processamento dos dados e das características de cada abordagem. A complementaridade entre métodos estatísticos e de *Machine Learning* permitiu obter uma visão mais completa do comportamento da variável dependente, contribuindo para previsões mais robustas e úteis para a tomada de decisão.

6. CONCLUSÃO

Este estudo teve como objetivo avaliar vários métodos de previsão das vendas mensais de um seguro de saúde, comparando abordagens tradicionais de séries temporais com métodos de *Machine Learning*. Os resultados demonstraram que a escolha do modelo, o pré-processamento dos dados e a seleção adequada de variáveis exógenas são fatores determinantes para a qualidade das previsões. O modelo que obteve o melhor desempenho foi o modelo ARIMAX(2,1,2), com a inclusão das variáveis exógenas "IPC total – taxa de variação homóloga" e "Nº de pessoas desempregadas", aplicado à série com *outliers* corrigidos. Este resultado evidencia o valor da incorporação de informação externa na modelação de séries temporais, especialmente no contexto de previsão de vendas, onde fatores macroeconómicos exercem influência direta sobre a procura.

Para além da contribuição metodológica, este trabalho apresenta também implicações práticas relevantes. A aplicação do modelo ARIMAX permite substituir o processo atual de previsão, realizado em Excel, por uma abordagem automatizada, mais robusta e menos suscetível a erros. Esta automatização pode libertar recursos internos e aumentar a eficiência operacional da empresa. A melhoria da precisão preditiva contribui ainda para uma definição mais realista de metas comerciais, com impacto direto na gestão das equipas de vendas e no planeamento financeiro. A integração de variáveis macroeconómicas permite antecipar variações na procura com base em tendências externas, oferecendo uma ferramenta útil para ajustar estratégias comerciais e de marketing. Adicionalmente, a metodologia desenvolvida é facilmente adaptável a outros produtos da seguradora, como seguros automóvel ou de vida, o que reforça a sua aplicabilidade e valor estratégico.

Entre as limitações identificadas, destacou-se a dificuldade em encontrar estudos para incluir na revisão de literatura relacionados com a venda de seguros, em particular, do seguro de saúde. Outra limitação está relacionada com a recolha de dados exógenos com periodicidade mensal para as variáveis de interesse. Variáveis importantes como o PIB ou taxa de desemprego não estão disponíveis, com periodicidade mensal, para todo o período analisado, o que levou à escolha de *proxies* como o índice de preços no consumidor e o número de pessoas desempregadas. Outro desafio foi a restrição do

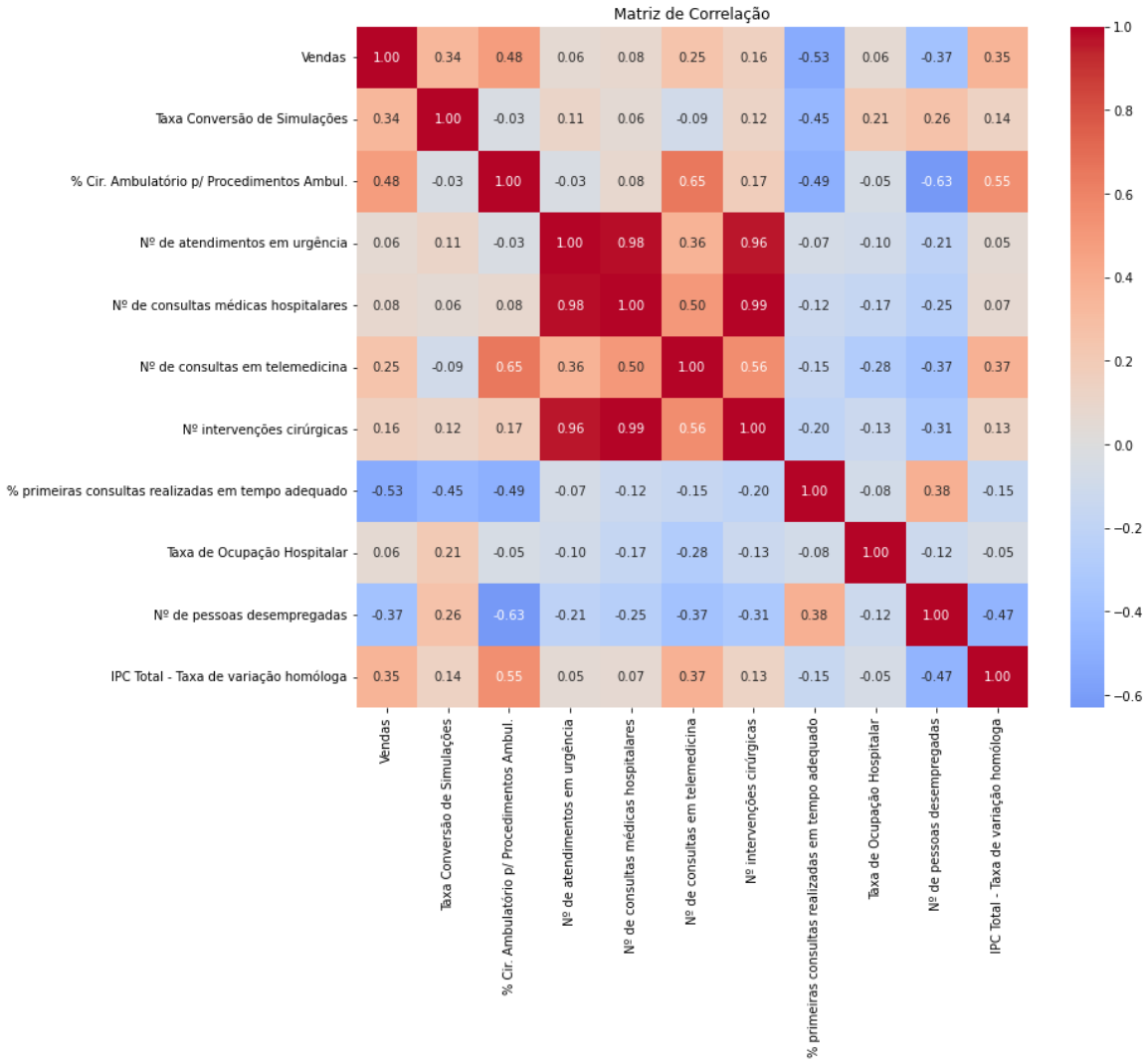
período histórico disponível (dados apenas desde 2016), que poderia ser alargado para melhorar o ajuste e a robustez das previsões.

Para trabalhos futuros, seria interessante avaliar outros modelos de *Machine Learning*, como, por exemplo, redes neuronais, e testar diferentes combinações de parâmetros nos modelos analisados. Outra sugestão é a inclusão de variáveis exógenas adicionais. Na literatura analisada, alguns modelos de previsão incorporam fatores relacionados com os hábitos de saúde dos indivíduos, os quais, neste caso de estudo, também poderiam ser considerados. Exemplos relevantes incluem o consumo de tabaco e os níveis de stress da população, que, quando analisados ao longo do tempo, podem fornecer informações valiosas para melhorar a precisão dos modelos preditivos. Além disso, seria relevante incluir uma variável que represente o posicionamento da empresa face à concorrência, de modo a captar o impacto das dinâmicas competitivas no desempenho comercial. A inclusão do DEI (*Daily Economic Indicator*) poderá igualmente enriquecer o modelo, ao refletir as condições macroeconómicas e as flutuações da atividade económica que influenciam a procura por seguros.

Em suma, este trabalho reforça a importância de combinar diferentes técnicas de modelação e de realizar um tratamento cuidadoso dos dados para melhorar a previsão, oferecendo uma base sólida para análises futuras e para a tomada de decisão estratégica na área comercial.

ANEXOS

Anexo 1 – Matriz de correlação



Anexo 2 – Repositório de Código e Dados

O ficheiro de dados em Excel e o código desenvolvido em Python no âmbito deste trabalho encontram-se disponíveis publicamente no seguinte repositório GitHub: https://github.com/anamargaridajesus/health_insurance_forecast

REFERÊNCIAS

Badshah, W., & Bulut, M. (2020). Model selection procedures in bounds test of cointegration: Theoretical comparison and empirical evidence. *Economies, 8*(2), 49. https://doi.org/10.3390/economies8020049

Bontempi, G., Ben Taieb, S., & Le Borgne, Y. A. (2013). Machine learning strategies for time series forecasting. In M. A. Aufaure & E. Zimányi (Eds.), *Business intelligence: eBISS 2012* (*Lecture Notes in Business Information Processing*, Vol. 138, pp. 62–77). Springer. https://doi.org/10.1007/978-3-642-36318-4_3

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Caiado, J. (2022). *Métodos de Previsão em Gestão com Aplicações em Excel* (3.ª ed.). Edições Sílabo.

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). "STL: A seasonal-trend decomposition procedure based on Loess". *Journal of Official Statistics*, 6(1), 3–73.

Dancey, C., & Reidy, J. (2020). *Statistics without Maths for Psychology*. Pearson. Consultado em 5 de abril de 2025, em www.pearson-books.com

Dash, S., Panigrahi, B. S., Sanikommu, V. V. B. R., Madhavi, B. K. and Sahoo, S. K. (2024). "A comparative analysis of different machine learning techniques for medical insurance premium prediction", *1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)*, pp. 1–6, doi: 10.1109/IC-CGU58078.2024.10530731.

Desboulets, L. D. D. (2018). A review on variable selection in regression analysis. In *Econometrics* (Vol. 6, Issue 4). MDPI AG. https://doi.org/10.3390/econometrics6040045

Diao, L., & Wang, N. (2019). "Research on Premium Income Prediction Based on LSTM Neural Network". *Advances in Social Sciences Research Journal*, *6*(11), 256–260. https://doi.org/10.14738/assrj.611.7397

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, *74*(366), 427. https://doi.org/10.2307/2286348

Ferreira, A., & Figueiredo, M. (2012). *Boosting Algorithms: A Review of Methods, Theory, and Applications*, Capítulo 3, p.1

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Hannan, E. J., & Quinn, B. G. (1979). "The determination of the order of an autoregression". *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2), 190–195. https://doi.org/10.1111/j.2517-6161.1979.tb01647.x

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts. Consultado em 2 de junho de 2025, em https://otexts.com/fpp3/

Kaur, J., Parmar, K. S., & Singh, S. (2023). Autoregressive models in environmental forecasting time series: a theoretical and application review. In *Environmental Science and Pollution Research* (Vol. 30, Issue 8, pp. 19617–19641). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s11356-023-25148-9

Masih, J., Majumdar, A., & Sharma, A. (2024). *What does the future hold for health insurance premiums in India: Insights from ARIMA models. Accountancy Business and the Public Interest, 40*(7), 49–65.

Neel, D. T. H. (2025). "Comparative study for Modeling and Forecasting life insurance premiums applying ETS, Holt Winter, NNETAR, and TBATS models". *Scientific Journal for Financial and Commercial Studies and Research, Faculty of Commerce, Damietta University*, *6*(2), 55-81. https://journals.ekb.eg/article_434169_a9cc7da69ad95d8b4e6f6858c87f590e.pdf

Orji, U. & Ukwandu, E. (2024). "Machine learning for an explainable cost prediction of medical insurance", *Machine Learning with Applications*, 15,100516, https://doi.org/10.1016/j.mlwa.2023.100516.

Ostertagová, E., & Ostertag, O. (2013). *Forecasting using simple exponential smoothing method. Acta Electrotechnica et Informatica, 12*(3), 62–67. https://doi.org/10.2478/v10198-012-0034-2

Plevris, V., Solorzano, G., Bakas, N. P., & ben Seghier, M. E. A. (2022). INVESTIGATION OF PERFORMANCE METRICS IN REGRESSION ANALYSIS AND MACHINE LEARNING-BASED PREDICTION MODELS. *World Congress in Computational Mechanics and ECCOMAS Congress*. https://doi.org/10.23967/eccomas.2022.155

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79. https://doi.org/10.58496/bjml/2024/007

Sijie, L., Sia, F., Alfred, R. and Moung, E. G. (2024). "Ensemble machine learning method for health insurance premium prediction", *IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pp. 642–646, doi: 10.1109/IICAIET62352.2024.10730721.

Stancu, S., Petrică, A.-C., & Ghițulescu, V. (2017). Stationarity-The Central Concept in Time Series Analysis. *International Journal of Emerging Research in Management &Technology*, *6*, 3000. https://doi.org/10.23956/ijermt/V6N1/107

Tibshirani, R. (1996). "Regression shrinkage and selection via the Lasso". *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

Ulyah, S. M., Mardianto, M. F. F., & Sediono, A. (2019). "Comparing the performance of seasonal ARIMAX model and nonparametric regression model in predicting claim reserve of education insurance", *6th International Conference on Research, Implementation and Education of Mathematics and Science* (ICRIEMS 2019) (pp. 122–130). Yogyakarta State University. https://doi.org/10.2991/icriems-19.2019.19

Veiga, C. P. da, Veiga, C. R. P. da, Girotto, F. M., Marconatto, D. A. B., & Su, Z. (2024). "Implementation of the ARIMA model for prediction of economic variables: evidence from the health sector in Brazil". *Humanities & Social Sciences Communications*. https://doi.org/10.1057/s41599-024-03023-3