Tourism demand forecasting in Portugal´s municipalities: an explainable machine learning approach

Catarina Neves

Dissertation written under the supervision of professor Ana Guedes

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa, January 2024.

[Page intentionally left blank.]

Tourism demand forecasting in Portugal´s municipalities: an explainable machine learning approach

Catarina Neves

January 2024

Supervisor: Professor Ana Guedes

Abstract

In recent decades, the tourism industry has experienced remarkable growth, making it essential to provide accurate forecasts of tourism demand for the efficient allocation of resources and continued growth of the industry. This thesis presents a machine learning approach to forecast tourism demand in Portugal. However, forecasting in the tourism sector faces a challenge: compromising the balance between model accuracy and interpretability. While some highly accurate models lack transparency, making them difficult to understand, this study addresses this concern using the Tree SHAP method. By identifying the contributions of the features, this approach offers a globally interpretable model, increasing the reliability of tourism demand forecasts. This thesis aims to answer: 1) What are the tourism demand trends in Portugal?; 2) What are the key features that most significantly influence tourism demand for different tourist accommodations in Portugal´s municipalities?; 3) How to trigger valuable insights in tourism demand forecasting models via the application of explainability strategies? For this purpose, non-public data from Turismo de Portugal and additional variables, such as population, are used to train an XGBoost model. The main predictors of demand in Portugal were revealed, including summer, population, and number of beds. This study has practical implications for policymakers and management teams in the marketing and tourism sectors, providing valuable information for decision-making in the sector.

Keywords: Tourism demand forecasting, Tourist overnight stays, Bed occupancy rate, Explainable AI, Tree SHAP, XGBoost.

[Page intentionally left blank.]

Tourism demand forecasting in Portugal´s municipalities: an explainable machine learning approach

Catarina Neves

January 2024

Supervisor: Professor Ana Guedes

Resumo

Nas últimas décadas, a indústria do turismo tem registado um crescimento significativo, sendo essencial o fornecimento de previsões precisas da procura turística para a alocação eficiente de recursos e para o crescimento contínuo da indústria. Esta tese apresenta uma abordagem de aprendizagem automática para prever a procura turística em Portugal. No entanto, a previsão no setor do turismo enfrenta um desafio: comprometer o equilíbrio entre a precisão do modelo e a sua interpretabilidade. Embora alguns modelos altamente precisos careçam de transparência, tornando-os difíceis de compreender, este estudo aborda esta preocupação utilizando o método Tree SHAP. Ao identificar as contribuições das variáveis, esta abordagem oferece um modelo globalmente interpretável, aumentando a credibilidade das previsões da procura turística. Esta tese pretende responder a: 1) Quais são as tendências da procura turística em Portugal?; 2) Quais são as características chave que influenciam de forma mais significativa a procura turística por diferentes alojamentos turísticos nos municípios de Portugal?; 3) Como acionar conhecimentos valiosos em modelos de previsão da procura turística através da aplicação de estratégias de explicabilidade? Para este propósito, são utilizados dados não públicos do Turismo de Portugal e variáveis adicionais, como a população, para treinar um modelo XGBoost. Os principais fatores de previsão da procura em Portugal foram revelados, incluindo o verão, a população e o número de camas. Este estudo tem implicações práticas para os responsáveis políticos e para as equipas de gestão nos setores do marketing e do turismo, fornecendo informações valiosas para a tomada de decisões no setor.

Keywords: Previsão da procura turística, Dormidas de turistas, Taxa de ocupação de camas, Inteligência Artificial Explicável, Tree SHAP, XGBoost.

[Page intentionally left blank.]

Contents

1 Introduction **1**

2 Literature Review **2**
2.1 Emerging Trends and Challenges in Tourism 2
2.2 Tourism Demand Forecasting 3
2.2.1 Related Work 4
2.3 The Role of Interpretability in Tourism Demand Forecasting 8

3 Methodology **9**
3.1 Data Collection 9
3.2 Data Preparation 10
3.2.1 Data Preprocessing 10
3.3 Modelling 12
3.4 Interpretability 17

4 Results **18**
4.1 Exploratory Data Analysis 18
4.2 Occupancy Forecasting 25
4.3 Model interpretability 34

5 Discussion **38**

6 Conclusion **42**
6.1 Limitations and Future Work 43

Appendix: Figures **49**

List of Tables

1 Description of the variables sources. 9
2 Data dictionary of the final dataset. 12
3 Parameters chosen for the models. 15
4 Error metrics to evaluate performance. 17
5 Descriptive analysis of the key numeric variables of interest. 25
6 Best parameter configuration of the models. 26
7 Performance metrics results of the models. 27

List of Figures

1 Distribution of the train and test set. 13
2 Demand behavior over time. 18
3 Demand behavior over the months of each year. 19
4 Total overnight stays by NUTS II in mainland Portugal (2015-2022). 20
5 Demand behavior for Lisboa over the months of each year. 20
6 Classification frequency of accommodation facilities (%). 21
7 Demand behavior by type of accommodation facility. 22
8 Bed occupancy rate by classification over time. 22
9 Supply behavior over time. 23
10 Triangle Correlation Heatmap. 24
11 Comparison of Residual Plots between the models. 28
12 Comparison of True vs Predicted values on Train and Test sets over time. . . . 29
13 Distribution of the mean errors for the 15 municipalities with high error on Test set. 30
14 Distribution of the mean errors by classification for the 4 municipalities with high error on Test set. 31
15 Distribution of the mean errors for the classifications on Test set. 31
16 Distribution of the mean errors for the 15 municipalities with low error on Test set. 32
17 True vs Predicted values for the accommodation facilities in Lisboa on Test set over time. 33
18 Tree SHAP implementation integrated into XGBoost to explain the entire dataset. 35
19 Force plot illustrating the impact of each feature on the prediction for hotel establishments in Lisboa during August 2022. 36
20 Force plot illustrating the impact of each feature on the prediction for hotel establishments in Lousada during December 2022. 37

21 Force plot illustrating the impact of each feature on the prediction for rural tourism in Vimioso during August 2022. 38
22 Distribution of Bed Occupancy Rate. 49

1 Introduction

In recent decades, the tourism industry has witnessed a growing trend, becoming a crucial sector of a country's economy. As a result, it has become essential for both government agencies and the private sector to develop an appropriate demand management plan (15). Anticipating the future volume of demand in the tourism industry correctly is a key component of this plan, since miscalculations can have serious consequences. Underestimating demand can lead to customer loss and dissatisfaction, overcrowded facilities and stations, and rapid depreciation of facilities. On the contrary, overestimating demand can result in an unjustified increase in companies' costs, maintaining the high quality of unsold seats and increasing total general expenses (21). Therefore, accurate modelling and forecasting of tourism demand is vital for the effective allocation of resources and the continued growth of the tourism industry (48). These forecasts offer valuable insight into both short-term operational strategies and long-term resource optimisation (49).

Portugal has been standing out in the tourism industry and is increasingly becoming a trend due to its diverse attractions, cultural richness, recognised gastronomy, a broad range of diverse accommodation options and other factors. Furthermore, the economic impact of this growing tourism sector has been significant. The increase in the number of tourists and investment in tourism-related companies has played a crucial role in boosting local economies. In 2022 the tourist accommodations in general registered 28.9 million guests, which provided 77.2 million overnight stays growing by 80.7% and 81.1% respectively compared to the previous year (19). This growth is linked to the substantial capital investments required by tourist enterprises, actively stimulating the local economies in which they are built.

This tourism activity contributes to the increase of the national Gross Domestic Product (GDP), strengthening local revenues through compatibility with other existing industries such as restaurants, and generating numerous job opportunities. In 2022, the GDP registered a notable growth, increasing by 6.7% in volume, the highest rate since 1987, and in nominal terms, GDP increased by 11.4% in 2022, reaching around 239.2 billion euros (19).

Given the importance of this sector, precise tourism demand forecasting has become a prominent topic in academic research, and it plays a crucial role in predicting future economic development. However, there are several approaches to forecasting, and it is worth noting that no single forecasting model is suitable for all types of tourism destination data with different time horizons (23).

In this context, this thesis aims to address the specific demand forecasting challenges within the tourism industry. To accurately predict Portugal´s future economic development, a ma-

chine learning approach is initially implemented to predict the demand in the various tourist accommodations of each municipality. However, one of the many challenges in forecasting tourism demand is the compromise between the accuracy of the model and its interpretability. While certain forecasting models can be highly accurate, they often lack transparency, since are insufficient in terms of interpretability, leaving their results less understandable to human comprehension. This insufficient explanation can hinder the wider acceptance of research findings in the tourism sector (46).

To overcome this issue, an explainable machine learning technique will be used in this study. This approach aims to provide more transparent and understandable insights, improving the overall reliability of tourism demand forecasts and aiding decision-making. It is vital to note that tourism demand forecasting is a dominant field of research, but it still offers room for improvement and refinement. With this regard, this thesis focuses on the following research questions:

1. What are the tourism demand trends in Portugal?
2. What are the key features that most significantly influence tourism demand for different tourist accommodations in Portugal´s municipalities?
3. How to trigger valuable insights in tourism demand forecasting models via the application of explainability strategies?

2 Literature Review

2.1 Emerging Trends and Challenges in Tourism

Tourism, as defined by the Division. and Organization. (9), is a social, cultural, and economic phenomenon related to the movement of people to places outside their usual place of residence, pleasure being the usual motivation. This dynamic sector has witnessed evolving trends and faced various challenges over the years.

One notable emerging trend is digital tourism, which is set for continued growth in the future, driven by technological advances and evolving consumer preferences. The future of digital tourism offers exciting possibilities, including the integration of voice and chatbot assistants, social media integration, and sustainable practices. By embracing these evolving trends, companies can promote sustainable growth and offer travellers exceptional experiences for the next years (40).

In addition, all travellers, especially Generations Y and Z, have shown a growing awareness and concern for sustainability. The COVID-19 pandemic has increased their ecological awareness, during their booking behaviors (2).

Tourism is currently facing challenges, having been severely affected by the crisis triggered by the COVID-19 pandemic and then, just as the sector was beginning to recover, it was hit again by the economic and social consequences of Russia's war in Ukraine, which dealt a further blow to recovery prospects. In this climate of uncertainty in the world, and in the tourism sector in particular, transformative actions are necessary to drive recovery and set tourism on the path to a more resilient, sustainable, and inclusive future, and for this, they will have to face new challenges (1).

The COVID-19 crisis laid several vulnerabilities in the tourism sector, challenging the practices of the tourism industry, drawing attention to issues such as poor risk management in the travel industry and the global spread of diseases, but it also allows us to reflect on tourism, leading it to a transformation towards responsible, sustainable, and socially innovative tourism (technofunc). As John Smith, CEO of ABC tourism, remarked, "The pandemic has forced the tourism industry to realize that they need to be able to adapt and change rapidly to meet the shifting needs of their customers. This agility will be critical in the post-pandemic world".

The conflict in Ukraine, coupled with economic sanctions on Russia, has introduced another challenge for the tourism sector. The war has led to rising prices in oil, gas and food, which will affect restaurant and transport costs with an impact on already high inflation rates, especially in Europe, which together with the fact that leisure travellers don't take expensive holidays, greatly damages the tourism sector in general (10).

The future of tourism will be determined and facilitated by technology and allow small tourism players to compete with giants. The fusion of information and communication technologies will enable tourism businesses to become more efficient and competitive (technofunc).

2.2 Tourism Demand Forecasting

In the field of forecasting, accurate predictions depend on the availability and analysis of time series data (5). Tourism demand forecasting has become essential in the tourism industry, and it provides important implications for destination policymakers and tourism practitioners (7). Due to its importance, it has attracted a great interest among scholars, making tourism demand forecasting a prominent and increasingly explored topic in academic research (34). Also is a dominant research theme in the tourism field and one of the most popular measurements is

tourist arrival[1] (Önder (51); Önder and Gunter (52); Onder and Wei (34)).

Although tourism demand forecasting clearly has benefits, it also presents innumerable complexities and challenges. Despite the variety of models for forecasting, each method has its own set of advantages and disadvantages. Furthermore, it is relevant to state that the suitability of a forecasting method can vary according on the specific characteristics of tourism destination data and the time horizon involved (45).

2.2.1 Related Work

Since the 1960s, the field of forecasting domain has relied on linear statistical methods such as ARIMA models. However, in the late 1970s and early 1980s, it became apparent that linear models are not suitable for many real applications (17). Recently, machine learning models have become a formidable alternative to classical statistical models in the forecasting community (5). This change in the forecasting approach has been especially relevant in the context of tourism, where various characteristics are typically forecasted in most studies to understand the dynamics of tourism demand. These variables include the number of trips from the origin country to the tourism destination, the amount of money that is spent by the whole tourists in a year or the average amount of money that each tourist spent in the destination region, and the number of nights that each tourist stays at the destination (13).

Machine learning has introduced versatility to forecasting, allowing the development of advanced techniques such as artificial neural networks (ANN). Although studies have shown the potential to improve forecast accuracy with these machine learning approaches, it is important to emphasise that they may not offer theoretical explanations for associations between tourism demand data and other variables (41). This limitation, evidenced by the lack of transparency and interpretability of machine learning models, makes their results less comprehensible to human understanding. This shortcoming not only hinders informed decision-making and the formulation of strategies in the tourism sector, but also emphasises the need for a balance between forecasting accuracy and the interpretability of forecasting methodologies.

Moreover, it should be recognized that these techniques require some consideration, since no single method consistently outperforms others across all scenarios (34).

In this field of tourism demand forecasting several methodologies are used, including time series models, econometric models, and artificial intelligence (AI) models.

[1] number of tourists arriving at a particular destination.

Time Series Models

Time series models are a fundamental tool for forecasting demand in the tourism sector. Traditional time series models include techniques, such as the naïve model, exponential smoothing model, autoregressive–moving-average (ARMA) models, and structural time series model (36). These time series models offer several advantages in forecasting accuracy by their ability to predict using only historical tourism demand data, allowing them to be easy to implement and effectively predict trends and seasonality, however, they rarely consider the influencing factors of tourism demand, which may result in a loss of important information (50). This advantage can be considered a limitation, since these models are unable to take explanatory factors into account, which makes it difficult to incorporate changes in policies and events.

A study by Oh and Morzuch (33) applied time series models to predict the tourism demand for Singapore. They examined the monthly travelers' arrival and based their modeling on several determinants, including historic data of arrivals. The models used covered a wide range of approaches, such as Naïve I, Naïve II, Linear regression, Winters's model, Autoregressive Integrated Moving Average (ARIMA), and Sine-wave regression. Their conclusions were insightful, highlighting the influence of different performance statistics, such as mean absolute percentage error (MAPE), mean absolute error (MAE), root-mean-square error (RMSE), and the length of the forecast horizon, which can, in turn, choose different models as the best model. They also pointed out that the model that performs the best during the training set may not have the best performance in the test set and changing the length of the prediction period may have an impact on what the best model is. The best performing model on the test set was a combined model of two ARIMA models with an RMSE of 8,988.2. As a result of their study, the authors concluded that structural models generally provide less accurate forecasts compared to univariate models since they have more reasons for forecast error.

Lin et al. (26) forecasted the tourism demand in Taiwan using the aggregate number of tourists who arrive in Taiwan every month. In the study, they applied three methods: time series, ANNs and multivariate adaptive regression splines. To assess the performance of these models approaches, was employed metrics like RMSE, mean absolute deviation and MAPE for method comparison. In their study, time series models outperformed the other methods.

Econometric Models

Econometric models have a causal approach and seek to analyse the relationship between tourism demand (the output variable) and its influencing factors (input variables), such as economic, social, demographic, and more. The objective in econometric tourism demand forecasting is to find the most influential variables in tourism demand (15). The strength of econometric

models lies in their ability to estimate the contribution and importance of influencing factors related to income, prices, etc. In contrast to time series models, econometric approaches offer practical insights into what drives tourism demand. In addition to their advantages, they have some limitations, such as depending on the number of explanatory variables, these models require extensive data collection and forecasting can be sensitive to model specification, especially for complex models.

Hilaly and El-Shishiny (18) evaluated various econometric models and analysed the pros and cons of each. The advances in econometric modeling were discussed and they reach to the same conclusion of many studies that no single forecasting model can generate the most accurate forecasts in all situations. They also introduced some new models, such as the tourism technical analysis system model and AI techniques. Notably, they concluded that when explanatory variables can be obtained or estimated accurately, AI techniques tend to give the best results. However, for measuring the impacts of explanatory variables in economic and policy issues, econometric modeling approaches are more useful.

AI Models

AI and Machine Learning (ML) are increasingly becoming powerful tools in several sectors, particularly in areas of the economy where sales and marketing play a crucial role, according to a report by McKinsey Global Institute analysis (30). Tourism belongs in these sectors, which means that AI has an extraordinary potential for this sector, to better understand and resolve tourism-related issues (12)).

By using these methods and approaches, tourism can discover previously unknown patterns, correlations, insights and predict to answer future-oriented questions (44). However, also face challenges in their complexity and implementation, along with the "black box" nature of their results, which raises interpretability issues, as previously noted.

In AI models, historical time series data on tourism demand is required and explanatory variables can be incorporated, depending on the specific model used. Some common explanatory variables include economic variables (e.g., GDP, exchange rates, and consumer prices), demographic data (e.g., population size, age distribution), and social media data (e.g., tourist sentiment, travel intentions, and social media engagement).

- Regression models

Semeida (39) carried out a comparative study between multiple linear regression and generalised linear modeling to predict travel demand in low populated areas of Northeast Egypt. The

author showed that the generalised linear model is superior in terms of accuracy and considered factors such as the number of trips per person per year, distance, population, area, income, travel time, travel cost, and trip frequency.

The challenge of forecasting tourism demand in Hong Kong was addressed by Wu et al. (47), using a Gaussian process regression model. To forecast the monthly international tourist arrivals, their study tested many variables in the model, including travel demand from each origin country, income levels of these origins, destination prices, transportation costs, foreign exchange rates, and the population of the origin countries. The sparse model revealed that it was more effective in comparison to the ARIMA, v-SVM and g-SVM models.

- Support vector machines (SVM)

Pai and Hong (35) introduced a multi-factor SVM with the aim of forecasting annual number of visitors. The authors to increase prediction accuracy, combined the SVM with a neural network. Their approach considered variables such as the service price, foreign exchange rates, population, market expenses, gross domestic expenditure, and the average hotel rate.

Lin and Lee (27) investigated the forecasting of tourism demand in Taiwan, comparing different forecasting methods. Multivariate adaptive regression splines (MARS), ANN and SVR were implemented to forecast the monthly number of tourist arrivals. The average error rate revealed the superior performance of the SVR model over the other models, with a mean error rate of 3.61%. The ANN model presented a sub-optimal performance with a mean error rate of 7.08%, while the MARS model performed the worst with a mean error rate of 11.26%. The key factors in this study included average hotel prices, the number of hotel rooms, international flight capacity, GDP, consumer price index (CPI) and foreign exchange rates.

- XGBoost models

Li et al. (25) investigated the application of machine learning in analysing large volumes of social media data to predict tourism demand in Sydney, Australia. The authors used algorithms such as Extreme Gradient Boosting (XGBoost), SVR, Random Forest Regressor and Linear Regression as forecasting models to predict tourist arrivals. They concluded that the XGBoost model outperformed the others, with an MAE of 1432 and a MAPE of 3.93%, surpassing the other models which obtained an MAE metric of over 2000 and a MAPE of over 7%. In their study, variables relating to Christmas, the business environment, general news, security and the exchange rate were the top five predictors of tourist arrivals.

- ANN models

The challenge of forecasting tourism demand in the north of Portugal was addressed using ANN (13). Their model is based on using the previous 12 values to calculate each forecast value.

Tsai et al. (43) applied two multiple-temporal-unit neural networks (MTUNN) and parallel ensemble neural networks (PENN) to forecast the travel demand by railway passengers, focusing on daily and monthly passenger demand. These networks were constructed using two categories of data, including temporal features and level shifts. By comparing the networks with MLPs showed the better performance of the two networks by the mean squared error and the MAPE measures.

2.3 The Role of Interpretability in Tourism Demand Forecasting

In the context of tourism, many demand forecasting models have achieved high accuracy, but are often insufficient in terms of interpretability. Machine learning models, although highly accurate, remain incomprehensible to human understanding. These complex models fall into the category of black box models, which are systems that do not reveal their internal mechanisms and cannot be understood by examining their parameters (32). This insufficient explanation prevents the acceptance of research results in the tourism sector. Researchers face the challenge of making tourism demand forecasts interpretable in a compelling way and it is worth mentioning that there is a lack of articles that prioritize interpretability in existing tourism demand forecasting studies (46).

This limitation is crucial to providing a clear and transparent understanding of tourism demand forecasts, allowing the results to be easily communicated and understandable to tourism professionals, managers and other relevant stakeholders. Interpretability is highly relevant when you consider the complexity of the tourism sector, where making informed decisions is key to developing effective marketing strategies, managing destinations and allocating resources. Moreover, by prioritising interpretability, researchers can increase the trust and acceptance of the results.

It is important to note that, as Molnar (32) mentioned, the higher the interpretability of a machine learning model, the easier it is for a human being to comprehend why certain decisions or predictions have been made. For this reason, the Machine Learning community has started to turn to techniques focused on interpretability (20). These techniques usually adopt two approaches to reach interpretability. Firstly, there are ML models designed to be intrinsically interpretable (intrinsic) due to their simplicity, achieving interpretability by restricting the

complexity of the machine learning model, such as short decision trees or sparse linear models. Secondly, there are techniques that provide posthoc explanations for the predictions made by complex models, through the application of methods that analyse the model after training, including approaches such as permutation feature importance. Furthermore, the explanation of a machine learning model can be classified as local or global, depending on whether they explain individual predictions or the overall behavior of the entire model (32). Global explanations focus on knowing what patterns are present in general, while local explanations involve knowing the reasons for a specific decision (11).

3 Methodology

3.1 Data Collection

In this study, data collection focused on acquiring structured time series data from Turismo de Portugal. The dataset has a multidimensional granularity, with monthly data covering the years 2015 to 2022, and focuses exclusively on the geographical scope of Portugal. This approach extends to the municipal level, providing a localised perspective. And it also includes the classification of accommodation facilities in each municipality.

The data collection process involved identifying the sources of data that contribute to the study. Firstly, through direct contact with Turismo de Portugal, access to tourism data, such as supply and demand data, was ensured and ethical considerations were taken into account, such as the fact that information from some municipalities is subject to statistical confidentiality, so this data should only be used for the development of the thesis. And finally, additional variables were incorporated into the study through various sources detailed in Table 1.

Table 1: Description of the variables sources. These variables were incorporated to enrich the original dataset.

Variables	**Source**
Population	Pordata according to Censos
Bed Occupancy Rate	Instituto Nacional de Estatística (INE)
Number of Holidays	Python 'holidays' package
Lockdown	Ministério da Administração Interna (31))
CPI	INE
GDP	INE
Crime Rate	INE

3.2 Data Preparation

This time series dataset was examined, originally containing 84,408 observations and 11 variables, with the outcome variables being *overnight stays* and *bed occupancy rate*. Data quality assurance was fundamental throughout this process.

3.2.1 Data Preprocessing

The original dataset is clean, with no missing values, duplicates, or unusual outliers. All the potential outliers identified were considered reasonable, contextualized, and justified due to the unique characteristics of each municipality, recognizing the disparities inherent in the tourism industry. However, a peculiar occurrence was identified on a specific datapoint corresponding to rural tourism in the municipality of Soure in December 2015. This entry presented a value of 0 for crucial variables, namely *overnight stays*, *guests*, and the *number of beds*. To deal with this anomaly, it was decided to impute these values using the average of the previous values for each of the respective variables. The purpose of this imputation strategy was to preserve the time trend and maintain the consistency of the dataset.

For the purpose of this work, only *local accommodations*, *rural tourism*, *hotels*, *tourist apartments*, *apartment hotels*, *inns* and *tourist villages* were considered. Additionally, a new classification, *hotel establishments*, was introduced by grouping *hotels*, *apartment hotels* and *inns* according to Regime Jurídico dos Empreendimentos Turísticos (RJET) – Decreto-Lei n.º 39/2008.

Feature engineering was performed to enrich the original dataset. A *population* feature was added, representing the number of people in each municipality per year. *Bed occupancy rate* (BOR) was calculated based on formal concepts of tourism, capturing the relationship between overnight stays and the number of beds. The formula is as follows:

$$\text{BOR} = \frac{\text{Number of Overnight Stays}}{\text{Number of available beds} \times \text{number of days in the specific month}} \times 100 \quad (1)$$

Equation 1 - Formula for bed occupancy rate according to INE.

Three additional features, *number of holidays*, *lockdown*, and *season*, were incorporated. The former accounted for the number of national holidays in each month and year, while the latter introduced a binary feature signifying COVID-19 lockdowns. The days affected by lockdowns were treated as holidays, preventing their recurrence in future trend analysis. In 2020, the lockdown days were identified as April, May, November, and December, while in 2021, they were observed in January, February, March, and April. Additionally, the categorical feature *season* was included to represent the season of the respective month, which provides more

insights about potential seasonal patterns in the data.

Two key economic indicators were also included in the dataset. The Consumer Price Index (*CPI*) to measure the general price level in Portugal in the NUTS II regions. And Portugal's annual Gross Domestic Product (*GDP*), which was distributed evenly over the months, assuming a consistent GDP throughout the year. This data provides information on economic conditions in Portugal and can be useful in the modeling process, as mentioned in the section 2.

Finally, a new variable, *crime rate*, was introduced to assess the security levels across all municipalities over the available years. Given that the crime rate information is only provided on an annual basis, an assumption was made that the crime rate remains constant over the months, similar to the treatment of the *GDP* feature.

After the initial data pre-processing stages, the next crucial phase in preparing the dataset for tourism demand forecasting was to determine the optimum number of lagged inputs (NLIs). However the determination of NLIs can be a time-consuming effort (4), identifying lag relationships in time series data is critical for tourism demand forecasting, given the changing influence of features across different lags (22). The initial step involves selecting the lag-order of time series variables, often achieved through methods like the Granger causality test (24).

The Granger causality test was applied to assess whether the past values of the independent variables can predict the target variables *overnight stays* and *bed occupancy rate*. The lagged features further contribute to the analysis by adding the following considerations: lagged overnight stays (t-1) which represent the number of overnight stays in the previous month for the same municipality and classification; lagged independent variables (t-1) for the variables guests, number of establishments, number of beds and number of rooms, which reflect values from the previous month for the same municipality and classification.

With these transformations applied to the data, the dataset has suffered significant changes, resulting in a total of 76,873 rows and 25 columns. This reduction in rows is caused by the exclusion of the month of January 2015 due to the incorporation of lagged variables. The Table 2 provides a detailed description of each variable, including the respective data types.

Table 2: Data dictionary of the final dataset. Description of the variables with the respective datatypes.

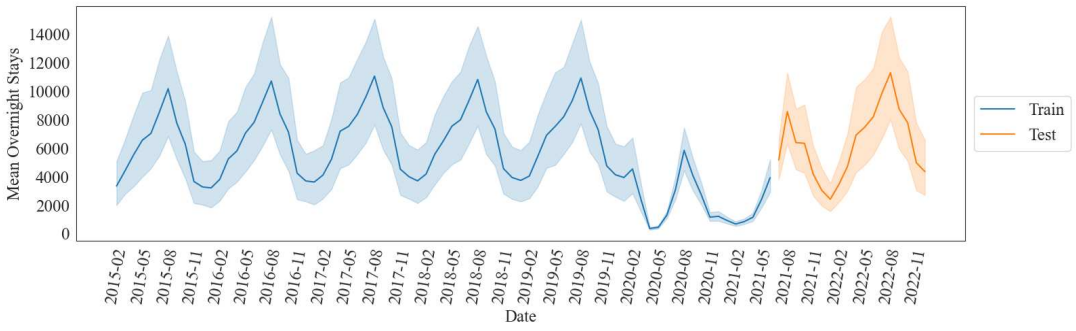
Variable	**Description**
Date (datetime64)	Date of the observation.
Year (int64)	Year associated with the observation.
Month (int64)	Numerical month of the year.
Name month (object)	Name of the month in full.
NUTSII (object)	Designation of NUTS level II.
Municipality (object)	Identification of the municipality.
Population (int64)	Number of people per year.
CPI (float64)	Monthly economic indicator.
GPD (int64)	Monthly economic measure of GDP.
Crime Rate (float64)	Monthly crime rate.
Number of Holidays (int64)	Number of national holidays.
Lockdown (int64)	Binary value indicating lockdown.
Season (object)	Classification of the season.
Classification (object)	Classification of the accommodation facility.
Overnight Stays (int64)	Total number of overnight stays.
Guests (int64)	Total number of individuals.
Number of establishments (int64)	Total number of establishments.
Number of beds (int64)	Total number of beds available.
Number of rooms (int64)	Total number of rooms available.
Bed Occupancy Rate (float64)	Percentage of beds occupied.

Note: The lagged features are also included and represent the previous month's values for the same municipality and classification of the current feature.

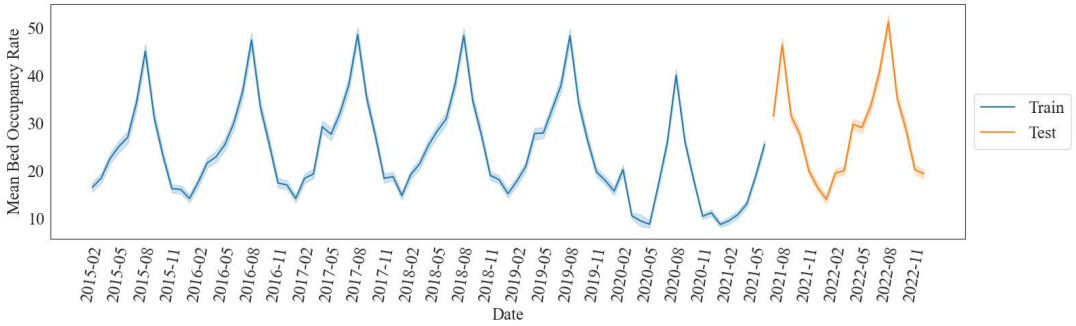
3.3 Modelling

In the modeling phase, the Extreme Gradient Boosting (XGBoost) algorithm was chosen as the forecasting model using the *xgboost package*. Is known for its robust implementation of gradient-boosting decision trees and stands out as an efficient and fast algorithm that can help understand complex data and make better decisions (25).

To apply this method, the dataset was divided into training (approximately 80% of total samples) and test sets (approximately 20% of total samples), with the training set containing data until June 2021 and the test set covering from July 2021 to December 2022 (Figure 1).



(a) Distribution of the sets for the models with the *overnight stays* as the target variable.



(b) Distribution of the sets for the model with the *bed occupancy rate* as the target variable.

Figure 1: Distribution of the train and test set. The train set contains data until June 2021 and the test set contains from July 2021 to December 2022.
Note: The split was the same for all the models. The X-axis represents dates, spaced at intervals of 3.

Motivated by the results of Oh and Morzuch (33), a simple univariate XGBoost model was chosen as a benchmark to evaluate the performance of the model that predicts the number of overnight stays. This reference is based on the relationship between the past and the present, using only overnight stays in the last months, *overnight stays (t-1)*, as a forecasting factor. The reason for this choice is due to the authors' view of the effectiveness of univariate models and the potential pitfalls associated with more complex models.

Next, a model with XGBoost algorithm was implemented with *overnight stays* as the target variable. The selected independent variables incorporated essential historical characteristics related to tourism in Portugal. These included features such as the *number of establishments* and the *number of beds*, both used in the current months, along with their *lagged versions* from previous months. Additionally, the model incorporated features that capture temporal aspects, including the *year*, *month*, *season*, *lockdown* and the *number of holidays*. And other indicators,

namely the *crime rate* and the *population* described in Table 2 from section 3.2.1. To fit this model, the objective function was set to 'count:poisson', since is tailored for count data like overnight stays. A randomized search with a Time Series Split with 5 splits was employed to optimize hyperparameters.

After a varied benchmarking process and by investigating numerous models and parameters, the XGBoost algorithm was used not only to train models with an absolute target, such as using *overnight stays* as the target, but also to develop a model with a relative target variable, such as the *bed occupancy rate* in Portugal, since there are challenges related to modeling with the absolute target model. This model uses the same set of features as the model with *overnight stays* as the target to predict tourism demand. To fit this model, the objective function was set to 'reg:gamma', due to the nature of the target variable- *bed occupancy rate*. In the Appendix, Figure 22 displays the left-skewed distribution of this variable, ranging from 0 to 100, which makes 'reg:gamma' more suitable for capturing the characteristics of the data. A randomized search was employed to optimise hyperparameters, using a Time Series Split with 5 splits to deal with the temporal nature of the data. This decision was motivated by the temporal dependencies present in the dataset, ensuring that the model is evaluated on different time periods during the cross-validation process. In addition, a post-processing step was introduced to refine the predictions. A function was implemented to adjust the model's predictions to the valid range of [0, 100]. Given the nature of the *bed occupancy rate*, this step ensures that the predictions align with the requirements of the target variable.

The hyperparameters selected for the two target models, *overnight stays* and *bed occupancy rate*, were based on key parameters such as the learning rate, the maximum depth of the tree and the regularisation terms:

- **Base score:** represents the initial prediction value on the average value of the response variable before building the trees. It can be adjusted during tuning, allowing the model to explore different starting points and assess their impact on overall performance.
- **Learning rate:** controls the shrinkage of each tree's contribution. Smaller values require more iterations but can improve generalization. It can explore different step sizes to find the ideal balance between model accuracy and computational efficiency.
- **Number of estimators:** determines the number of trees to be built, increasing this number may improve performance but also increase training time which can lead to overfitting.
- **Maximum depth:** defines the maximum depth of each tree. Deeper trees can capture more complex interactions but may overfit. Smaller values were used to assess the complexity of the model, avoiding very deep trees that could lead to overfitting.

- **Subsample rate:** controls the fraction of samples used for fitting each tree. Smaller values introduce randomness and can prevent overfitting. The grids chosen, present in Table 3, allow the model to be trained on a smaller fraction of data in each iteration, avoiding overfitting.
- **Colsample by tree:** controls the fraction of features used for training each tree. Like subsample, it introduces more randomness and helps to reduce overfitting.
- **Alpha and Lambda :** introduce L1 and L2 regularisation to control the complexity of the trees. They help to prevent overfitting by penalising large coefficients.
- **Minimum child weight:** influences the minimum sum of weights needed in a child. The grids were chosen to control the minimum weight needed in a child node, avoiding partitions with too few instances.
- **Gamma value:** specifies the minimum loss reduction required to make a further partition on a leaf node. The grids were selected to control the minimum loss reduction needed to make an additional partition, adding regularisation to the model.

Table 3 represents the various combinations of values chosen for the key parameters.

Table 3: Parameters chosen for the models. Parameters with a description of the corresponding value grids.

Parameter	**Model Overnight Stays**	**Model BOR**
Base score	[0.5, 0.6, 0.7]	[0.5, 0.6, 0.7]
Learning rate	[0.01, 0.05, 0.1, 0.15]	[0.01, 0.05, 0.1, 0.15]
Number of estimators	[50, 100, 150, 200]	[50, 100, 150, 250, 300, 350]
Maximum depth	2 to 6	2 to 10
Subsample rate	0.6 to 0.9	0.6 to 1
Colsample by tree	[0.6, 0.8, 1]	[0.6, 0.8, 1]
Alpha and Lambda	[0.1, 0.2]	[0.1, 0.2]
Minimum child weight	[5, 10, 20]	[5, 10, 20]
Gamma value	[0, 0.1, 0.5]	[0, 0.1, 0.5]

Note: The BOR model corresponds to the model in which *bed occupancy rate* is the target variable.

Finally, to evaluate the performance of the XGBoost models, error metrics were used, such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE) and Coefficient of determination (R^2), represented in Table 4. The RMSE measures the average magnitude of errors between predicted and actual observations and penalises larger errors more severely due to the quadratic term. The MAE calculates the average of the absolute errors, providing a direct measure of the magnitude of the errors without squaring them, making it less sensitive to extreme values, and penalises errors in the same way as the RMSE. MAPE expresses the errors as a percentage of the actual observations, penalising in proportion to the scale of the observations, which means it penalises larger actual values more, but can be sensitive to cases with small actual values. The R^2 represents the proportion of the variance explained by the model, ranging from 0 to 1, where higher values indicate a better fit. The R^2 penalises unexplained variance, highlighting the quality of the fit.

As indicated by the section 2, the most used metrics to this type of problem include MAE, MAPE and RMSE. This preference arises from certain limitations associated with R^2, despite being widespread use. R^2 is very sensitive to outliers, where extreme values can have a large impact on the metric, distorting the interpretation of the model's performance.

In the context of tourism, events such as pandemics often lead to extraordinary or unusual values, influencing the model´s performance significantly and resulting in inaccurate evaluations. Additionally, the R^2 provides an overall measure of model fit, but may not effectively highlight performance variations across different data segments. In tourism forecasts, demand can vary significantly between regions, seasons, or types of establishments. This metric may not fully capture the variations in these diverse conditions, complicating the understanding of the model's effectiveness in specific contexts.

Therefore, when using R^2, it is crucial to extend the analysis with other metrics to obtain a more comprehensive evaluation of the model's performance. This ensures an extensive evaluation, especially given the unique challenges and characteristics associated with forecasting tourism demand. It should also be noted that in order to compare models with different target variables, which in turn have different types of representation, it is essential to use metrics that take into account the scale of the data and are not influenced by its magnitude. So, to assess the performance of these models, only metrics that are robust to variations in scale will be considered, namely MAPE and R^2.

Table 4: Error metrics to evaluate performance. MAE, MAPE, RMSE and R^2 and their respective formulas.

Metric	**Formula**
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
MAPE	$\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{y_i}$
RMSE	$\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$
R^2	$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

3.4 Interpretability

As a next step, an explanatory tool was implemented to tackle the challenge mentioned in the section 2 concerning insufficient explanations in the tourism sector.

Interpretability tools have been developed to help data scientists and machine learning practitioners better understand of how ML models work (20). The most popular methods are LIME and Shapley values and provide posthoc explanations, allowing researchers to identify the importance of the input variables and, in turn, to add an interpretation to the model prediction (8).

The tool used was SHAP (SHapley Additive exPlanations) via the *SHAP package*, that is widely used and provide both local and global explanations. SHAP is a post-hoc explanation technique for black box ML models, which offer a game-theoretic approach to explain the output for any ML model. This approach assigns importance scores to each feature for each prediction (28). These scores are based on the notion of Shapley values from cooperative game theory, they ensure a fair distribution of "credit" over the input features for each prediction, also these scores elucidate how each feature impacts predictions. Features with positive SHAP values positively impact the prediction, while those with negative values have a negative impact. The magnitude is a measure of how strong the effect is (20).

The Shapley value is the only method that satisfies properties of Efficiency, Symmetry, Dummy and Additivity (32).

- Efficiency: The feature contributions must add up to the difference of prediction and the average prediction.

- Symmetry: The contributions of two feature values should be the same if they contribute equally to all possible coalitions.
- Dummy: A feature that does not change the predicted value should have a Shapley value of 0, regardless of which coalition of feature values it is added to.
- Additivity: In the context of multiple predictions functions, the summation of Shapley values can be calculated for each prediction or using both prediction functions.

The Tree SHAP algorithm was specifically used as a specialized implementation of SHAP values, since it was designed for tree-based models, such as XGBoost. This choice was made to ensure accurate and efficient explanations for the model's predictions in the context of tourism demand forecasting, providing information on the contribution of each feature to a specific prediction.

4 Results

4.1 Exploratory Data Analysis

The foundation of study's analysis lies in understanding the temporal evolution of demand. The behavior of demand, reflected in overnight stays over the period analysed, shows through Figure 2 a pattern of relative stability with small improvements initially until 2020. The year 2017 stands out as the year with the highest demand between 2015 and 2022, with an average of approximately 6844 overnight stays. A significant disruption occurred in 2020 and 2021, as the global COVID-19 pandemic and associated restrictions led to a drastic decrease in demand. These difficult circumstances led to a massive decrease in overnight stays during these years. In 2022, demand recovered, marking a return to some appearance of normality, and recorded an average of approximately 67245 overnight stays.

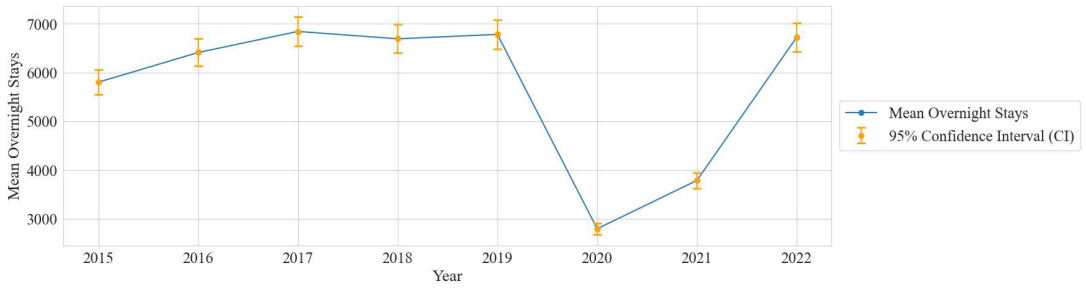


Figure 2: Demand behavior over time. Mean Overnight Stays per year shows a stable pattern across 2015 to 2019. In 2020, due to COVID-19 there was a drop and, in 2022, a recovery with a result returning more similar to previous years.

In Figure 3 shows that demand for overnight stays has a seasonal pattern, with fluctuations observed over the months of each year. Seasonality is present, characterized by pronounced peaks during the summer months. This increase in demand is in line with typical tourist behavior, as individuals often seek accommodation during the summer season.

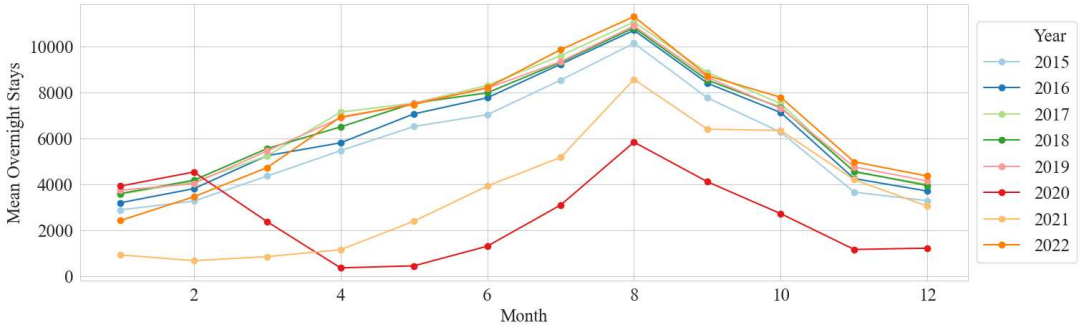


Figure 3: Demand behavior over the months of each year. Mean Overnight Stays per month indicates a seasonal pattern throughout the month. Seasonality is present during the summer months, which is as expected.

From 2015 to 2022, the total number of overnight stays in the various regions (NUTS II) reveals distinct patterns in tourism activity. Algarve takes the lead, with a substantial total of 135.7 million overnight stays. Lisbon follows, with a significant 112.1 million overnight stays, which highlights the capital's attraction for visitors. On the other hand, Alentejo has a comparatively lower number of overnight stays, with a total of 19.3 million (Figure 4).

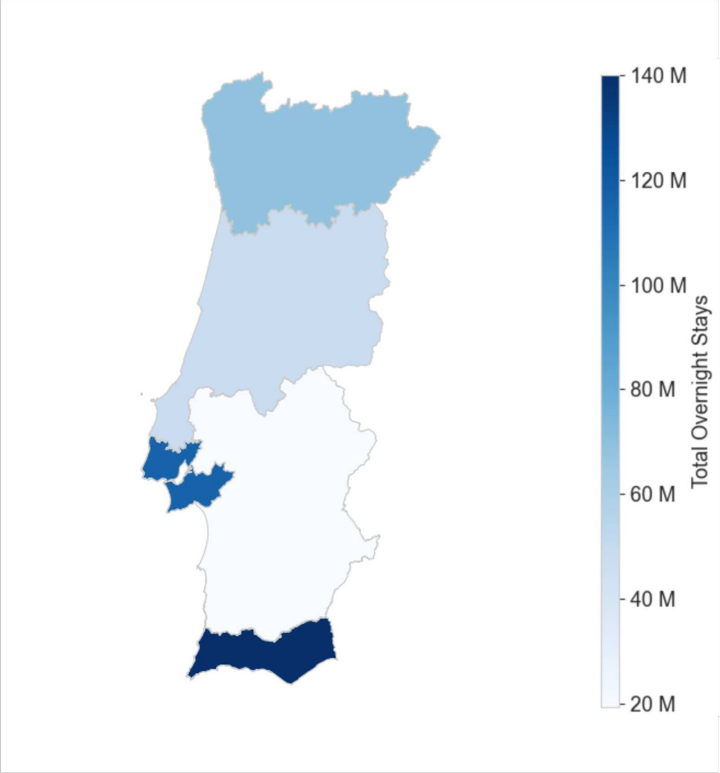


Figure 4: Total overnight stays by NUTS II in mainland Portugal (2015-2022). Total Overnight Stays over the years show that Algarve is the region with the most overnight stays, followed by Lisboa.

When analysing the municipalities, Lisboa stands out as the municipality with the highest demand, reflecting its popularity among tourists. Also is observed in Figure 5 that warmer weather tends to attract more visitors. The peak of demand was observed in August 2019, with a high average aligned with expectations of 481329 overnight stays.

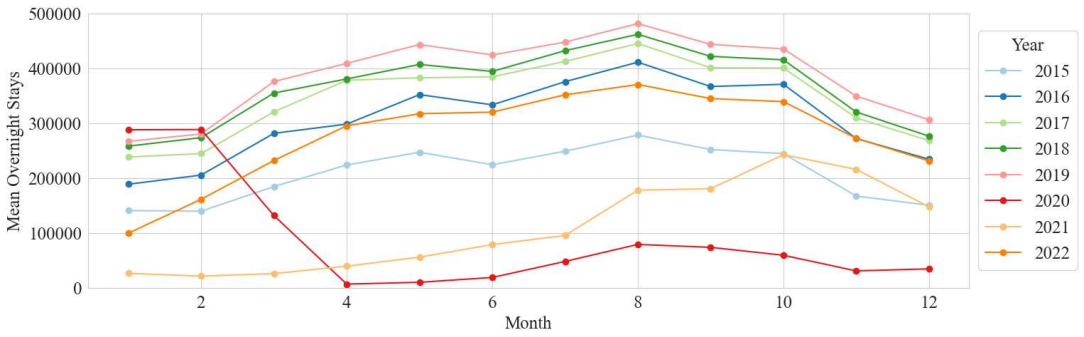


Figure 5: Demand behavior for Lisboa over the months of each year. Mean Overnight Stays per month show that, in almost every year, the summer months are the ones chosen by tourists to visit Lisbon.

The frequencies of the classifications of tourism data, represented in Figure 6, reveal a diverse range of accommodation. The majority of the data is attributed to local accommodations, representing 31.3%. Rural tourism follows close behind, comprising 31.0% of the data, with 24091 observations. Hotel establishments account for 29.6%. Tourist apartments, with 5.4%, and tourist villages, with 2.7%, represent smaller but significant segments. Despite the frequency of accommodation classifications, hotel establishments emerge as tourists' preferred choice, collecting a total of around 327.0 million overnight stays during the period from 2015 to 2022. This is followed by local accommodation, with a notable difference of 269.3 million overnight stays, which means that it registers a total of 57.7 million overnight stays (Figure 7).

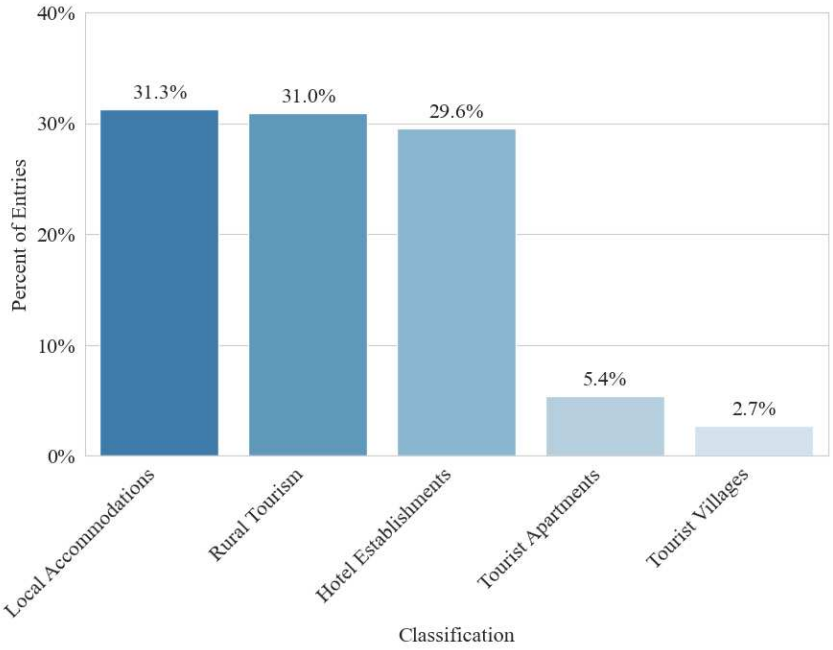


Figure 6: Classification frequency of accommodation facilities (%). The percentage of entries in the data for each classification from 2015 to 2022 shows that local accommodations is the accommodation facility with the most data, followed by rural tourism with 31% of the data.

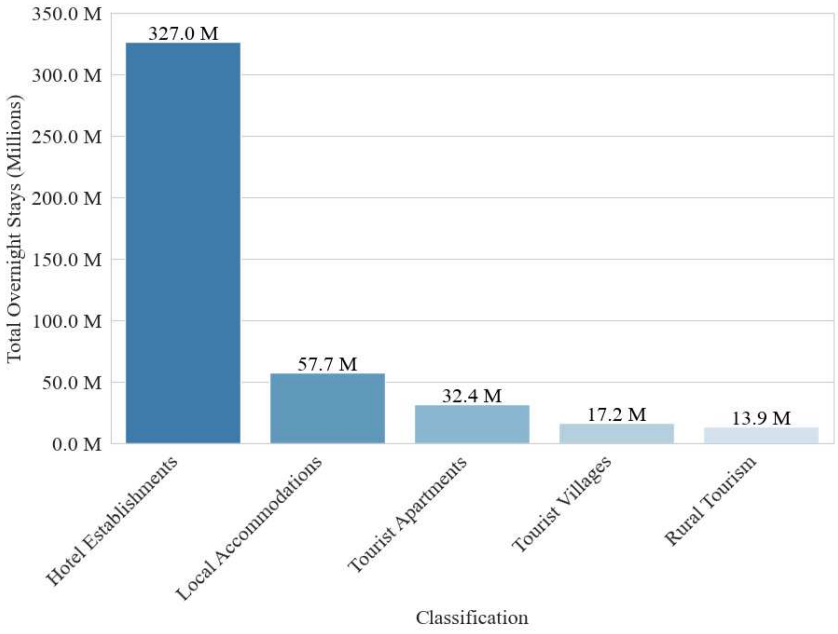


Figure 7: Demand behavior by type of accommodation facility. Total number of Overnight Stays from 2015 to 2022 shows that hotel establishments are the preferred choice by tourists, with a major difference in the number of overnight stays compared to the other classifications.

As can be observed in Figure 8, the mean bed occupancy rate reveals a considerably expected decline during the pandemic years, indicative of the challenges faced by the tourism industry. However, in 2022 a positive trend emerges, showing a collective recovery in all classifications, with an emphasis on hotel establishments. It can also be seen that hotel establishments consistently exhibit a prominent position, being the classification with the highest number of occupied beds in most years. A deviation from this trend occurred in 2019, when tourist villages outperformed all classifications in every year, registering the highest average bed occupancy rate of approximately 40%.

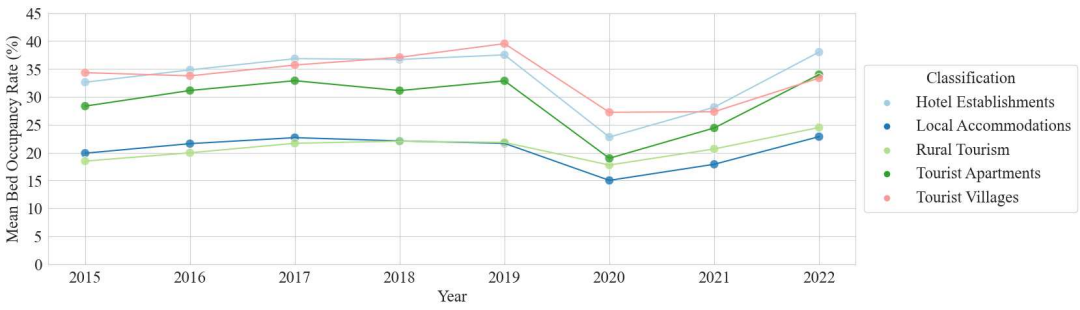
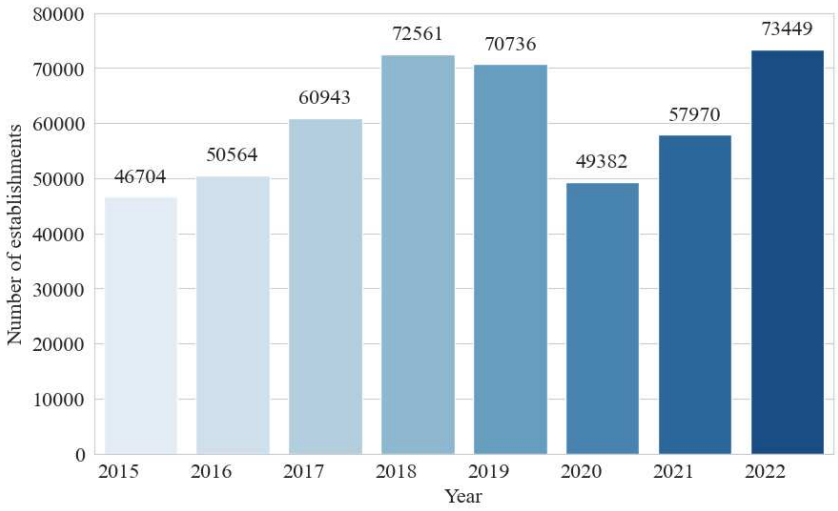
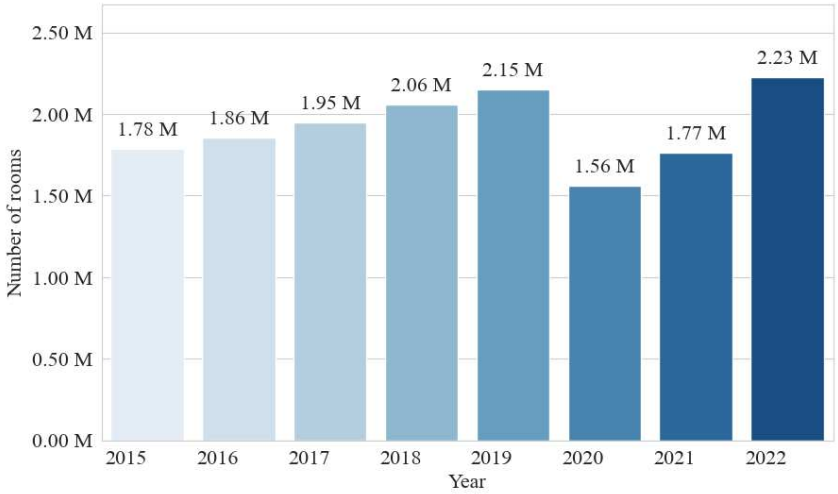


Figure 8: Bed occupancy rate by classification over time. Mean bed occupancy rate per year shows an expected drop during the pandemic years. In 2022 there was a recovery in this rate in all classifications.

Looking at the supply side of tourism, the recovery from the serious crisis generated by COVID-19, which severely affected all activities linked to tourism, is noteworthy. Despite this, 2022 was the year in which the most increases were observed in the main indicators relating to the supply of tourist accommodation, including the number of establishments and the number of rooms, showing the highest values for these indicators. This positive trend reflects a recovery in the tourism industry, signifying a gradual return to normality and an encouraging revitalization of tourism-related activities (Figure 9).



(a) Number of Establishments available over time.



(b) Number of Rooms available over time.

Figure 9: Supply behavior over time. The total number of establishments and rooms available per year shows a notable recovery in 2022, after the impacts of the COVID-19 pandemic.

Through descriptive analysis, relationships within the dataset were explored. Figure 10 shows that the variables *number of rooms*, *number of beds* and *guests* and their respective lagged variables exhibit the highest correlations with the target variables *overnight stays* and *bed occupancy rate*, indicating their significance in predicting demand. In addition, the variables *population* and *lockdown* are also included in the highest correlations with the *bed occupancy rate*. The presence of multicollinearity was observed and verified in some variables through the high variance inflation factor (VIF). In addition, the associations between the categorical variables were assessed using Cramer's V. The variable *classification* showed the greatest association with *overnight stays*, with a value of approximately 0.31. The variable *season* was the most associated variable with *bed occupancy rate* with a value of 0.23.

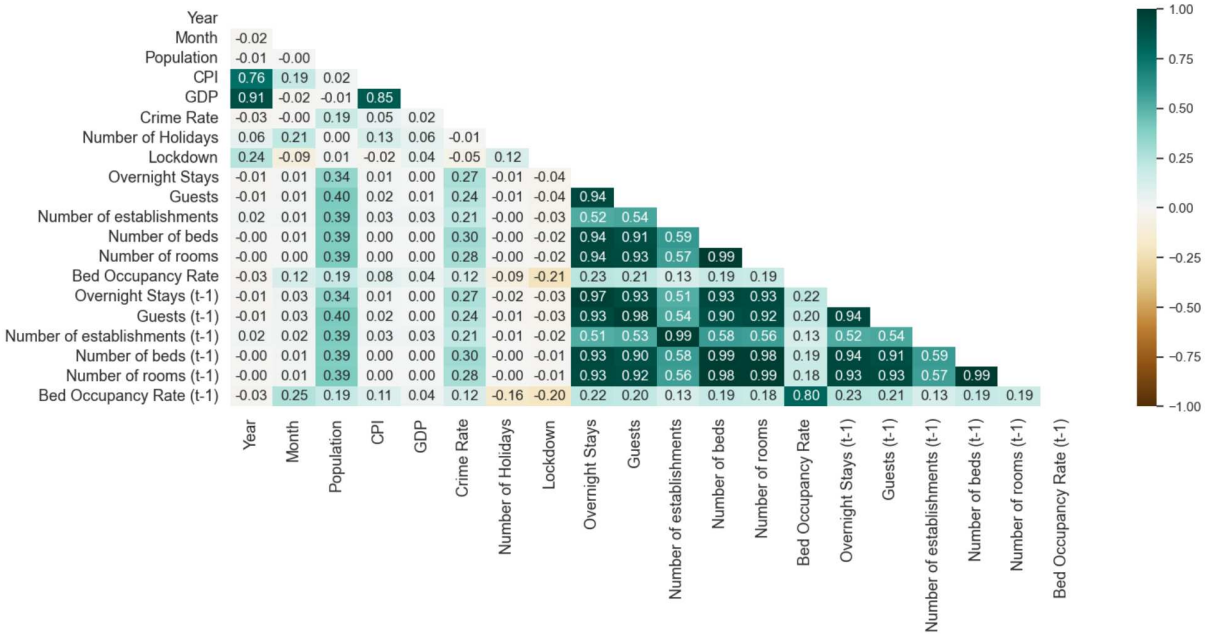


Figure 10: Triangle Correlation Heatmap. Through observing the correlations between variables show that the variables *number of rooms*, *number of beds* and *guests* are the ones with the highest correlations with *overnight stays* and *bed occupancy rate*.

By examining the key variables of interest based on the high correlations of these variables with targets and the targets themselves (Figure 10), it can be seen in Table 5 that the *population* and *crime rate* variables registered a mean of 36437 and 2.4‰, with standard deviations of approximately 60888 and 0.8‰, respectively. The *overnight stays* and *guests* recorded a mean of approximately 5760 and 2161, with standard deviations of 36977 and 13409, respectively. The *number of establishments*, has a mean of around 6, accompanied by a standard deviation of 19. The variables *number of beds* and *number of rooms* reveals a mean of 444 and 197, along with a standard deviation of 2041 and 927, respectively. These standard deviations indicate that there is a diversity in these variables across establishments. Finally, the *bed occupancy rate* shows a

mean of 25%, with a standard deviation of 20%.

Table 5: Descriptive analysis of the key numeric variables of interest. The key variables of interest, based on the correlations, are *population*, *crime rate*, *overnight stays*, *guests*, *number of establishments*, *number of beds*, *number of rooms* and *bed occupancy rate.*

Variables	**Mean**	**Standard Deviation**	**Minimum**	**Maximum**
Population	36,437.11	60,888.26	384	558,461
Crime Rate (‰)	2.39	0.83	0.72	6.82
Overnight Stays	5,760	36,977	1	1,164,529
Guests	2,161	13,409	1	469,083
Number of establishments	6.20	19.16	1	666
Number of beds	443.97	2,041.66	2	51,826
Number of rooms	197.24	926.76	1	24,868
Bed Occupancy Rate (%)	25.34	19.67	0.03	100

4.2 Occupancy Forecasting

As mentioned in section 3.3, a benchmark was carried out and the results of the models were compared, predicting demand in the test set and evaluating the model's performance by comparing the predicted demand with the actual demand.

The resulting models were fitted using a randomized search with a combination of parameters to improve performance in predicting demand, using *overnight stays* and *bed occupancy rate* as target. The selection of the hyperparameters was done by choosing the best cross-validated RMSE score. The optimal configuration of the best parameters is displayed on Table 6.

Table 6: Best parameter configuration of the models.

Parameter	**Model Overnight Stays**	**Model BOR**
Base Score	0.6	0.7
Learning Rate	0.1	0.1
Number of estimators	200	400
Maximum depth	5	9
Subsample rate	0.9	0.9
Colsample by tree	0.8	0.8
Alpha and Lambda	0.2	0.2
Minimum child weight	20	10
Gamma value	0.5	0.5

Based on the performance metrics results represented in Table 7, the Overnight Stays model outperforms the benchmark model in all metrics, except MAPE: lower RMSE, indicating fewer incorrect predictions; significantly lower MAE, which indicates a smaller absolute error and a higher R^2, indicating a larger proportion of the variance is explained by the model. The MAPE is slightly higher, reflecting a higher percentage of error relative to the actual values. So, it should be noted that the Benchmark model performs worst, which contrasts with the point of view of Oh and Morzuch (33).

When analysing the Overnight Stays model, the model has a much higher RMSE and MAE values compared to the Bed Occupancy Rate (BOR) model, indicating a greater deviation between the predicted and actual values for overnight stays. However, as noted in section 3.3, the performance of this model can only be meaningfully compared with the others using MAPE and R^2, with MAPE being favoured due to consideration of the limitations associated with R^2, as detailed in section 3.3. This suggests that the Overnight Stays model provides a less accurate relative representation of the data, due to the higher MAPE, but according to R^2 it captures the variance slightly better.

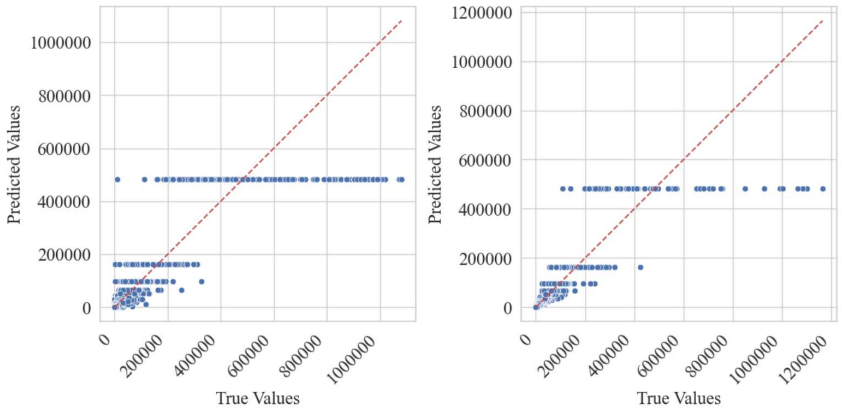
Table 7: Performance metrics results of the models. Benchmark, Overnight Stays and the BOR model.

Performance Metric	**Benchmark Model**	**Overnight Stays Model**	**BOR Model**
RMSE	16300.96 / 17419.51	13003.55 / 13183.51	7.14 / 7.14
MAE	1959.82 / 2159.23	1537.95 / 1664.63	4.86 / 5.04
MAPE	1.53% / 0.98%	1.62% / 1.08%	0.41% / 0.32%
R^2	0.80 / 0.79	0.88 / 0.88	0.87 / 0.87

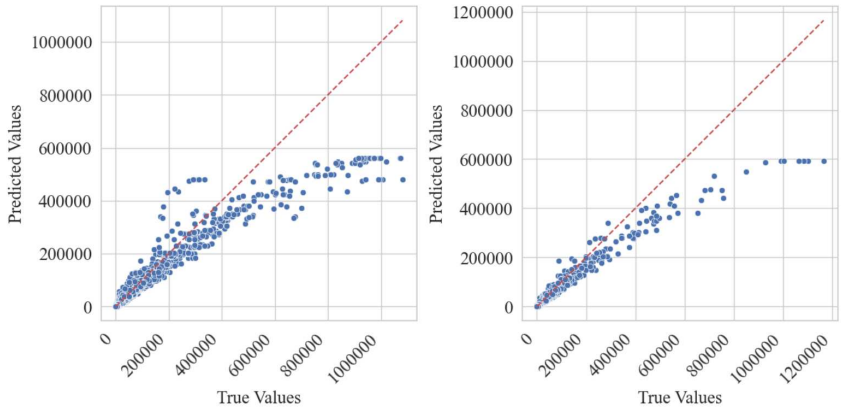
Note: The metrics represent the performance on the training set / testing set.

The superior performance of the Overnight Stays model compared to the benchmark model is more evident when examining the residual plots. For lower values, the benchmark model demonstrates lower accurate predictions, resulting in larger errors. Moreover, at higher values, both models face challenges in accurately predicting the number of overnight stays, with the benchmark model facing even greater difficulty in this regard (Figures 11a and 11b). Although the overall performance of the Overnight Stays model is robust, it struggles when confronted with higher values, particularly from around 400,000 overnight stays, displaying a pattern similar to a logarithmic curve, as illustrated in Figure 11b. This issue arises due to the significant disparities in overnight stays, ranging from 1 to 1,164,529, as can be seen in Table 5. The presence of these high overnight stays introduces complexity into the model.

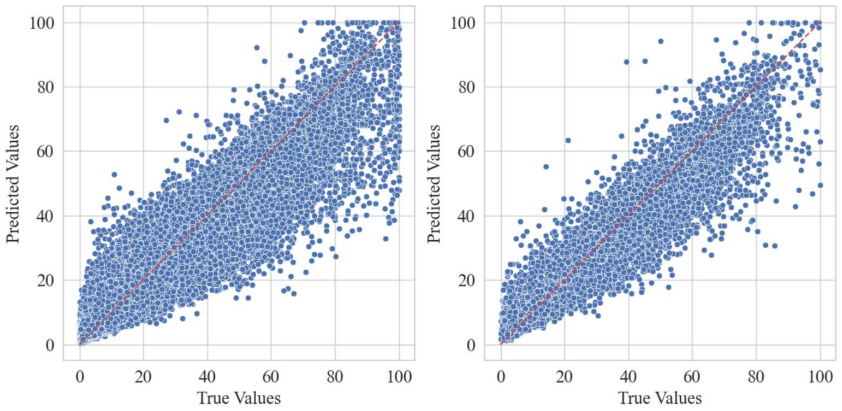
As a solution to overcome this limitation, a transition to using the Bed Occupancy Rate (BOR) model is proposed. The *bed occupancy rate* serves as an alternative target variable, offering the advantage of normalization and the limit of a range from 0 to 100. This approach aims to mitigate the challenges associated with the wide range of overnight stays by providing a more manageable and normalized metric for model training and forecasting. By looking at the Figure 11c, this model revealed to be more efficient than using *overnight stays* as a target.



(a) Residual Plots for Train set (left) and for Test set (right) of Benchmark model.



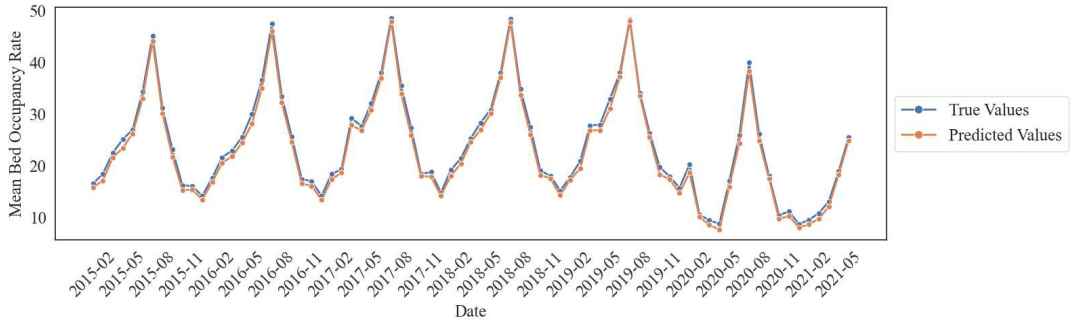
(b) Residual Plots for Train set (left) and for Test set (right) of Overnight Stays model.



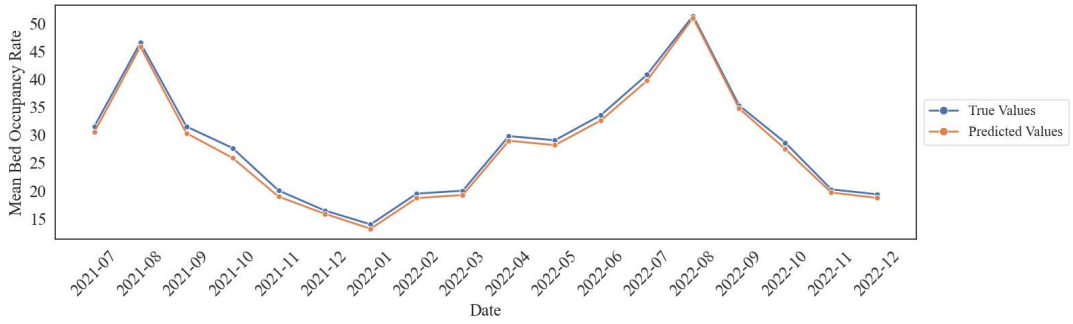
(c) Residual Plots for Train set (left) and for Test set (right) of BOR model.

Figure 11: Comparison of Residual Plots between the models. The residual plots show superior performance of the Overnight Stays model compared to the benchmark.

By analysing only the chosen model, BOR model, the predictions obtained by the model can be seen in the training set and the test set through Figure 12a and Figure 12b which show the performance of the predictions over the months of the training set and the test set using the mean bed occupancy rate. It reveals that the model predicts values consistently, suggesting a good capture of patterns over time in both sets. When there is a greater divergence of values, the predicted values tend to be slightly higher than the actual values for the bed occupancy rate. Furthermore, when comparing the two sets, the model shows significantly more accurate predictions in the training set.



(a) True vs Predicted values on Train set over time.
Note: The X-axis represents dates, spaced at intervals of 3.



(b) True vs Predicted values on Test set over time.

Figure 12: Comparison of True vs Predicted values on Train and Test sets over time. Mean bed occupancy rate of the predicted and true values for each month of each year in the train and test set show that the model has an overall good performance in both sets, with marginally better predictions in the train set.

Figure 13 demonstrates the most substantial prediction errors in the municipalities on test set, with Cadaval, Monforte, Salvaterra de Magos and Ourique being the municipalities with the most pronounced discrepancies, registering a mean MAPE of approximately 167%, 158%, 125% and 124%, respectively. Moreover, it should be noted that in these municipalities, local accommodations show the highest mean errors among all classifications for Cadaval and Ourique. For Monforte, the classification tourist villages is the one with more errors and for

Savaterra de Magos is rural tourism (Figure 14). It should be noted that Cadaval and Salvaterra de Magos only have information available on local accommodations and rural tourism, Monforte has no hotel establishments in the period analysed in the study, and Ourique has no tourist villages. This emphasises the imbalance of data between classifications, as can be seen in Figure 6.

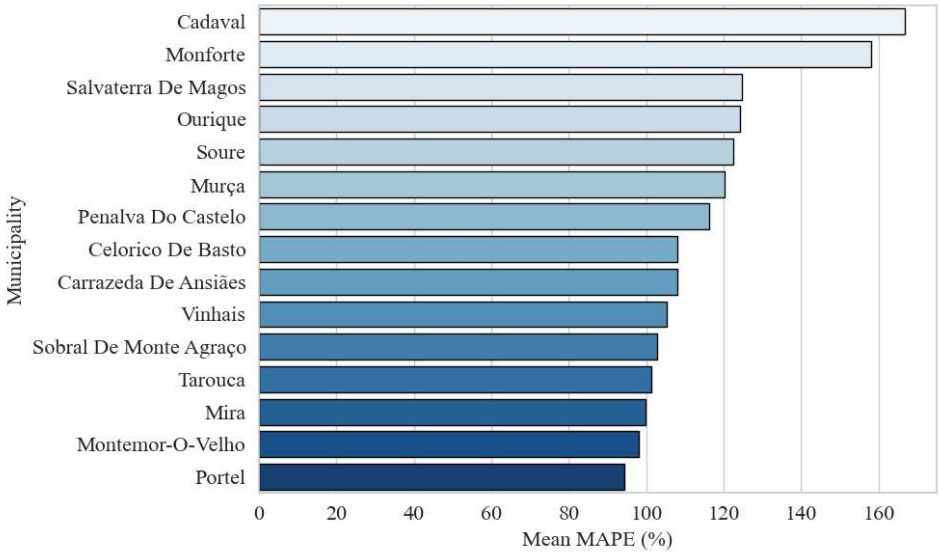


Figure 13: Distribution of the mean errors for the 15 municipalities with high error on Test set. The mean errors show that Cadaval, Monforte, Salvaterra de Magos and Ourique are the municipalities with the most associated errors.

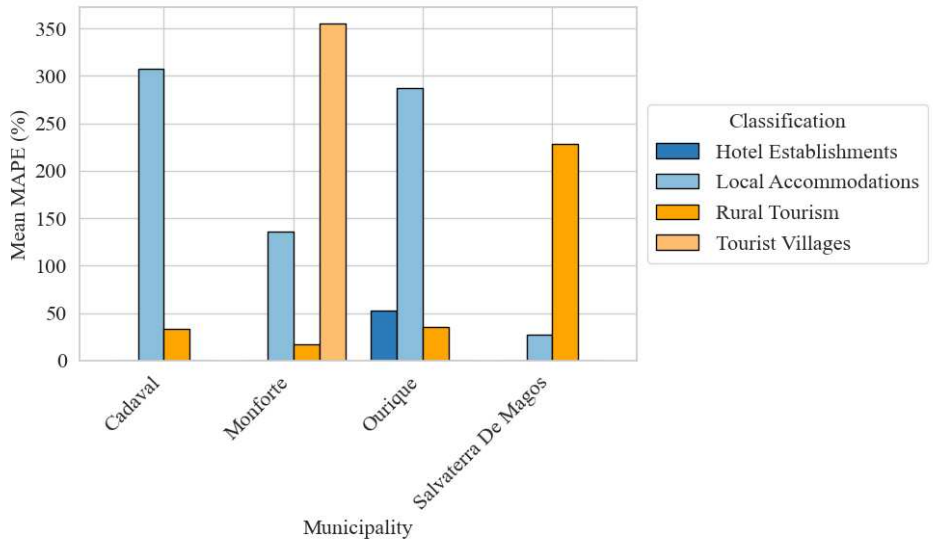


Figure 14: Distribution of the mean errors by classification for the 4 municipalities with high error on Test set. For Cadaval and Ourique, local accommodations are the classification with more errors.
Note: None of the municipalities in the Figure have any information available about the tourist apartments classification.

In addition, the Figure 15 shows that tourist villages have the highest average MAPE, approximately 56%. This is followed by local accommodation with a value of 37%, while the accommodation with the smallest error is hotel establishments, with a value of around 21%.

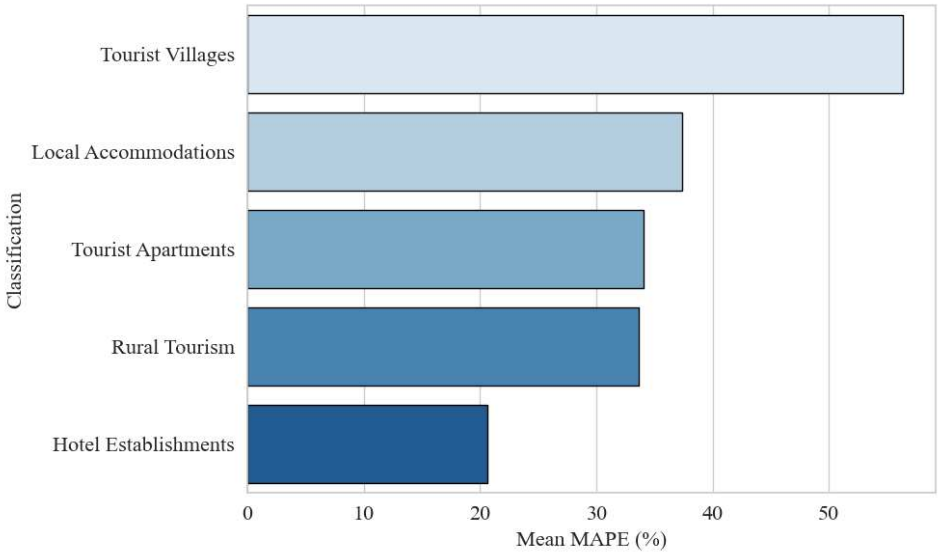


Figure 15: Distribution of the mean errors for the classifications on Test set. The mean MAPE of the classifications show that tourist villages have the highest mean error.

On the other hand, Figure 16 shows the municipalities with the most correct forecasts, where the municipalities of Matosinhos, Porto, Calheta (R.A.M.), Valongo and Ribeira Grande hold the best positions, with mean MAPE of less than 10%.

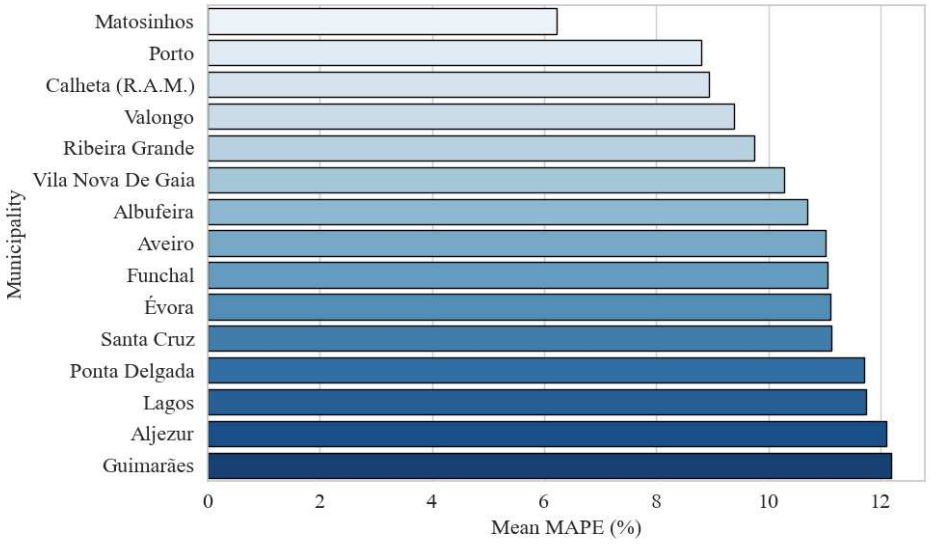
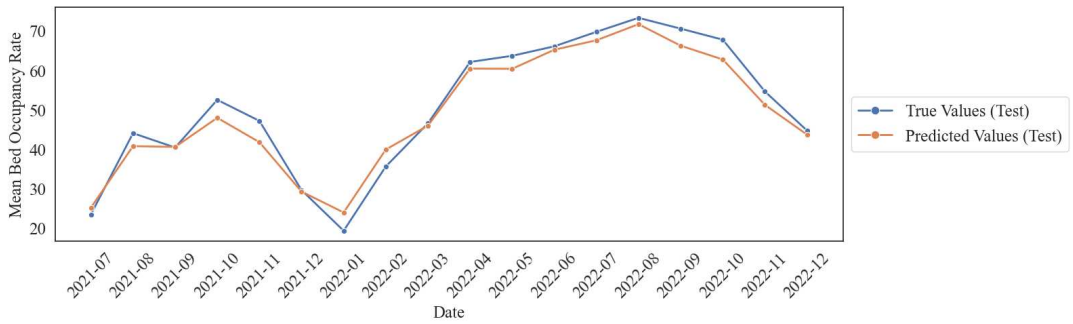
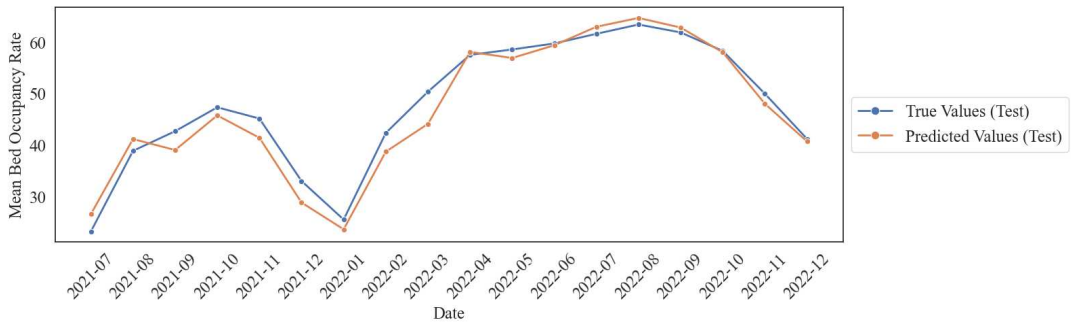


Figure 16: Distribution of the mean errors for the 15 municipalities with low error on Test set. The mean MAPE show that Matosinhos, Porto, Calheta (R.A.M.), Valongo and Ribeira Grande are the municipalities with less errors.

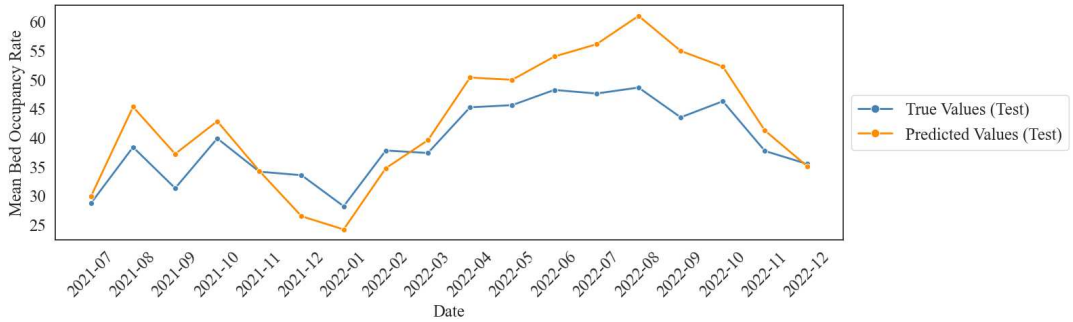
Given that Lisboa is one of the destinations most preferred by tourists, as can be seen in Figure 5, an investigation was carried out for each accommodation classification, and it should be noted that Lisboa has no information on tourist villages. Hotel establishments and local accommodations emerged as the categories with less residuals in the forecasts, indicating a more accurate capture of true value patterns. Conversely, for the tourist apartments and rural tourism, the predictions show a greater distance between the predicted and true values, suggesting less accurate results. The challenges in making accurate forecasts are closely linked to the complexity of the data. Rural tourism, faces additional complexity due to limited training data-available only for 2015, August to December 2020, January, May and June of 2021. This imbalance makes it challenging to forecast accurate values (Figure 17).



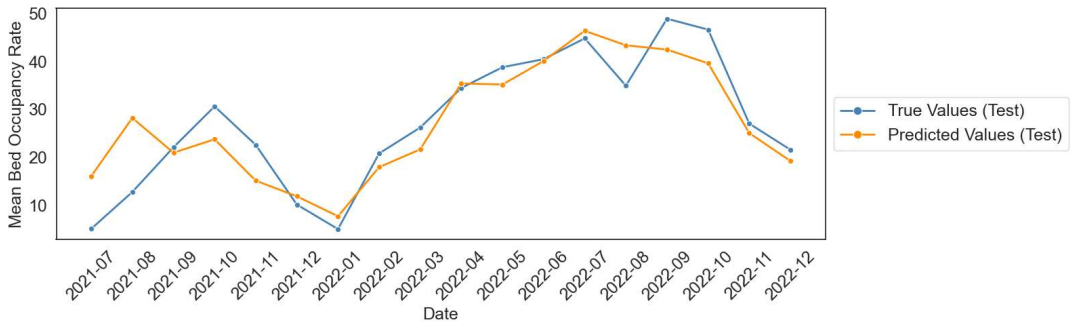
(a) True vs Predicted values for hotel establishments in Lisboa on Test set over time.



(b) True vs Predicted values for local accommodations in Lisboa on Test set over time.



(c) True vs Predicted values for tourist apartments in Lisboa on Test set over time.



(d) True vs Predicted values for rural tourism in Lisboa on Test set over time.

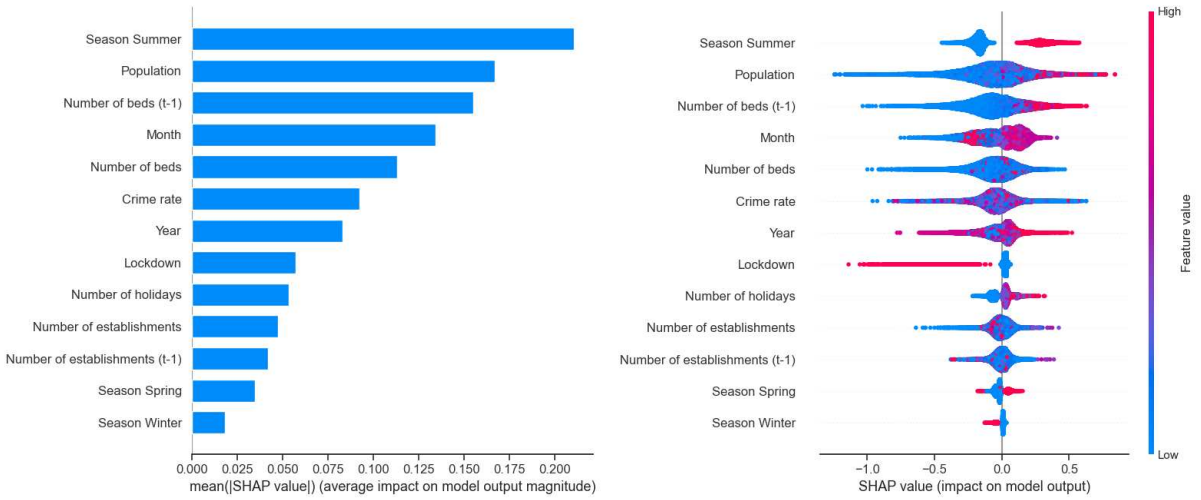
Figure 17: True vs Predicted values for the accommodation facilities in Lisboa on Test set over time. The mean bed occupancy rate for each month of each year for different accommodation facilities.

4.3 Model interpretability

The Tree SHAP outputs, represented in Figures 18a and 18b, offer a comprehensive analysis of each feature's contribution to the predictions built by the XGBoost model, providing valuable insights to further investigate the key features that influence tourism demand. By aggregating the importance scores for predictions, can be produced global explanations by averaging the magnitudes of SHAP values across the entire dataset. The density scatter plot of SHAP values for each feature reveals the extent to which each feature influences the model output for individual data points.

With these Figures, it can be seen that, the *season summer*, *population*, *number of beds (t-1)* in the previous months, *month* and *number of beds* in the current month emerge as the features that play major roles in determining the results, making them essential inputs. On average, the *season summer* leads to a change of 21 percentage points in the forecast result, followed by the *population*, which causes a change of 17 percentage points. The *number of beds (t-1)* also has a substantial impact, contributing to a 16 percentage points change in the result. In addition, the *month* impacts a round 13 percentage points in the result. Finally, the *number of beds* has a impact of 11 percentage points.

In Figure 18b, the features are not only ordered by their effect on the forecasts, but it is also possible to see the impact of the highest and lowest values of the features on the model's result. It can be observed that lower values for the *population* and *number of beds (t-1)* are mostly associated with a negative impact on the forecasts, while higher values for these features contribute positively to the forecasts. Also the *lockdown* variable shows the impact of lockdowns on forecasts. When the variable is set to 1, indicating periods of lockdown, it consistently exerts a negative influence on the forecast results, as expected.



(a) Bar chart of mean importance.

(b) Density scatter plot of SHAP values.

Figure 18: Tree SHAP implementation integrated into XGBoost to explain the entire dataset. The summer season, the population, the number of beds in previous months, the month and the number of beds are the most important features in predicting the bed occupancy rate.

From the global explanations provided by Figures 18a and 18b, the analysis of local explanations becomes fundamental for this study. This localised information offers a detailed understanding of how specific features, such as the *number of beds* and *month*, influence forecasts.

The values provided by Figure 19 refer to the August 2022 observation for hotel establishments in Lisboa, which has a bed occupancy rate of 73.34%. It is relevant to note that, in all the available data, there is not a single observation with a bed occupancy rate above 80%, which is related to the fact that Lisbon is the municipality with highest demand, but yet has enough supply to cover this demand. When interpreting the values in the figure, it is essential to realise how each variable contributes to the model's prediction in relation to the expected average value. As all the values present are considered "higher values", it suggests that, for this observation, these features increased the prediction. Analysing each feature individually, it can be seen that:

- ***Crime rate* (4.93‰):** the SHAP value for Crime Rate suggests that, compared to the expected average, the higher crime rate in August 2022 contributes positively to the predicted bed occupancy rate for hotel establishments in Lisboa. For this specific observation, a higher crime rate is associated with higher predicted bed occupancy.
- ***Month* (8- August):** the SHAP value for the month of August suggests that, compared to the expected average, the fact that the observation corresponds to August increases the bed occupancy rate prediction. This is in line with the typical trend in which demand can be higher during the summer months.

- ***Population* (546923):** for this observation, the higher population compared to the expected average, in Lisboa in August 2022 contributes positively to the predicted bed occupancy rate.

- ***Season Summer* (1):** the SHAP value for the summer season suggests that, compared to the predicted average, the fact that the observation takes place in the summer season contributes positively to the predicted bed occupancy rate. This is also consistent with the trend of increased demand for tourism and accommodation during the summer.

- ***Number of beds (t-1)* (50928):** the SHAP value for the number of beds in the previous month indicates that, compared to the expected average, having a higher number of beds in the previous time month contributes positively to the predicted bed occupancy rate. This implies that historical data on bed availability influences the forecast.

Analysing this specific observation suggests that all of the features with higher than average values increase the model's prediction of bed occupancy rate.

Figure 19: Force plot illustrating the impact of each feature on the prediction for hotel establishments in Lisboa during August 2022. The impact of each feature suggests that all the features contribute positively to the model´s prediction.

Secondly, the only municipality that achieved a 100% occupancy rate of available beds in 2022, Lousada, was studied in the months of October and December. The analysis will focus on observing the month of December 2022 in Lousada's hotel establishments. The values provided by Figure 20, are not all "higher values", which means that there is a positive and negative impact of the variables on the model's prediction:

- ***Crime Rate* (1.62‰):** a higher crime rate has increased the prediction of the bed occupancy rate. This can be unintuitive, and it is essential to consider this in the context of the learnt patterns of the model.

- ***Number of establishments* (1):** the SHAP value indicates that the existence of one establishment in Lousada is associated with a higher predicted bed occupancy rate for that particular observation.

- ***Population* (47760):** A higher population has contributed positively to the predicted bed occupancy rate. Larger populations might indicate more potential demand.

- ***Year* (2022):** the year 2022 being a higher value has contributed positively to the predicted bed occupancy rate. This may suggest that the model has learnt specific trends for the year 2022.

- ***Season Summer* (0):** not being in the summer season contributed negatively to the expected bed occupancy rate. Summer seasons are often associated with greater tourist activity, so the absence of this season has a negative impact.

- ***Month* (12- December):** the month being December has decreased the prediction. December might be a period with lower tourism activity compared to other months.

The values for these variables suggest that having a single establishment, a higher crime rate, a larger population and the specific year 2022 are associated with a higher predicted bed occupancy rate. On the other hand, the absence of summer and the month of December are associated with a lower predicted bed occupancy rate.

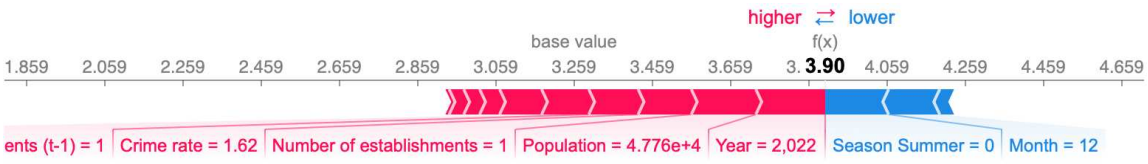


Figure 20: Force plot illustrating the impact of each feature on the prediction for hotel establishments in Lousada during December 2022. The absence of summer and the month of December are associated with a lower predicted bed occupancy rate.

Contrary to the previous cases, the Figure 21 shows a case where the bed occupancy rate is lower: Vimioso is the municipality that had the lowest mean bed occupancy rate in 2022. The case observed is for rural tourism in August, which was the month with the highest bed occupancy rate, 28.57%, in that year. When analysing each feature individually, the following can be seen:

- ***Number of holidays* (1):** having one holiday suggests that there might be increased tourism activity during that period, potentially leading to a higher bed occupancy rate.

- ***Month* (8- August):** August is typically a peak month for tourism, contributing to a higher bed occupancy rate.

- ***Season Summer* (1):** Summer is in line with August, reinforcing the expectation of increased tourism and higher bed occupancy.

- ***Population* (4143):** a lower population, lower than the expected average, may imply that Vimioso is a smaller municipality with fewer residents and the impact on tourism may

change, which in this case has decreased the prediction. However, smaller populations can lead to fewer tourists.

- ***Number of beds (t-1)* (79):** the number of beds in the previous month was lower than the expected average, which suggests a decrease in bed capacity, potentially influencing the current bed occupancy rate.
- ***Number of beds* (83):** a lower number of beds in the current month still indicates limited accommodation capacity.

These identified values such as the *month*, *season summer* and the presence of a holiday indicate a situation where contribute positively to the model´s prediction. However, the *population*, the *number of beds* and the *number of beds (t-1)* may contribute to a lower bed occupancy rate.

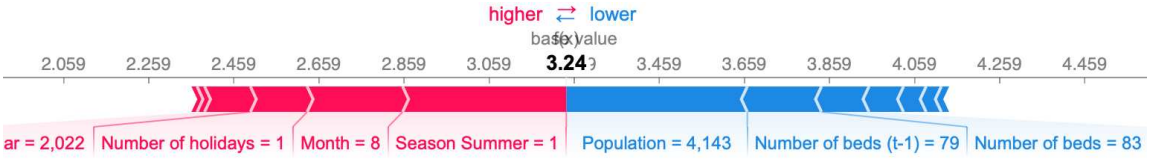


Figure 21: Force plot illustrating the impact of each feature on the prediction for rural tourism in Vimioso during August 2022. The impact of each feature suggests that the *population*, *number of beds (t-1)* and *number of beds* are below the expected average or have a negative impact on the model's prediction for the specific observation.

5 Discussion

An analysis of tourism demand was carried out, which made it possible to assess the trends in the demand for overnight stays in Portugal. There was a regularity in overnight stays over the years studied, which was only interrupted by the unexpected event of COVID-19 in 2020 and 2021. In 2022, with the end of the pandemic, the pre-covid values returned, indicating a regular trend in demand for overnight stays in the future (Figure 2). This analysis demonstrated that there is significant seasonality, with the summer months being the most attractive for tourists (Figure 3), who tend to prefer the Algarve and Lisbon regions (Figure 4) with hotel establishments as their first choice of accommodation facility, followed by local accommodations (Figure 7). This study of the dynamics of demand provides valuable insights into the trends shaping Portugal's tourism industry.

This thesis presents a methodological framework for analysing tourism demand. This framework focused on forecasting tourism demand by applying a machine learning model designed to extract valuable knowledge through the application of explainability strategies. The

study involves the implementation of an XGBoost model and the development of SHAP to interpret the model's decision-making process. Using these methods, the key features that most significantly influence tourism demand were investigated.

The model's performance metrics, as shown in Table 7, such as RMSE, MAE, MAPE and R^2, suggest that the model with *overnight stays* as a target is robust, as it outperforms these metrics, except MAPE, compared to the benchmark model, making it a more reliable and accurate tool for this study objective compared to the benchmark model. The effectiveness of this model shows a certain pattern regarding performance, where it demonstrates superior performance for smaller values of the target variable and encounters some challenges when predicting higher values. This phenomenon is expected given the wide range of values for the target variable, *overnight stays*.

These disparities emphasise the complex nature of forecasting in a dataset characterised by a significant amount of variability. The challenges in forecasting higher values are associated, in particular, with municipalities characterised by extensive overnight stays, a pattern evident in classifications such as hotel establishments and local accommodations. These accommodation facilities, which contribute to a considerable number of overnight stays, introduce a complexity layer that requires careful consideration in the modelling process. In line with this complexity, Hilaly and El-Shishiny (18) reach the same conclusion drawn by numerous studies, emphasising that no single forecasting model can generate the most accurate forecasts in all situations. The complexities of the dataset, especially regarding to municipalities with a high number of overnight stays, align with this perspective, highlighting the need for adapted modelling approaches to suit diverse and complex scenarios. In retrospect, a potential solution was implemented to address this challenge, providing a more adapted modelling approach. This solution was to use the *bed occupancy rate* as the target variable, since it offers the advantage of normalization and is limited to 0 to 100. This solution proved to be more accurate in terms of performance metrics compared to the other models performed, although it showed some limitations when it came to predicting bed occupancy rates above 90%, obtaining a higher prediction error. Another possible strategy could also be to consider training specific models for the most populous municipalities or those with the highest demand, considering their unique characteristics and dynamics, to improve forecast accuracy.

Despite the merits of AI models, they face challenges. They often require extensive parameter tuning and vast datasets on past tourist behaviour and other factors to train the model, which makes them susceptible to overfitting and slower to develop and implement. The results can be difficult to interpret due to the lack of a functional form, and although they can be highly accurate, they present challenges in transparency. To overcome these limitations that can hinder the acceptance of research results in the tourism sector, the SHAP technique was used. This

method provides more transparent and understandable information, improving the overall reliability of tourism demand forecasts, aiding decision-making and allowing the magnitude and importance of predictor variables to be extracted.

The inclusion of suitable independent variables when forecasting tourist demand is a major challenge due to the dynamic nature of tourists' tastes and attitudes (25). The study reveals that the topics related to tourism supply, such as the *number of beds* and the *number of beds (t-1)*, in the previous months, are one of the main predictors of *bed occupancy rate* (Figure 18). Demographic factors, such as *population*, show their relevance in shaping tourism demand, which is in line with what was mentioned in section 2. Temporal attributes such as the *season summer*, the *month* and the *year*, are also key features that influence the tourism demand. Furthermore, the study also suggests a positive impact of security measured through the *crime rates*, which is in line with empirical verification by tourism scholars (Cho (6); Gao and Su (14); Goh (16); Ribaudo and Figini (37); Richards (38)).

In contrast to the prevailing literature, the findings challenge the significance of some factors in tourism demand forecasting. Economic indicators such as *CPI* and *GDP*, often considered important in forecasting, are viewed as not crucial in this study.

In order to provide valuable insights for the municipalities' respective accommodation facilities when making strategic decisions in the tourism sector, a detailed analysis of local explanations was carried out. By focusing on local explanations, decision-makers obtain actionable information tailored to their specific contexts, allowing them to make informed choices that align with the observed impacts highlighted in the broader analysis. This differentiated perspective improves the decision-making process of municipalities and their accommodation facilities, promoting more effective strategies.

As an example, hotel establishments in Lisbon were examined during the month of August 2022 to obtain valuable information on the factors that determine the model's forecasts for that specific month and location (Figure 19). The positive SHAP value for the number of beds in the previous month suggests that historical bed availability plays a crucial role, where more beds in the previous month contribute positively to the predicted bed occupancy rate. Additionally, the positive SHAP values for the month of August and for the summer season indicate that during the summer months there is an increase in demand for hotel establishments in Lisbon, which is in line with the general trend of higher demand during the summer months. A higher population in Lisbon positively influences the expected bed occupancy rate, suggesting an increase in demand in densely populated areas. At odds with expectations, a higher crime rate contributes positively to the predicted bed occupancy rate, suggesting that crime is not a significant factor for tourists. It is worth noting that municipalities with higher crime rates also

have larger populations, which potentially leads to greater demand, since tourists prefer more populous municipalities due to their greater popularity.

Another example is for hotel establishments in Lousada during the month of December 2022 (Figure 20), where this municipality is the only one in 2022 to have achieved a bed occupancy rate of 100%. Despite December being conventionally considered a low season for tourism, the bed occupancy rate suggests a unique and perhaps untapped market scenario. The influence of seasonal patterns is evident in the lower values for Summer and December, indicating an off-peak season. The higher crime rate in Lousada did not act as a dissuading factor for tourists during this period, potentially challenging the frequent association between high crime rates and lower tourism. This may be due to what was mentioned above in the other example. The higher population and the presence of only one hotel establishment may contribute to higher demand, especially if Lousada is an attractive destination or hosts specific events. This situation of reduced supply in terms of the number of establishments may represent an opportunity for potential new hotel establishments, since the high population, together with the limited existing options, suggests an unexplored market demand. Furthermore, the higher value for the year 2022 emphasises the importance of considering annual trends, revealing that Lousada has maintained high occupancy rates even in a challenging year for the tourism industry. This tourism scenario in December 2022, marked by an occupancy rate of 100% during the low season and influenced by local factors such as the crime rate and population, highlights the need for local stakeholders and policymakers to understand these dynamics in order to optimise tourism strategies.

A final example is for rural tourism in Vimioso during August 2022, the municipality with the lowest mean bed occupancy rate in 2022 (Figure 21). This analysis shows that the observation is in line with the peak tourism season. The smaller population suggests that Vimioso may be a smaller municipality with fewer residents. Tourism can be influenced by factors such as natural attractions, cultural events or specific tourist activities. The lower values for the number of beds (current and previous month) indicate a potential constraint on accommodation capacity. A limited number of beds may be contributing to the low bed occupancy rate. The presence of a national holiday may contribute to an increase in tourism during this period. However, other factors such as the attractiveness of the municipality and promotional activities also play an important role, for example, in Vimioso in August there are local festivals with well-known national singers. The results suggest an interaction between seasonal patterns, local characteristics and accommodation capacity to influence the bed occupancy rate.

These detailed analyses of the different accommodation facilities by municipality over a given period can be valuable for understanding the dynamics that influence demand and provide actionable information, guiding potential improvements in operational efficiency and strategic

decision-making.

The application of the XGBoost model to tourism demand forecasting demonstrates the dynamism and complexity of this field. Its effectiveness lies in its ability to deal with complex and nonlinear relationships, modelling demand for new products or relatively new destinations with limited historical time-series data.

This study has practical implications for policymakers and management teams in the marketing and tourism sectors. Using the XGBoost model and implementing SHAP can help managers and policymakers capture the crucial factors that can influence their outcomes. By extracting information from the model's forecasts, they can make informed decisions.

6 Conclusion

The motivation for this thesis came from a critical analysis of the evolution of tourism demand forecasting methodologies over the years. Despite the variety of successful methods used to study this phenomenon, there is a notable gap, as mentioned in section 2, which concerns the lack of studies that integrate explainability into their analysis. This scarcity has not only challenged the reliability of existing research, but has also highlighted the need for a more transparent and interpretable approach. Consequently, this thesis aimed to fill this gap by introducing and applying an explainability framework to tourism demand forecasting. Since the explainability of the model ensures that decision-makers can make decisions with confidence, knowing the underlying factors that influence the results.

The results obtained from the study provide valuable information on the dynamics of Portugal's tourism industry in each municipality, which is an asset for policymakers and management teams in the marketing and tourism sectors. By using the XGBoost model and the SHAP technique, decision-makers have access to a comprehensive understanding of the main factors influencing tourism demand. And this transparency of the model not only increases the reliability of forecasts, but also facilitates more informed and strategic decision-making.

By focusing only on the policymakers, they can use the knowledge gained from the model's predictions to develop policies and initiatives that align with the specific dynamics of each municipality. For example, understanding the impact of seasonal patterns, demographic factors and local attributes allows for the implementation of strategies to optimise tourism promotion during peak periods or to address challenges during slower seasons. Meanwhile, management teams in the tourism industry can use the forecasting framework to optimise resource allocation, marketing campaigns and operational strategies. Knowing the specific factors that

drive demand in each municipality allows for more precise targeting of promotional efforts and the development of personalized services to meet the preferences of the various tourists. Ultimately, this increases the overall competitiveness of accommodation facilities and tourist services.

In conclusion, this thesis contributes to the tourism sector, not only by providing a new forecasting framework regarding its use in tourism forecasting, but also by emphasizing the importance of explainability in enhancing the robustness and applicability of tourism demand studies. The results obtained have practical implications that extend beyond theoretical knowledge, providing actionable information to policymakers and management teams so that they can define effective strategies, allocate resources wisely and contribute to the sustainable growth of the tourism sector.

6.1 Limitations and Future Work

It is essential to note that this thesis has limitations, since the testing of this forecasting framework is limited to Portugal and its municipalities. Therefore, the findings regarding the factors that influence demand may not be generalisable to other destinations.

The tourism demand forecasting is a dominant field of research, presenting opportunities for continuous refinement and improvement. This study can be improved and suggests potential directions for future exploration. Namely, the integration of external factors, such as events, transport costs and exchange rates, to enable a more comprehensive understanding of their impacts on tourism demand.

Incorporating data on events for each municipality throughout the years proves to be advantageous in forecasting tourism demand, given the impact of major events on visitor patterns and the seasonal dynamics of tourism. Major events often lead to an increase in demand for accommodations facilities, requiring measures in terms of room availability, pricing strategies and overall capacity management. Unfortunately, due to data granularity limitations and the unavailability of information on events in previous years through VisitPortugal, the inclusion of this variable was not feasible for this thesis.

By adding exchange rates to the dataset, the relative price of a destination can be determined. According to traditional economic theory, an increase in price can correspond to a decrease in demand and empirical studies have also confirmed this significant relationship between exchange rate and tourist arrivals (Wu et al. (47); Pai and Hong (35); Lin and Lee (27); Li et al. (25)).

Lin and Lee (27) emphasise the importance of average hotel prices as a key factor in their study, highlighting the insights that can be gained by incorporating hotel price dynamics into forecasting models. This strategic inclusion allows for a more detailed understanding of consumer behaviour in the context of travel decisions. Another topic considered important in forecasting tourism demand is the weather, which is considered to be one of the most important factors in deciding on a travel destination (Becken (3); Martín (29)).

Recognising the importance of these additional characteristics underlines the potential for future research to take advantage of this knowledge for a more accurate and comprehensive forecast of tourism demand.

References

[1] (2022). *OECD Tourism Trends and Policies 2022*. OECD.

[2] (2023). What trends offer opportunities or pose threats on the european outbound tourism market?

[3] Becken, S. (2013). Measuring the effect of weather on tourism: A destination- and activity-based analysis. *Journal of Travel Research*, 52:156–167.

[4] Bi, J. W., Han, T. Y., and Li, H. (2022). International tourism demand forecasting with machine learning models: The power of the number of lagged inputs. *Tourism Economics*, 28:621–645.

[5] Bontempi, G., Taieb, S. B., and Borgne, Y.-A. L. (2013). Machine learning strategies for time series forecasting.

[6] Cho, V. (2010). A study of the non-economic determinants in tourism demand. *International Journal of Tourism Research*, 12(4):307–320.

[7] Colladon, A. F., Guardabascio, B., and Innarella, R. (2019). Using social network and semantic analysis to analyze online travel forums and forecast tourism demand. *Decision Support Systems*, 123.

[8] Czerwinska, U. (2022). *Interpretability of Machine Learning Models: How Can One Explain Machine Learning Models?*, volume Part F1051, pages 275–303. Springer Nature.

[9] Division., U. N. S. and Organization., W. T. (2010). *International recommendations for tourism statistics 2008*. United Nations.

[10] do turismo de portugal, C. (2022). Efeitos turísticos da guerra da rússia na ucrânia.

[11] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

[12] Egger, R. and Gretzel, U. (2022). Tourism on the verge series editors.

[13] Fernandes, P. and Teixeira, J. (2008). Prediction tourism demand using artificial neural networks. *International Conference in European Union's History*.

[14] Gao, Y. and Su, W. (2019). Is the world heritage just a title for tourism? *Annals of Tourism Research*, 78:102748.

[15] Ghalehkhondabi, I., Ardjmand, E., Young, W. A., and Weckman, G. R. (2019). A review of demand forecasting models and methodological developments within tourism and passenger transportation industry. *Journal of Tourism Futures*, 5(1):75–93.

[16] Goh, C. (2012). Exploring impact of climate on tourism demand. *Annals of tourism research*, 39(4):1859–1883.

[17] Gooijer, J. G. D. and Hyndman, R. J. (1982). 25 years of time series forecasting.

[18] Hilaly, H. and El-Shishiny, H. (2008). Recent advances in econometric modeling and forecasting techniques for tourism demand prediction.

[19] INE (2023). Estatísticas do turismo 2022.

[20] Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Vaughan, J. W. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. Association for Computing Machinery.

[21] Labib, R. and Savard, G. (2013). Railway demand forecasting in revenue management using neural networks'.

[22] Law, R., Li, G., Fong, D. K. C., and Han, X. (2019). Tourism demand forecasting: A deep learning approach. *Annals of Tourism Research*, 75:410–423.

[23] Li, H., Hu, M., and Li, G. (2020). Forecasting tourism demand with multisource big data. *Annals of Tourism Research*, 83.

[24] Li, X., Pan, B., Law, R., and Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism Management*, 59:57–66.

[25] Li, Y., Lin, Z., and Xiao, S. (2022). Using social media big data for tourist demand forecasting: A new machine learning analytical approach. *Journal of Digital Economy*, 1:32–43.

[26] Lin, C. J., Chen, H. F., and Lee, T. S. (2011). Forecasting tourism demand using time series, artificial neural networks and multivariate adaptive regression splines:evidence from taiwan. *International Journal of Business Administration*, 2. GOOD FOR DISCUSSION, addition of variables.

[27] Lin, C.-J. and Lee, T.-S. (2013). Tourism demand forecasting: Econometric model based on multivariate adaptive regression splines, artificial neural network and support vector regression.

[28] Lundberg, S. M., Allen, P. G., and Lee, S.-I. (2017). A unified approach to interpreting model predictions.

[29] Martín, M. B. G. (2005). Weather, climate and tourism a geographical perspective. *Annals of Tourism Research*, 32(3):571–591.

[30] McKinsey Global Institute analysis (2018). Artificial intelligence (ai) has the potential to create value across sectors.

[31] Ministério da Administração Interna (2021). DossiÊ temÁtico: Covid-19: Compilação legislativa europeia, nacional e regional: de 2021-11-01 a 2021-11-30; jurisprudência.

[32] Molnar, C. (2019). Interpretable machine learning a guide for making black box models explainable.

[33] Oh, C. O. and Morzuch, B. J. (2005). Evaluating time-series models to forecast the demand for tourism in singapore: Comparing within-sample and postsample results. *Journal of Travel Research*, 43:404–413.

[34] Onder, I. and Wei, W. (2022). *Time Series Analysis: Forecasting Tourism Demand with Time Series Analysis*, volume Part F1051, pages 467–480. Springer Nature.

[35] Pai, P.-F. and Hong, W.-C. (2005). Research notes an improved neural network model in forecasting arrivals.

[36] Peng, B., Song, H., and Crouch, G. I. (2014). A meta-analysis of international tourism demand forecasting and implications for practice. *Tourism Management*, 45:181–193.

[37] Ribaudo, G. and Figini, P. (2017). The puzzle of tourism demand at destinations hosting unesco world heritage sites: An analysis of tourism flows for italy. *Journal of travel research*, 56(4):521–542.

[38] Richards, G. (2018). Cultural tourism: A review of recent research and trends. *Journal of Hospitality and Tourism Management*, 36:12–21.

[39] Semeida, A. M. (2014). Derivation of travel demand forecasting models for low population areas: the case of port said governorate, north east egypt.

[40] Sivarethinamohan, R. (2023). Exploring the transformation of digital tourism: Trends, impacts, and future prospects. pages 260–266. Institute of Electrical and Electronics Engineers (IEEE).

[41] Song, H. and Li, G. (2008). Tourism demand modelling and forecasting-a review of recent research. *Tourism Management*, 29:203–220.

[technofunc] technofunc. Challenges in the tourism industry.

[43] Tsai, T. H., Lee, C. K., and Wei, C. H. (2009). Neural network based temporal feature models for short-term railway passenger demand forecasting. *Expert Systems with Applications*, 36:3728–3736.

[44] Weihs, C. and Ickstadt, K. (2018). Data science: the impact of statistics. *International Journal of Data Science and Analytics*, 6:189–194.

[45] Witt, S. F., . S. H. (2002). *Forecasting tourism flows. In A. Lockwood S. Medlik (Eds.), Tourism and hospitality in the 21st century (pp. 106–118)*, volume Part F1051, pages 467–480. Elsevier Butterworth-Heinemann.

[46] Wu, B., Wang, L., Tao, R., and Zeng, Y. R. (2023). Interpretable tourism volume forecasting with multivariate time series under the impact of covid-19. *Neural Computing and Applications*, 35:5437–5463.

[47] Wu, Q., Law, R., and Xu, X. (2012). A sparse gaussian process regression model for tourism demand forecasting in hong kong. *Expert Systems with Applications*, 39:4769–4774.

[48] Yang, X., Pan, B., Evans, J. A., and Lv, B. (2015). Forecasting chinese tourist volume with search engine data. *Tourism Management*, 46:386–397.

[49] Yang, Y., Guo, J., and Sun, S. (2021). Tourism demand forecasting and tourists' search behavior: evidence from segmented baidu search volume. *Data Science and Management*, 4:1–9.

[50] Yang, Y. and Zhang, H. (2019). Spatial-temporal forecasting of tourism demand. *Annals of Tourism Research*, 75:106–119.

[51] Önder, I. (2017). Forecasting tourism demand with google trends: Accuracy comparison of countries versus cities. *International Journal of Tourism Research*, 19:648–660.

[52] Önder, I. and Gunter, U. (2016). Forecasting tourism demand with google trends for a major european city destination. *Tourism Analysis*, 21:203–220.

Appendix: Figures

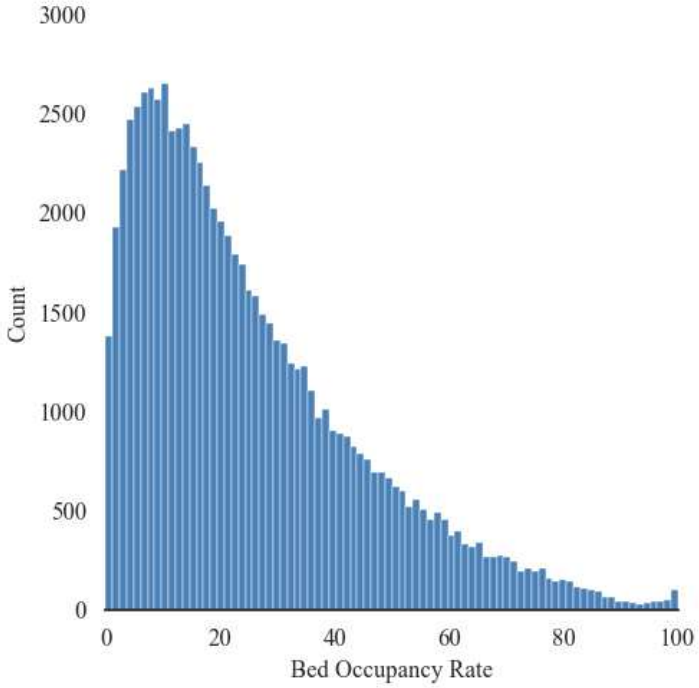


Figure 22: Distribution of Bed Occupancy Rate. It can be seen the left-skewed distribution of the variable.