LUND UNIVERSITY
School of Economics and Management

Master's Programme in Data Analytics and Business Economics

Modeling Airbnb Prices in the Maltese Islands

A Machine Learning Approach

by Gabriella Camilleri

DABN01
Master's Thesis (15 credits ECTS)
May 2023
Supervisor: Simon Reese

Abstract

The digital platform Airbnb has gained popularity in a number of countries particularly in the Maltese islands. Striking a balance in setting a price that is competitive and also renders a good profit can be a challenge. In this thesis a model is developed to predict the price of a listing in the Maltese islands for September 2022 through a machine learning approach whereby five types of models are considered. K Nearest Neighbours sets a baseline, while linear regression, a random forest, gradient boosted trees and neural networks are assessed in search of the model that is most generalisable beyond training data. Findings from this research conclude that gradient boosting specifically CatBoost model gives the best performance achieving an R^2 of 0.77.

Additionally the same models are re-fitted but incorporating additional walkable distance features to carefully identified points of interest namely historical sites, beaches, nightclubs, the capital city and bus stops. The results attained indicate that none of of the walkable distance features heavily contribute to explain any variance in the price of listings in the Maltese islands and only a slight improvement in model performance in some of the models considered is reported. Further to this, while retaining the additional distance features, training of the neural network is leveraged by pre-training the model on data that corresponds to another Mediterranean touristic island of Crete and a slight improvement is reported in model performance over the model solely trained on data for Malta from an R^2 of 0.66 to 0.67. This result opens a window for further research that seek to reap the benefits of transfer learning.

Keywords: Airbnb, Maltese Islands, Machine Learning, Geospatial Data, Transfer Learning

Acknowledgements

Resilience is a trait this research and a number of people throughout this experience have particularly taught me. I would like to express my appreciation for the continuous valuable guidance of my supervisor Simon Reese. I would like to also show my gratitude to my family and friends who have supported me throughout this learning experience. Finally I would like to thank the contributors of the free open map data source at OpenStreetMap(OSM) utilised in this research.

GOVERNMENT OF MALTA

FONDI.eu

The research work disclosed in this publication is partially funded by the Endeavour II Scholarship Scheme. The project may be co-funded by the ESF+ 2021-2027

May be co-funded by the European Union

ENDEAVOUR SCHOLARSHIPS SCHEME

Contents

List of Figures

List of Tables

1

Introduction

The tourism industry is a key contributor to the Maltese economy. In fact, prior the pandemic, in 2019 , it is claimed that the tourism industry contributed to 16.9 % of total jobs in the islands (OECD, 2022). Despite the pandemic, in 2022, tourism expenditure in Malta rebounded quickly (European_Commission, 2023) getting very close to pre-pandemic levels and amounting to a 91% recovery from €2.2Bn in tourism expenditure in 2019 to €2Bn in 2022 (NSOMalta, 2023) and yearly tourist arrivals reached 83% of pre-pandemic levels (Deloitte, 2023).

In recent years, the establishment of peer-to-peer accommodation digital platforms such as Airbnb have taken by a storm the accommodation sector in the tourism and hospitality industry in a number of countries. Platform such as Airbnb operate based on the so-called Sharing Economy business model which involves the peer-to-peer provision or sharing of goods or services through community-based online platforms (Hamari et al., 2016). Airbnb is a leading global online platform allowing individuals to advertise and rent a variety of property types ranging from an entire property unit (such as an apartment or a villa) to just a room within a property. This platform has been rapidly expanding across hundreds of different countries particularly in small European touristic destinations including the Maltese islands. In fact, from a study carried out in 2019 across 167 countries, the Maltese islands reported the second highest number of Airbnb listings per capita (Adamiak, 2019).

Considering both the popularity of Airbnb listings in the Maltese islands and the economic implications of the tourism industry to the Maltese Economy this thesis develops a pricing tool that predicts with reasonable accuracy the nightly price of an Airbnb unit in Malta. The nightly price of an accommodation is key both to the host profitability and the choices made by a potential tenant (Gibbs et al., 2017). On Airbnb, hosts are free to set the nightly price as they deem appropriate. Nevertheless, setting prices that strike a balance between rendering a good flow of revenue and being competitive requires domain expertise and extensive experience (Kanakaris and Karacapilidis, 2023). Airbnb currently aids hosts by offering a smart pricing tool whereby a host sets a minimum and a maximum price for its listing and a price within these bounds are suggested based on a dynamic adjustment based on an estimated fluctuating demand (such as due to seasonality, yearly events and others). Nevertheless, the host still needs to make an informed decision on the minimum and maximum price bounds and there is still room for improvement for such tool (Yang, 2021). Further to this, there are always prospective investors who are looking in making an investment in acquiring or renovating a property and are in search for some key features of a property that maximizes their potential profits (Yang, 2021).

The pricing tool developed in this thesis is the first up to the best of the author's knowledge explicitly targeting the Maltese islands that utilises a machine learning approach. Consequently, multiple machine learning models will be fitted and tested on performance to predict nightly price. Two main techniques will be considered to enhance the performance of the models fitted namely including geospatial distances to potential places of interest and public transportation as well as considering a transfer learning approach by leveraging a pre-trained model trained on Airbnb listings in Crete: A competing Mediterranean holiday island destination.

1.1 Research Problem

The main research problem in this thesis is to set up a model that predicts reliable prices of Airbnb listings in the Maltese islands. As a result the focus of this research revolves around three main questions:

1. How effective is a machine learning approach to predict the nightly price of Airbnb listings in the Maltese islands?
2. Does the walkable distance to places of interest and public transport have an effect on the price of listings?
3. Can the transfer of knowledge from data relating to a similar touristic island to Malta be beneficial to enhance model training and predictions of Airbnb prices in Malta?

1.2 Outline of the Thesis

In Chapter 2 we will present an overview of the existing literature revolving around our research. In Chapter 3 a review of the methodological concepts of the Geospatial extension undertaken in this research and the prediction algorithms implemented to tackle our research question will be provided. Then in following chapter, Chapter 4, the data utilised in this research will be described and data pre-processing steps will be outlined. This research results are then presented in Chapter 5 and discussed in light of existing literature. Lastly, in Chapter 6, concluding remarks of this research, limitations and potential future extensions are highlighted.

2

Literature Review

2.1 Price Determinants of Sharing Economy based Accommodations

Within the accommodation sector in the hospitality industry, price has been considered as crucial since early days (Gibbs et al., 2017). As claimed by Gutentag (2013), Airbnb has shaken up the traditional market for accommodation in the tourism industry up to a point leading conventional incumbent hotels in certain regions to drop their prices (Zervas et al., 2017). Although existing literature studying the price determinants of Airbnb listings in a variety of approaches in large urban cities around the world is abundant (Chen and Xie (2017), Gibbs et al. (2017), Teubner et al. (2017), Gunter and Önder (2018), Toader et al. (2022), Bernardi and Guidolin (2023)), existing literature specifically focusing on Airbnb price prediction models for Airbnb listings in well-known touristic leisure areas is rather scarce and even more scarce for Airbnb listings in touristic islands. Nevertheless, leisure tourism areas in particular in Southern Europe are among the most popular with Airbnb hosts as claimed by Adamiak (2018).

In reviewing the limited literature within specifically leisure touristic islands of researchers who in the past have attempted to better understand what drives the nightly price of an Airbnb accommodation, two main groups of Airbnb listings price determinants are identified namely: Property characteristics and Host characteristics. Lorde et al. (2019) who carried out a study among the islands of the Caribbean in North America and Suárez-Vega and Hernández (2020) who carried a similar study but on price determinants of listings in the Southern European island of Gran Canaria both confirm that property attributes such as the number of bedrooms, the number of bathrooms, air conditioning and pool facilities have a relevant impact on the price of listings. Additionally, Lorde et al. (2019) identify host attributes particularly having a Superhost status, host response time and review ratings as significant price determinants of Airbnb listings in the touristic leisure islands of the Caribbean while Suárez-Vega and Hernández (2020) spot the count of listings managed by a host to have a remarkable impact on price. Existing literature on Airbnb listings in the Maltese Islands (Ellul, 2019; Fearne, 2022), focus on identifying price-characteristic patterns of listings and identify the number of people a listing can accommodate, number of bedrooms, pool facilities and sea-view availability as main drivers of price. Further to this, Fearne (2022) identifies that prices of listings corresponding to an entire property, having bed and breakfast service and located in the Southern Harbour which includes the capital Valletta are on average higher.

2.2 Price Prediction Models

Such findings in previous literature just discussed are derived in a variety of approaches that evolved over time. Earliest research concerning the price of Airbnb listings concerned setting up models to specifically examine the effect of a variety of attributes on the price of Airbnb listings. A widely implemented approach for such studies undertook a hedonic regression approach (Chen and Xie, 2017; Deboosere et al., 2019; Gibbs et al., 2017; Liang et al., 2019; Lorde et al., 2019; Önder et al., 2019; Teubner et al., 2017). This also applies to the two existing studies on Maltese Airbnb listings (Ellul, 2019; Fearne, 2022). Other researchers started to look into other ways of modelling Airbnb price listings such as through a Machine Learning Approach. The shift from hedonic regression models to alternative approaches such as machine learning is stimulated by the greater ability of the latter approach to explain variability in the dependent variable (Sainaghi, 2021). Furthermore, hedonic regression methods have a weaker ability in handling outliers and feature engineering is human-based giving rise to potential bias in the model (Yazdani, 2021).

Liu (2021) studied and compared the performance of a number of price prediction models based on property and host attributes namely KNN, Multiple Linear Regression, Lasso regression, Ridge regression, random forest and gradient boosting methods to predict the price of Airbnb listings in Amsterdam and identified XGBoost as the model with best performance with an R^2 score of 0.6321. Yang (2021) considered property attributes and host attributes to develop a price prediction model for Airbnb listings in Beijing through a machine learning approach by considering XGBoost and Neural Network as potential price prediction models and also identified XGBoost as the best performing model with an R^2 score of 0.6549. Other researchers in addition to property and host attributes incorporate the use of textual data by extracting features from guest reviews when taking on a machine learning approach. For example, Rezazadeh et al. (2021) extract features through a sentiment analysis on guest reviews and consider linear regression, K-Means clustering, Support Vector Regression, Neural Networks and tree-based models to deduce a price prediction model for Airbnb listings in New York City and identified the Support Vector Regression model to achieve the best performance, reporting an R^2 score of 0.6901.

2.2.1 Enhancing models with Spatial Information

A number of researchers have set eye on the potential use of spatial information to explain house prices (Cellmer et al., 2019; Hill and Scholz, 2017; Mathur, 2020) in the sphere of real estate. This incorporation of spatial information has been extrapolated to the tourism industry (Chica-Olmo et al., 2020) whereby literature has shown that the location of an accommodation such as its proximity to points of interest, access to public transportation and proximity to the airport plays an important role in the stay experience of a guest (Yang et al., 2018). Researchers have incorporated geospatial data in their studies in explaining Airbnb price listing in a variety of ways. Some researchers undertake a spatial econometrics approach (Ellul, 2019; Gyódi and Nawaro, 2021) to account for spatial auto-correlation between listings based on their geographical proximity by considering a spatial lag in the hedonic price model constructed and confirmed that prices of Airbnb listings are spatially dependent on the prices of neighbouring listings. Others, (Chica-Olmo et al., 2020; Dudás et al., 2017; Gyódi and Nawaro, 2021), take on a different approach to incorporate geo-spatial data in predicting Airbnb prices. In the article by Dudás et al. (2017) the impact of proximity to the city centre on Airbnb listings in Budapest is studied through multi-band raster maps by incorporating a distance variable computed based on the Manhattan distance and it is concluded that price of listings is not correlated with its proximity to the city centre. Chica-Olmo et al. (2020) studied the impact of location related variables on Airbnb price listings in Málaga, Spain by incorporating distance variables from the Airbnb listing to the city centre, the beach and nearest place of interest and reported an improvement in variance in the listing price explained by the Ordinary Least Squares(OLS) regression model by incorporating distance variance over the OLS regression model that did not include distance variables. Further to this, including proximity to points of interest in terms of Euclidean distances is also very common in the literature (Ayouba et al., 2020; Crisci et al., 2022; Gyódi and Nawaro, 2021; Jiao and Bai, 2020; Yang, 2021). In particular, Gyódi and Nawaro (2021) have considered incorporating location variables based on Euclidean distances in an OLS model for each listing to a central point in the city centre and the nearest metro station and concluded that distance to the city centre has a significant impact on listing prices on six out of ten European urban cities the study was carried out in while distance to the nearest metro station had a significant impact on price in four cities. Due to the differences that exist between different cities, clearly there is no common consensus whether proximity to certain attractions and centrality of a listing influences the price.

Another approach to incorporate geospatial data is a network-based approach. Schwarzová (2020) considered the street graph network in the urban city of Prague, Czech Republic to compute the walkable distance from Airbnb listings to the nearest restaurant, park, supermarket and public transportation station. More recently, Sun et al. (2022) also considered a network based approach to determine factors that influence the spatial distribution of Airbnb listing in Suzhou, China. Both Schwarzová (2020) and Sun et al. (2022) justify the choice of this approach by claiming that walkable distance is more realistic then considering a direct distance based on some metric such as Euclidean distance. In this thesis a similar network-based approach is undertaken with some modifications by selecting amenities that are more sought after by a tourist visiting a holiday island rich in history and well sought after its beaches.

2.2.2 Enhancing models through Transfer Learning

Transfer learning is the process of having a machine learning algorithm that boosts its performance in a target task by learning from one or multiple similar application scenarios. Implementation of transfer learning has been experimented with by a variety of researchers in a variety of fields (Pan and Yang, 2010). Reviewing implementation of transfer learning in the tourism sector a common approach of transfer learning implementation is to utilise an open source established pre-trained model and fine tune it to your area of study. Such approach is widely implemented particularly in text classification tasks to analyse guest reviews (Ambolkar et al., 2022; Mousavi and Zhao, 2022) and image classification tasks to analyse listing images (Nguyen et al., 2018; Zhang et al., 2022). Apart from well established open source pre-trained models another approach is to perform pre-training on some available data from a domain that is similar to the domain under study. Existing literature undertaking this approach across differing markets to attain a final model to predict Airbnb prices in a market of interest is not found up to the author's knowledge at the time of writing. However a similar approach in predicting Electricity prices through a transfer leaning approach between different markets across different countries is identified in a number of studies. For example, Gunduz et al. (2023) pre-train a neural network on day-ahead electricity price data within one country to predict day-ahead electricity in another country across 4 different markets; Germany, France, Belgium and Nordics and report an improvement in model performance within all 4 markets over models trained and tested on a single market. Similarly, Van den Hurk (2021) undertakes a transfer learning approach based on Neural Networks to predict house prices between prices in an existing housing market and prices withing a newly constructed housing market which also reported an improvement in model performance through knowledge transfer. A similar approach will be employed in this study.

3

Methodology

In this section a theoretical overview of the methodological concepts employed in this thesis will be provided.

3.1 Geospatial Data

One of the primary ideas explored in this thesis is to investigate the potential benefits of integrating geospatial data to enhance the predictive ability of the Machine Learning models that will be implemented. Distances are calculated based on a GIS (Geographical Information System) Street Network rather than through a standard distance measure such as Euclidean distance. We therefore first provide an overview of this Network-Based approach to compute distances.

The Maltese islands are well-known for their beaches, nightclubs, and historical sites, according to the Trip Advisor website[1]. Taking this into account and general domain expertise, information of how far an Airbnb listing is located from the nearest beach, nearest historical site and nearest nightclub are considered. Furthermore distance to the nearest bus stop and to the Capital City are also considered as potential information relating to the accessibility and centrality respectively of a listing.

Distances to an Airbnb listing are computed based on a road network of the islands. For each Airbnb listing, we have available the longitude and latitude. Implementation of this road-network approach to calculate distances is based on the OSMnx library in Python (Boeing, 2017). Road network for any city in the World can be accessed via the *graph* module within OSMnx that is linked to an API of OpenStreetMap. Further information on the latter is provided in Chapter 4. Being represented by a graph, the street network consists of nodes and edges where nodes represent street junctions and edges (a collection of nodes linked together) represent streets. Furthermore, when retrieving the street network, the user can specify the type of network, namely drivable, walkable or bikable network. A walkable network is selected on the basis that the Maltese islands are rather small and walkable distance would be a more meaningful choice. Similarly for comparison, walkable distance is considered for the island of Crete. Being a walkable network the direction of edges (streets) is not integrated within the network. To better visualise a road

[1] https://www.tripadvisor.com/Tourism-g190311-Malta-Vacations.html accessed on 5 April 2023.

network, the road network for the Maltese islands is illustrated in Figure 3.1 where nodes are displayed in red and edges are displayed in white.

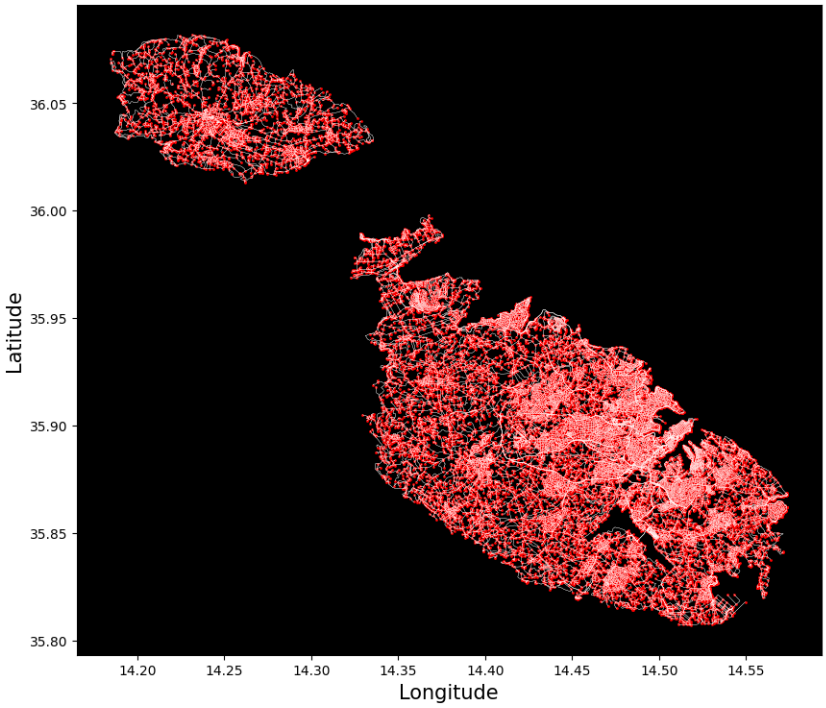
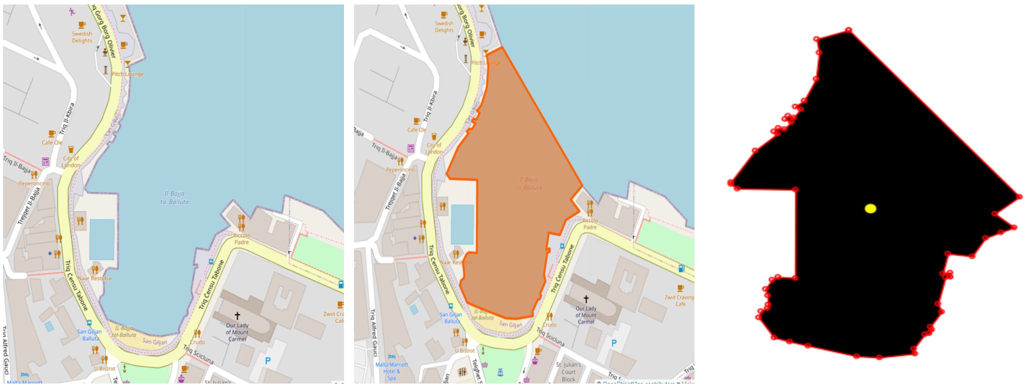


Figure 3.1: Street Network of the Maltese Islands

The walkable street network is combined with location data on points of interest also extracted from an API of OpenStreetMap. Further details on extraction of location data is provided in Chapter 4. A point of interest can be extracted in the form of a node or edge. An edge does not require to be linear but can also be in the form of an enclosed shape referred to as a polygon. A polygon is a collection of nodes that enclose an area for example a beach. When the location data on a point of interest is available in terms of a polygon the centroid of the polygon is computed as a representation location point of this amenity. To better visualise this concept of a polygon, on the left Figure 3.2 illustrate a bay in Malta ('il-Bajja tal-Balluta'). Location data is available as a polygon that covers the area of the bay as shown in the middle figure. The figure on the right displays the raw nodes that illustrate the boundary of the polygon in red and the computed centroid in yellow.

Given the location data of points of interest the shortest walkable route between any particular point of interest and any Airbnb listing can be computed. For that purpose, firstly a node on the street network that is closest to this centroid of a point of interest is found. The same applies when a point of interest is provided as a node. The function *nearest_nodes* from the *distance* module in OSMnx package is used for this purpose. Similarly, for each Airbnb listing the nearest nodes on the street network is also computed. Once we have each amenity of interest and each Airbnb listing represented by a node on the network then the shortest walkable distance between them is computed using the *shortest_path* function also found in the *distance* module. Shortest distance is computed based on the Dijkstra algorithm.



(a) Bay as illustrated in OpenStreetMap.

(b) Polygon representation of Bay as illustrated in OpenStreetMap.

(c) Raw Coordinates of Polygon and Centroid.

Figure 3.2: Polygon Data Representation

3.2 Prediction Algorithms

In order to determine the most effective model for a regression task aimed at predicting the nightly price of Airbnb listings in the Maltese Islands, a variety of machine learning models will be explored. The ”No Free Lunch Theorem” is a well known concept in the field of Machine Learning and highlights that there is no machine learning model that performs best across all problems and more than one model should be experimented with to identify the best performing model for the particular problem at hand (Kuhn and Johnson, 2013). The approach undertaken in this thesis is thus an experimental one whereby a variety of models are fitted, experimented with and compared on the basis of their performance. In total, 5 different models: A K-Nearest Neighbour model (KNN), a Linear Regression (3 variations), a Random Forest, a gradient boosted decision tree (2 variations) and a Feed-Forward Neural Network (2 variations) are considered. An overview of the methodology of each of these models will now be presented.

3.2.1 K-Nearest Neighbours Algorithm

KNN is the benchmark model that is selected in our research to which the performance of other algorithms can be compared. This model is selected as a benchmark model as it is a relatively simple model (Lindholm et al., 2022).

In order to accommodate both quantitative and qualitative input variables in our data, the Gower distance (Gower, 1971) is used to calculate the distance between the input vectors of two data points. That is, the distance between two listings (observations) a and b is given by:

$$d(x_a, x_b) = 1 - \frac{\mid x_a - x_b \mid}{max(x_i) - min(x_i)} \text{ for } i = 1, ..., n \tag{3.1}$$

for a numerical predictor X where n is the number of observations and

$$d(x_a, x_b) = \begin{cases} 0, & \text{if } x_a = x_b \\ 1, & \text{otherwise.} \end{cases} \quad (3.2)$$

for qualitative variables. The distance between two listings, a and b is then determined by computing an average of all distances for all numerical and categorical predictors between the two listings. Unlike traditional KNN models based on Euclidean distance that require input normalisation, the Gower distance formulation computes the normalised distance for numerical features and hence input normalisation is not required.

3.2.2 Linear Regression

Linear regression is also used despite its simplicity since it can be considered as a stepping stone to more advanced models (Lindholm et al., 2022). In order to address over fitting concerns, both l1 and l2 penalized versions (henceforth referred to as Lasso and Ridge regression) of linear regression are also considered. While when fitting standard linear regression model with no regularization input normalization is not performed as it is not required, when regularized models were implemented input normalisation is opted for to ensure that penalisation effect due to regularization is consistent across all input features irrelevant of their value range.

3.2.3 Random Forests

So far, individual basic models trained on the entire training data have been discussed. As a next step, we consider a model from the set of so-called ensemble models which take advantage of what Lindholm et al. (2022) refer to as the 'wisdom of the crowds' by combining the output of multiple individual basic models. A random forest regression is picked for the next step of our modelling phase particularly because of this bagging feature that declines the variance in predicted values (Kuhn and Johnson, 2013). A number of parameters can be specified when fitting a random forest. The parameters selected for tuning via random search followed by a grid search are detailed in Appendix B Table B.3. It is important to highlight that the default parameter value for $'max\_features'$ is equivalent to considering all features when in search for the best split which is equivalent to just bagging. Additionally, the *total_count*/3 was specifically selected in the random search for *'max_features'* since it stands for a third of the total number of features provided to the model based on the suggestion by (Liaw and Wiener, 2014). As a further means of regularization apart from considering specifying the maximum depth a base model tree can grow through *'max_depth'*, the parameter *min_samples_leaf* which specifies the minimum number of observations for a further split in the base models to be considered is also specified as a means to avoid over fitting .

3.2.4 Gradient Boosted Decision Trees

As reviewed in the Literature Review earlier, a number of researchers who experimented with different Machine Learning models to attain a prediction model for Airbnb listings prices in other cities around the World reported gradient boosting decision tree models to triumph among other machine learning models. Inspired by promising results from existing literature two variations of gradient boosting decision tree models will be considered and discussed. The general concepts underlying gradient boosting will be outlined based on the framework set by Lindholm et al. (2022) and Wallin (2022). Let us suppose we have a regression task with the set of inputs and output denoted by $\{\mathbf{x}_i, y_i\}_{i=1}^n$ and the goal is to find function $m(\boldsymbol{X})$ such that the cost function denoted by J and evaluated as:

$$J(m(\boldsymbol{X})) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, m(\boldsymbol{x}_i)) \tag{3.3}$$

is minimised. This is achieved such that at each iteration a new base model is constructed sequentially on the basis of previous base models such that at iteration say b the base model is given by:

$$m^{(b)}(\boldsymbol{X}) = \sum_{k=1}^{b} \alpha^{(k)}\tilde{m}^{(k)}(\boldsymbol{X}) \tag{3.4}$$

where $\tilde{m}^{(k)}(\boldsymbol{X})$ are basis functions and $\alpha^{(k)}$ is the weighting assigned to each respective base model. The key concept behind gradient boosting is to minimise the cost function in a way that imitates the gradient descent algorithm. This is attained such that at iteration say b the negative gradient of the cost function evaluated for predictions attained from the previous model, $b-1$ is computed and denoted as follows:

$$\boldsymbol{d}^b = -\nabla_c J(c)|_{c=m^{(b-1)}(\boldsymbol{X})}. \tag{3.5}$$

Elements resulting in $\boldsymbol{d}^b$ are then used to fit a regression where the inputs are the initial $\boldsymbol{x}_i$ and outputs are the corresponding elements in $\boldsymbol{d}^b$, d_i^b. The derived regression is $\tilde{m}^{(b)}$ and the new updated boosted model is now denoted by:

$$m^b = m^{b-1} + \gamma\alpha^b\tilde{m}^{(b)}. \tag{3.6}$$

In this way iteration b is completed. One of the most popular implementations of gradient boosting is XGBoost. This algorithm additionally supports regularized gradient boosting and parallelization and will be one of the employed candidate model in this thesis. As part of the parameter tuning exercise undertaken in fitting the XGBoost model several means of regularization are considered. These include considering a lasso and ridge regularization on the weights through parameters *reg_alpha* and *reg_lambda* respectively, a limit on the depth of each of the set up base models through parameter *max_depth*, random selection of observations prior training per base model through the parameter *subsample* and a random selection of features considered for training per tree that is set through the parameter *colsample_bytree*.

Further to this another variation of boosted decision trees will be considered namely, CatBoost. CatBoost was developed after the widely implemented gradient boosting algorithm XGBoost (Chen and Guestrin, 2016) with two major improvements that will be discussed in detail in the next section.

CatBoost

CatBoost stands for categorical boosting and as the name suggests it is best known for its support to categorical features in the data. In tree-based models a common way of handling categorical features is by one-hot encoding (Prokhorenkova et al., 2018) such as XGBoost and random forests. However, one-hot encoding is not ideal when having categorical features as this approach induces sparsity and potentially negatively impacts model performance. In fact, Hastie et al. (2009) claims categorical variables especially with a large number of levels should be avoided with tree-based models. Apart from sparsity another issue with categorical variables and tree based models is that each category in a categorical variable is considered independent from the other categories in the same categorical variable and when a tree-based model makes a choice of feature selection, dummy variables resulting from one hot encoded categorical variables are less likely to be selected than other numerical variables especially at the initial levels of the tree due to one-hot encoding. This will in fact be seen in our empirical results in Chapter 5. Moreover, Hastie et al. (2009) states that if factor variables with high number of categories are used as is, then this may result in splits that fit the training data excessively well and hence lead to overfitting.

CatBoost is a variation of gradient boosted decision trees that does not require categorical variables to be one-hot encoded. This algorithm handles categorical variables by a technique called Ordinal Target Encoding. It transforms each category into a numerical value in a way such that categories with similar target value distribution are assigned similar numerical values while categories with different target distributions are assigned more differing numerical values. This technique helps transforming categorical features into numerical variables with minimal information loss (Prokhorenkova et al., 2018). This is the first major advancement of CatBoost over traditional XGBoost.

To explain the concept of Ordinal Target Encoding employed by CatBoost on the basis of the theoratical framework set up by Prokhorenkova et al. (2018), let D be the original data set and similarly as before let the input and target variables be denoted by $\{\mathbf{x_i}, y_i\}$ for $i \in \{1, ..., n\}$ observations. Let $\ell(y_i, m^b(\mathbf{x}_i))$ denote the loss function where m^b is the function constructed iteratively at the b^{th} iteration in the gradient boosting algorithm that attempts to predict $\mathbf{y}$. At iteration $b+1$ we are after finding new function m^{b+1} such that $m^{b+1} = m^b + \tilde{m}^{b+1}(\mathbf{x})$ where $\tilde{m}^{b+1} = \underset{\tilde{m} \in M}{\arg\min} \mathbb{E}[\ell(y, m^{b+1})]$ where M is the set of potential decision trees and $\tilde{m}^{b+1}$ is the one that minimizes the expected loss. Firstly, observations in D are randomly permuted to attain permutation say σ. Suppose that the k^{th} element of σ is denoted by $\sigma(k)$ and let $D_k = \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{k-1}$ be the inputs for observations prior the k^{th} observation in σ. Suppose that the i^{th} variable $\mathbf{x}^i$ is a categorical variable. Then the categorical value of the variable i for observation k denoted by x_k^i is transformed into a numerical feature by CatBoost using the following equation:

$$\hat{x}_k^i = \frac{\sum_{\mathbf{x_j} \in D_k} \mathbb{1}_{\{x_j^i = x_k^i\}} y_j + ap}{\sum_{\mathbf{x_j} \in D_k} \mathbb{1}_{\{x_j^i = x_k^i\}} + a} \tag{3.7}$$

where p is the average value of the target variable (Micci-Barreca, 2001) and

$a > 0$ is a parameter set by the user. In this way computation of the target statistic $\hat{x}_k^i$ excludes y_k and overcomes the issue of so-called target leakage (Prokhorenkova et al., 2018) and consequently over fitting. On the other hand when it comes to determine if decision tree $\tilde{m}^{b+1}$ minimises the expected loss, dataset D is utilised.

An further modification in CatBoost is the implementation of the so called 'Ordered Boosting Technique' instead of the classical gradient boosting technique. In traditional gradient boosting when computing the leaf value in tree building this is dependent on all observations in this leaf leading to bias since these observations were initially used to train (build) the model so far on. To overcome this issue random permutation is again considered. Then models are created based on differing training data say for example Model M_j is trained on all observations that are listed before the j^{th} observation in the attained random permutation such that when it comes to compute the residual for observation j, this is computed based on a model M_j that was not trained using this observation.

Furthermore another differing feature of CatBoost from other boosting algorithms is that trees are built in a symmetric manner such that nodes at the same depth of the tree have the same set threshold for splitting. An example of a symmetric tree is displayed in Figure 3.3. This makes the algorithm computationally faster. Such trees are termed 'oblivious' trees in literature (Prokhorenkova et al., 2018).

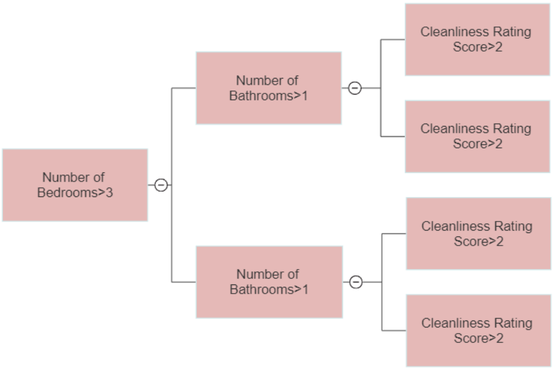

Number of Bedrooms>3
Number of Bathrooms>1
Number of Bathrooms>1
Cleanliness Rating Score>2
Cleanliness Rating Score>2
Cleanliness Rating Score>2
Cleanliness Rating Score>2

Figure 3.3: An oblivious tree

Additionally, CatBoost at tree-level employs a minimal-variance sampling technique on observations for which different probability of selection is assigned to different observations on the basis of assigning higher probabilities to observations that are more informative such that accuracy at split level is maximised. Further details on this technique can be found in Ibragimov and Gusev (2019).

3.2.5 Feed-Forward Neural Network (NN)

The final machine learning model used in this thesis is a fully-connected neural network (NN) model from the deep learning field. The architecture of a NN consists of a chain of linear regression models and activation functions that inject an element of non-linearity to the model. Inputs are normalised prior fitting the NN to avoid having some weights being updated at different rates since the learning speed is proportional to input magnitude. During the experimentation phase of the NN architecture, we began by considering a single hidden layer, then added a second and eventually a third. However, we found that one hidden layer was the most appropriate for our regression task by observing validation loss.

While experimenting with NNs it is observed that the model is overfitting to the training data and some regularization is incorporated in the model. Dropout is applied to the hidden layer as it regarded a powerful regularization technique that is computationally cheap (Goodfellow et al., 2016).

With regards to the activation function applied to the hidden layer there is no common consensus which works best and a researcher is encouraged to undertake a trial and error approach to identify the activation functions leading to best model performance (Goodfellow et al., 2016). Therefore two different activation functions are considered and experimented with. Furthermore, since we are handling a regression task the linear activation function is considered for the output layer. A linear activation function for the output is particularly ideal as it does not saturate while complementing well gradient optimization algorithms for learning (Goodfellow et al., 2016).

3.2.6 Neural Network Based Transfer Learning

The success of machine learning models has been accelerated with the era of 'Big Data' (Goodfellow et al., 2016). Deep learning models such as neural networks are well known to generalise better to unseen data when availed with a larger training data (Gunduz et al., 2023) and data dependence is a burden to deep learning models when compered to more simple traditional machine learning models (Tan et al., 2018) such as the ones previously discussed in this Section.

So far we have discussed machine learning models for which models are trained and tested on data that pertains to a single source: in our study data pertaining to the Maltese islands. Given that the data set for the Maltese Islands is relatively small a transfer learning approach is explored in attempt to better the performance of a neural network fitted and tested on data from a single source.

To better understand the concept of transfer learning the foundations of the theoretical framework behind transfer learning will be outlined on the basis of the framework outlined by Pan and Yang (2010). Transfer learning can be regarded as the transfer of information from one domain (scenario) which is referred to as the source domain to another domain which is referred to as the target domain. Let us denote by D_S, the source domain and D_T the target domain. Each domain is composed of two elements, input features, X and a marginal probability distribution $P(X)$. Each domain has a corresponding so-called task which consists of the output and and learned predictive function. The source task and the target task are denoted by T_S and T_T respectively while the learned predictive functions are denoted by f_S and f_T . Then formally transfer learning can be defined as the learning of f_T in D_T utilising information captured in D_S and T_S.

The most difficult question to answer in the sphere of transfer learning is which scenarios are capable of reaping the benefits of transfer learning without falling in the pitfall of so called negative transfer (Pan and Yang, 2010) which occurs when transfer learning has a negative impact on the performance of the model in the target domain. Selection of Crete data is made on the basis that Crete just like the Maltese islands is a well-known touristic island with common natural characteristics like climate and beaches (Vladimirova, 2011). Furthermore, as claimed by (Ellul, 2019) Crete is identified as a competing holiday destination to the Maltese islands. Further to this we performed some initial explanatory data analysis to identify similarities in the two domains found in Section 4.3.

In our case we decide to perform what is termed in literature (Pan and Yang, 2010) as parameter-based transfer where a model trained on the source domain is re-utilised for the target domain. An important claim must be made here that in this implementation an assumption that parameters pertaining to the learnined predictive function in the source domain, f_S are relevant for predictions in the target domain, D_T. Firstly, a feed forward NN is trained on the data for Crete which is identified as our source domain. The best architecture of the network is determined in a similar approach as outlined earlier when initially fitting a NN on data for Malta whereby 1 hidden layer is first considered, followed by 2 and later by 3 hidden layers. Based on the mean squared error performance on test data a model with two hidden layers is selected. Hyperparameter tuning is then performed to find the optimal parameters for the NN with 2 hidden layers. For transfer learning purposes, a NN can have some layers whose corresponding weights are frozen and others are not. The main struggle in this approach is to identify which part of the trained model captures knowledge that is of potential use to our target task (Yang et al., 2020). Having said that literature suggests that the earlier layers are are the most generalisable and potentially contain the most potential transferable knowledge (Tan et al., 2018). Based on these suggestions, the first layer in the trained NN on Crete data is accordingly 'frozen' to retain information learned during training on the data set for Crete. The term 'freezing' is used in the sphere of transfer learning to refer to the case when the weights in a layer are not trained during the learning phase of a NN (not updated during back propagation) but are kept fixed. Then we proceed to fine tune this NN having its first layer frozen on data for Malta which is identified as our target domain, D_T.

3.3 Metrics

In light of metrics evaluated in previous related literature and hence to facilitate comparison of our results the R^2 value is deemed as a suitable evaluation metric throughout our studies to assess model performance of models investigated.

4

Data

We will next discuss and provide a preliminary analysis of the data utilised.

4.1 Listings Data

Airbnb listings data for Malta is attained from *insideairbnb.com* which is an organisation that collects Airbnb data for multiple cities worldwide on a quarterly basis with the purpose of advocating the impact of peer-to-peer accommodations on residential communities. Data available for the Maltese islands is rather limited and corresponds to only data scraped at a one point in time in September, 2022. This limited data availability is a challenge and for this reason we consider utilising a data set of listings corresponding to the island of Crete in Greece to implement a transfer learning model as discussed in Chapter 2&3. Listings data for Crete is available at a quarterly basis from March 2022 till March 2023. Airbnb listings data for Crete is also acquired from *insideairbnb.com*. Listings data set for each of the islands describes physical features of each Airbnb listing, location details of the listing as well as features concerning the host of the listing. A complete list of variables in the listings data set post data cleaning and pre-processing can be found in the Appendix A in Table A.1. The total number of listings in Malta and Crete is 8,505 and 113,891 respectively. The data pre-process of the listings data is summarised in Figure 4.1 and explained in further detail below.

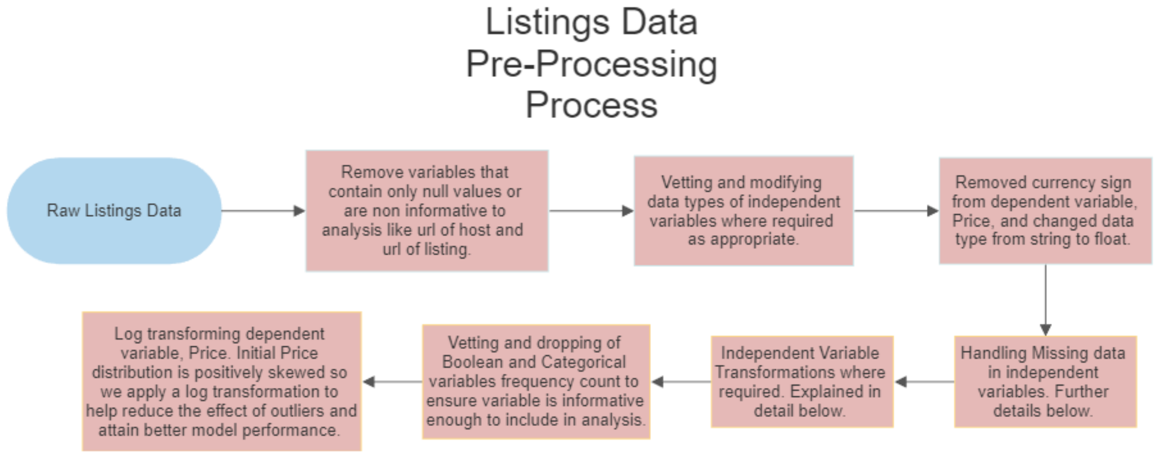

Listings Data Pre-Processing Process
Raw Listings Data
Remove variables that contain only null values or are non informative to analysis like url of host and url of listing.
Vetting and modifying data types of independent variables where required as appropriate.
Removed currency sign from dependent variable, Price, and changed data type from string to float.
Handling Missing data in independent variables. Further details below.
Independent Variable Transformations where required. Explained in detail below.
Vetting and dropping of Boolean and Categorical variables frequency count to ensure variable is informative enough to include in analysis.
Log transforming dependent variable, Price. Initial Price distribution is positively skewed so we apply a log transformation to help reduce the effect of outliers and attain better model performance.

Figure 4.1: Listings Data Pre-processing Workflow

4.1.1 Handling of Missing Data

Our initial listings data both for Malta and Crete contained missing data. Missing data frequency for listings data in Malta is shown in Table 4.1.

Variable Name	Data Type	Missing Data Count	Missing Data Percentage
Host_response_time	Categorical	866	10.18%
Host_response_rate	Numerical	866	10.18%
Host_acceptance_rate	Numerical	515	6.06%
bedrooms	Numerical	266	3.12%
beds	Numerical	165	1.94%
Number_of_baths	Numerical	24	0.28%
review_scores_rating	Numerical	1469	17.27%
review_scores_accuracy	Numerical	1492	17.27%
review_scores_cleanliness	Numerical	1492	17.27%
review_scores_checkin	Numerical	1492	17.27%
review_scores_communication	Numerical	1492	17.27%
review_scores_location	Numerical	1492	17.27%
review_scores_value	Numerical	1492	17.27%

Table 4.1: Missing data in Maltese Listings Data

When proportion of missing data is small applying traditional imputation methods such as mean imputation is a sensible choice (Bennett, 2001). However, Bennett (2001) continues to claim than when missing data percentage exceeds 10%, this might lead to biased analysis. Since some of our variables have more than 10% of missing data, we consider a tree-based model as an alternative to perform imputation of missing data as suggested by Hastie et al. (2009) specifically a random forest imputation based method. Waljee et al. (2013) have claimed that random forest imputation performed better than traditional imputation methods such as mean imputation, K nearest neighbour imputation and multivariate imputation by chained equations when data is missing completely at random [1], while Tang and Ishwaran (2017) further claimed that performance of random forest imputation is better when compared with other imputation methods even for data missing not at random[2]. From further analysis it is deduced that in our scenario missing data in variables relating to review scores corresponds to listings with no guest reviews in the last 12 months. There could be various reasons a listing is not yet reviewed perhaps the listing has a price which is too high and is not being booked in which case data is not missing completely at random among others. No knowledge on missing data of other variables is known. On the basis of this limited knowledge on missing data a random forest Imputation is deemed suitable.

Prior to performing this imputation an ordinal encoding is performed for the categorical variable *host_response_time*. This is suitable since this variable has a reasonable ordering in its categories. Then random forest imputation consists of an iterative process and works as follows. To start off the imputation algorithm, the

[1] Missing data is independent of observed data as well as unobserved missing data values.

[2] Missing data is correlated with the value of the missing values.

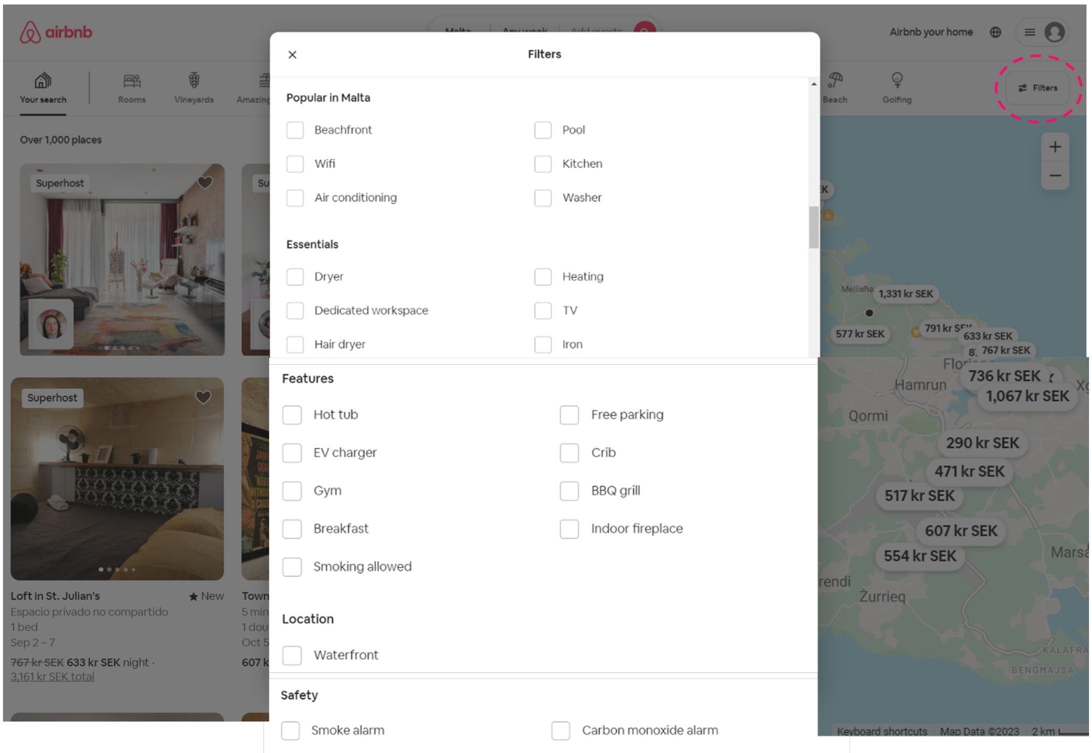
mean value of each variable with missing data is computed and imputed in place of the missing data. Now note, that since our data set has multiple variables with missing data, imputation is performed in ascending order of missingness starting with the variable that has the lowest percentage of missing data say variable X_1.Then a random forest is trained on all observations that do not have missing values in variable X_1 and the target variable of the random forest is the variable X_1. The trained random forest is then used to perform predictions of variable X_1 for observations with missing values in this variable. In other words, the training data consist of observations that do not have missing observations in X_1 while the test data are observations that have missing values in X_1. At this point there are no observations with missing values in variable X_1. Next we repeat the same procedure for the variable with the second least missing values. This is repeated for all variables with missing values. At this stage one iteration of the random forest imputation algorithm is complete. This whole procedure explained is repeated until some stopping criterion is met or a the number of iterations pre-specified is met. In the upcoming iterations the random forests for say variable X_1 are trained on original data and imputed data for missing values of other variables such that at each iteration the random forests are trained in attempt to beat the quality of the imputation for that variable in the previous iteration. Missing data in the Crete data set is also handled with this approach. Random forest imputation is implemented in Python using the *MissForest* package.

4.1.2 Feature Transformations

Physical Properties of Listings

In the raw data set information about bathrooms within listings is provided by a single text variable *'bathrooms_text'* where a host specifies the number of bathrooms and whether it is shared (eg.'1 shared bath', '1 bath','2 shared baths', '2 baths' etc...). Variable *'bathrooms_text'* is split into two new variables: one specifying whether bathroom/s available is/are shared or not, namely (*'bathroom_shared'*), and another variable indicating the count of bathroom/s available, namely $Number\_of\_Baths$.

The amenities of a listing are provided in the raw data set as a list in a single variable. Airbnb provides each host the opportunity to select amenities that it offers at its listing out of a standard list of amenities. On top of this, the host can add further custom free-text amenities. Hence the number of unique amenities specified is quite large (1,198 for Malta and 3,388 for Crete). At this point it is decided to create boolean variables for each amenity that the Airbnb filtering tool on the site allows potential guests to select that are specifically relevant for listings in Malta. These amenities on the Airbnb filtering tool are grouped into 5 categories namely: 'Popular in Malta', 'Essentials', 'Features', 'Location' and 'Safety' and a snapshot of the filtering tool for potential guests is displayed in Figure 4.2. Each of the amenities displayed in Figure 4.2 is represented by a boolean variable indicating whether a listing has this amenity or not. The only exception is that both safety features are grouped into one boolean variable indicating the availability of at least one of a smoke alarm or a carbon monoxide alarm. The reason this is done is since individual boolean would have been not so informative. Furthermore, only one listing allowed smoking within its premises and hence this boolean variable is deemed non



Filters

Popular in Malta

□ Beachfront □ Pool

□ Wifi □ Kitchen

□ Air conditioning □ Washer

Essentials

□ Dryer □ Heating

□ Dedicated workspace □ TV

□ Hair dryer □ Iron

Features

□ Hot tub □ Free parking

□ EV charger □ Crib

□ Gym □ BBQ grill

□ Breakfast □ Indoor fireplace

□ Smoking allowed

Location

□ Waterfront

Safety

□ Smoke alarm □ Carbon monoxide alarm

Figure 4.2: Amenties Filters for Guests when browsing on Airbnb for listings in Malta

informative and dropped from the data set. The amenities suggested in the filtering tool for listings in Malta coincides with that of Crete. Thus a similar approach is undertaken to represent amenity information for listings in Crete.

Additionally, in the raw data set a categorical variable *property_type* is provided with 65 categories however some categories in this variable are not so informative since for example the most popular category for Malta is *Entire rental unit* and consists of 3,610 listings as shown in Figure 4.3. Other categories such as *'Entire villa'* are considered as informative and retained as is. On the rest of the listings with unclear property type an exercise is performed in attempt to gain valuable information on the type of property a listing is found in by analysing two free text variables initially presented in the raw data set namely *name* and *description* whereby a host is free to provide a name and a description to the listing. To give an example a search is made among listings whose *name* or *description* contains the word *apartment* or *flat* and property type is then identified *apartment*. The final categories for *property_type* are apartment, maisonette, house, villa, hotel, guesthouse bed and breakfast and other as shown in Figure 4.4. A similar modification is applied for data on listings in Crete. From Figure 4.4 it is observed that apartments are the most popular type of properties found advertised on Airbnb for listings in Malta. Same claim is made by Ellul (2019) for listings data in Malta for May, 2019. A similar exercise is performed for the variable *property_type* for data on listings in Crete. The corresponding plots of counts prior and post modification of the variable can be found in Appendix A Figures A.1 and A.2.

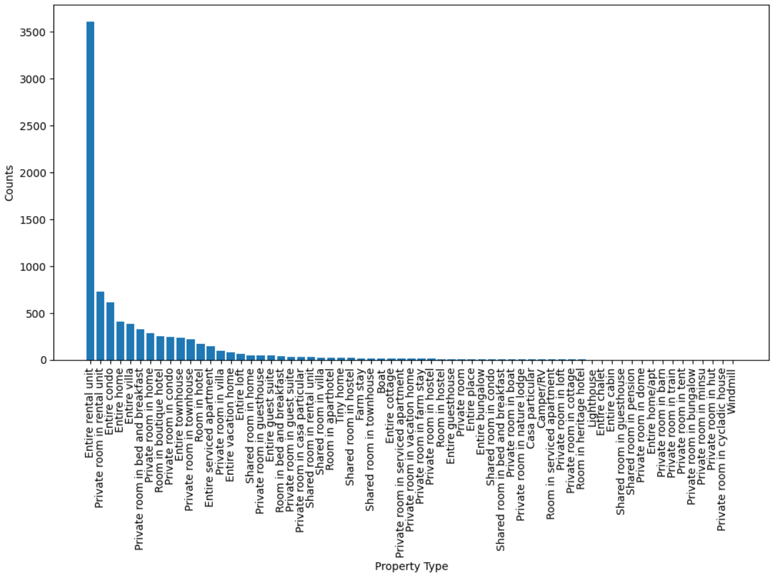


Figure 4.3: Count of listings by property type in raw data set

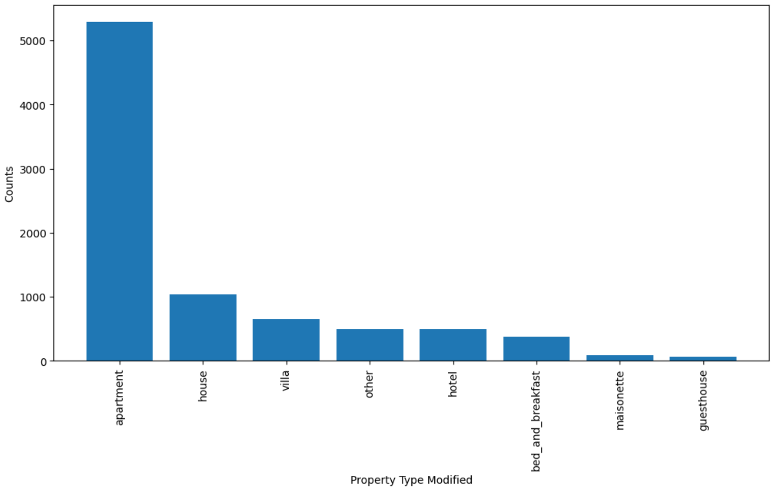


Figure 4.4: Count of listings by property type in modified data set

Host Features

In the raw data set we are provided with information on the specific day a host signed up as a host with Airbnb. This information is transformed into a new variable namely *host_no_days_active* which is the number of days between the day the host signed up with Airbnb and the day the data is scraped.

Additionally in the raw data set the modes utilised by hosts for verification purposes are provided as a list of modes for each listing in terms of sing variable. The possible modes of verification are *email*, *work_email* and *phone* or a combination of these. This information is transformed into two boolean variable *email* indicating whether host has provided an email (work or personal) for verification and another boolean variable *phone* indicating whether phone is provided. The latter variable is however dropped since only 8 listings did not provide a phone for verification and hence variable is not very informative.

Location Feature

Information concerning the locality in which a listing is located is provided as a categorical variable with initially 68 localities. Some localities are less popular than others as shown in Figure 4.5. A choice is made to group localities with less than 150 listings by region of where they are located in the Maltese islands indicating the location of the variable at region level instead. The count of listings for the resulting variable which is named *region_name* is displayed in Figure 4.6. From Figure 4.6 it is observed that the most popular location for Airbnb listings is Sliema. Ellul (2019) also claims that Sliema is the most popular locality for Airbnb listings when analysing data for listings scraped in May, 2019. A similar exercise is not required for data of listings in Crete since the variable capturing region information is non-informative for transfer learning purposes.

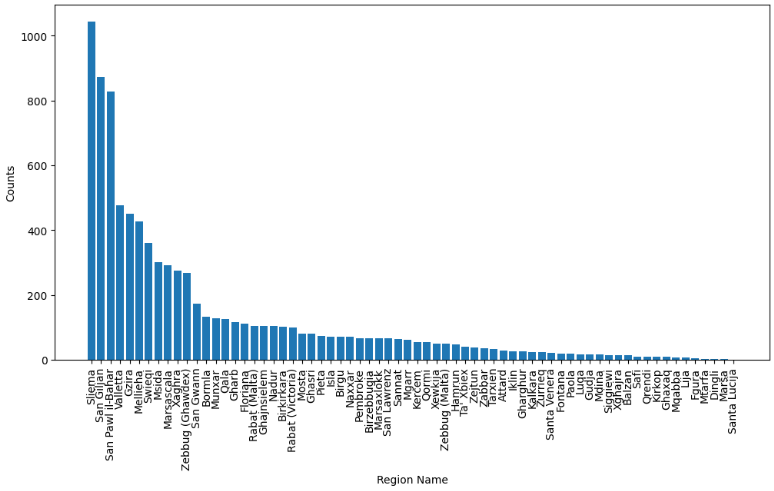


Figure 4.5: Count of listings by region in raw data set

One-hot encoding

Most machine learning algorithms do not accept categorical variables as input features. Therefore, unless otherwise specified one-hot encoding is performed on cat-

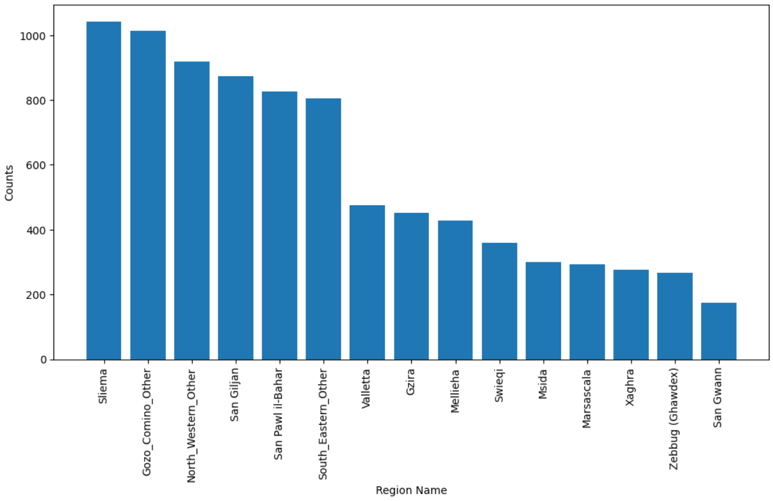


Figure 4.6: Count of listings by region in modified data set

egorical variables *property_type*, *room_type*, *region_name* and *host_response_time* as part of the pre processing of the data.

Target Variable Transformation

In Figure 4.7 we display the initial distribution of the target variable nightly price. It is clearly observed that the distribution is positively skewed and hence log transformation is applied to help diminish the effect on performance of outlying values.

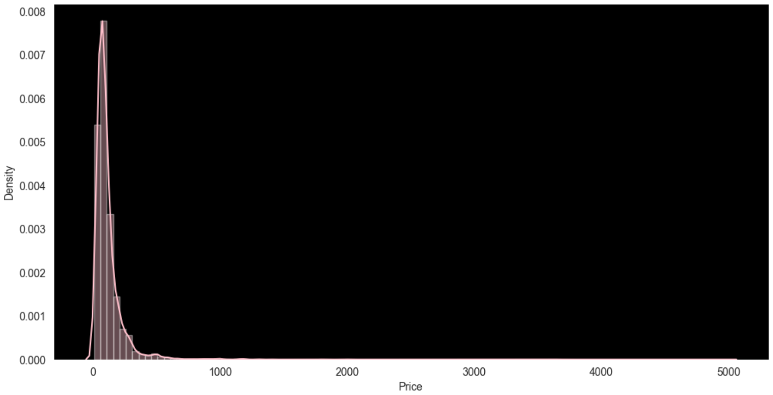


Figure 4.7: Listings Price Distribution in the Maltese Islands

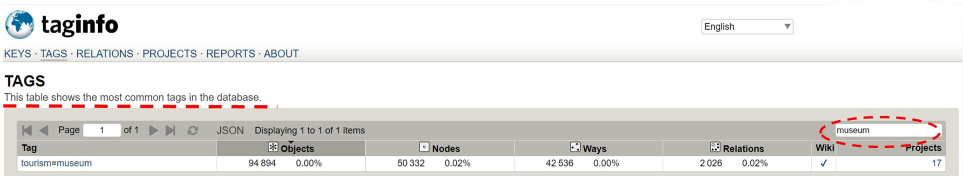
We also observe summary statistics for the target value prior log transformation in Table 4.2.

Variable Name	Mean	Standard Deviation	Minimum Price	Maximum Price
Price	114.26	128.43	9.00	5,000

Table 4.2: Summary Statistics for Target Variable

4.2 Additional Data

Additionally to the listings data we consider incorporating distance of listings to potential amenities of interest as explained earlier in Chapters 2 and 3. To compute these distances we acquire location data of amenities namely beaches, historical sites, bus stops and nightclubs from an API of *OpenStreetMap* namely *Overpass Turbo*. The OpenStreetMap database comprises of two main elements: nodes and edges. Every node and edge is assigned a *Tag* in *Overpass Turbo* by the contributors of the community and consequently this *Tag* is used to query a point of interest. Each *Tag* consists of two components: a key and a value. The key refers to a larger group a point of interest belongs to while the value is more specific. For example the tag for a museum can be as follows: 'tourism=museum' where tourism is the key and museum is the value. Contributors are free to give a *Tag* as they deem most appropriate and it is difficult to identify all the tags assigned by different contributors for say museums. However, *Taginfo*[3] is a site linked to OpenStreetMap data that provides statistics that are updated daily on tags in OpenStreetMap. This site is used to determine the best tags to use when querying each of the places of interest. For example when 'museum' is queried in *Taginfo* as shown in Figure 4.8 the most commonly used tags for museums are displayed.



TAGS

This table shows the most common tags in the database.

Displaying 1 to 1 of 1 items

Tag	Objects		Nodes		Ways		Relations		Wiki	Projects
tourism=museum	94 894	0.00%	50 332	0.02%	42 536	0.00%	2 026	0.02%	✓	17

Figure 4.8: Taginfo Screenshot for most popular tags in OpenStreetMap corresponding to museums

Choice of tags to query points of interest in OpenStreetMap are thus based on statistics from *Taginfo*. To compute distances of Airbnb listings to the capital city a prominent fixed point (node) in the city is selected for the Maltese Islands (Malta and Gozo) and similarly for Crete and distances from Airbnb listings were computed accordingly. Within the main island in Malta, the fixed point selected in Valletta is the Tritons' Fountain while within the smaller island, Gozo, the Cathedral of Assumption is taken as a fixed point in the capital city Rabat (Victoria). In Crete, the Heraklion Archaeological Museum in the capital city, Heraklion, is taken as a fixed point in the city. Choice of the latter is made based on the fact that this museum is located in the main square of the city, Eleftherias Square.

Furthermore, ethnicity data by locality for Malta is collected from the National Statistics Office in Malta and corresponds to data collected during the 2021 Population Census. Similar data for the island of Crete is requested from the Hellenic Statistical Authority in Greece however at the time of writing the Authority claimed that such data for 2021 Census population is not yet available. Data for Malta is provided in terms counts of residents by locality by ethnical background and total population counts by locality. This data is used to to compute a percentage proportion of non-Caucasian residents by locality and create a new variable within our

[3] https://taginfo.openstreetmap.org/

models that captures this information name *Residential_Ethnicity*.

Therefore final data set following pre-processing consists of three main components: Listings Data, Geospatial Data and Ethnicity Data. In total data is composed of 62 variables of which 30 are numerical, 28 are boolean and 4 are categorical.

4.3 Preliminary Analysis of Data for Malta and Crete

We perform a simple preliminary analysis to compare data for Malta and Crete. In Figure 4.9 we present a probability density of lnprice for Malta and Crete whereby listings for Crete are filtered for listings data scraped only in September 2022 for better comparability. Some similarity in the price distribution can be clearly observed although distribution for Malta is more symmetric around the mean. Furthermore, a table of summary statistics of the lnprice variable for the two domains is displayed in Table 4.3. In addition in Appendix A Table A.2 a table of correlation between lnprice and each of the variables is also provided for both Malta and Crete. We identify some variables to have very similar correlation with lnprice such as the number of bedrooms having a correlation of 0.53 for both Malta and Crete and the number of people a listing accommodates having a correlation of 0.54 for both islands too. Other variables such as the walkable distance to historical places has a differing correlation having a correlation of 0.12 and -0.02 for Crete and Malta respectively and the listing being a private room having a correlation of 0.008 and -0.37 for Crete and Malta respectively.

The variables in each of the data sets are the same both including listings data and additional geospatial distance data. However residential ethnicity variable will be excluded since as stated earlier data for Crete is still unavailable at the time of writing.

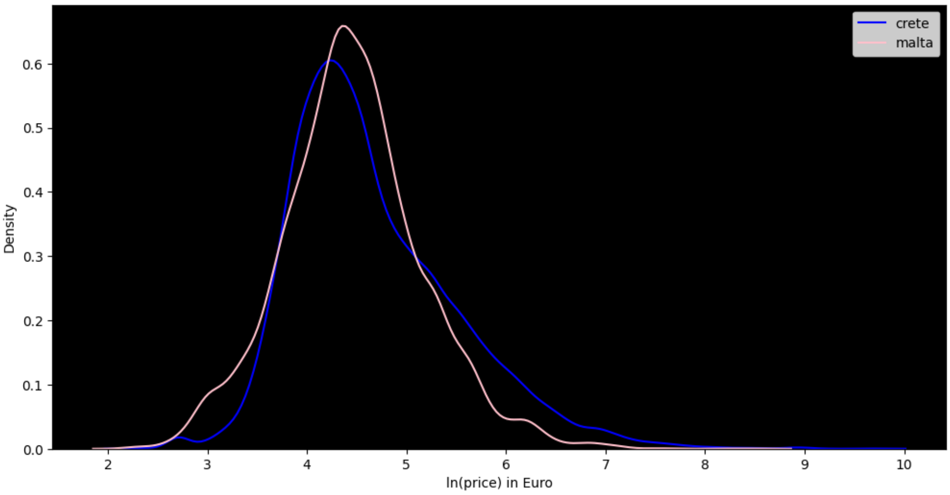


Figure 4.9: Probability density of lnPrice for the source domain (Crete) and target domain (Malta) as at September 2022

	Crete (Source)	Malta (Target)
Count of Observations	113,891	8,505
Mean	4.69	4.46
Standard Deviation	0.99	0.70
minimum	2.07	2.19
maximum	11.51	8.85

Table 4.3: Summary Statistics for ln (Target Variable) for Crete and Malta

5

Empirical Analysis

We will now outline the main findings of this research. For each model considered data was split into Training Data (80%) and Test Data(20%). No prior feature selection is made but rather we rely on intrinsic feature selection mainly regularization and tree based models. Intrinsic feature selection is a quick approach and enables the user to make feature selection taking into account the model objective for each model considered to strike a balance between feature sparsity and predictive performance of the model (Kuhn and Johnson, 2013).

5.1 Results from adding Geospatial Data

The initial findings of this research involve evaluating and comparing the predictive performance of various machine learning models using two types of data. The first type of data does not include distance features to nearby amenities of interest, while the second type includes distance features related to beaches, historical sites, the capital city, nightclubs, and bus stops. The objective was to analyze the impact of including these distance features on the predictive ability of the models. In total 8 distinct models have been fitted on each of the two types of data sets: KNN, Linear Regression, Ridge Regression, Lasso Regression, Random Forest, XGBoost, CatBoost, and a Neural Network. All models were implemented in Python using the libraries and modules specified in Appendix B Table B.1.

For KNN, the optimal number of neighbours, K, for this model is found to be 5 for both data sets considered. Selection of the optimal K is made based on a plot of cross-validation error versus number of neighbours shown in Figure 5.1.

Moving to linear regression models, when fitting initial linear regression model without regularization an issue can be identified where the model is struggling to identify categories in the same categorical variable setting very similar coefficient values to categories belonging to the same categorical variable. As a result, both Ridge and Lasso regularisation are explored and the optimal regularization parameters are found in Table 5.1. Parameter choice is made via a 5-fold cross validation on the basis of a grid search. Parameter grids considered are detailed in Appendix B Table B.2. Additionally, in Figure 5.2 a plot of the top 10 features with the highest absolute coefficient value following ridge regularization is illustrated. The same features have the highest absolute coefficient value for both data sets and sim-

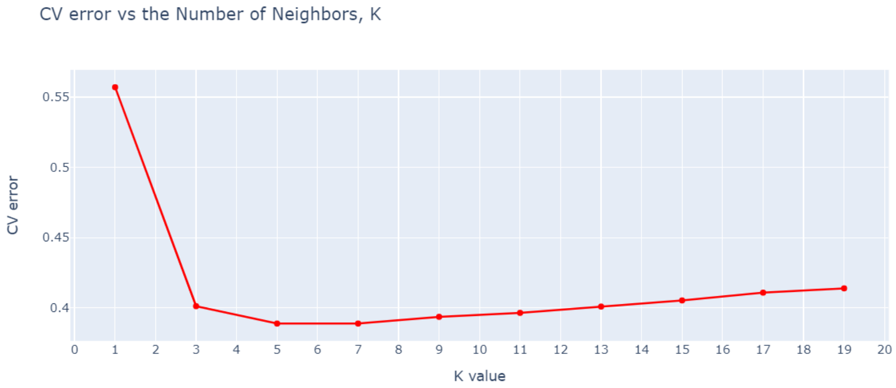
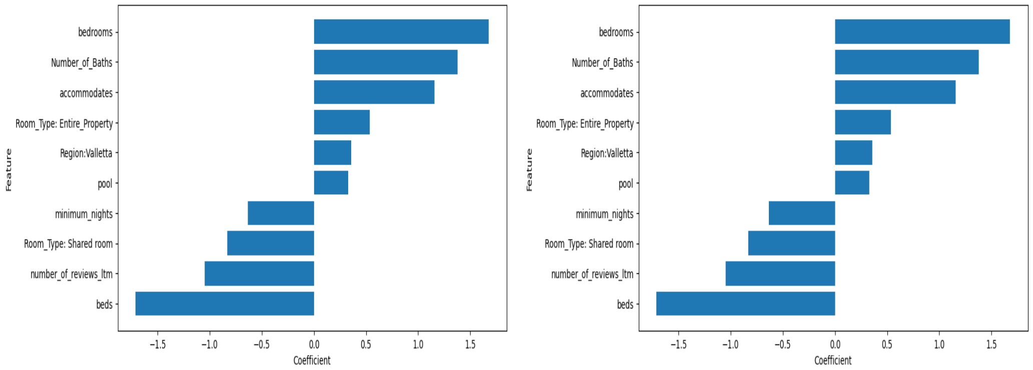


Figure 5.1: Plot of cross validation error versus number of neighbours in KNN model

Data	Ridge Regularization Parameter	Lasso Regularization Parameter
Without Geospatial features	0.5001	0.0001
With Geospatial Features	0.5001	0.0001

Table 5.1: Optimal regularization parameters for Linear Regression

ilar coefficient values with identical order by coefficient value is attained. None of the additional geospatial features are identified of excessive importance for price of a listing however the most important geospatial feature out of all the additional geospatial features is walkable distance to the capital with a coefficient of -0.27.

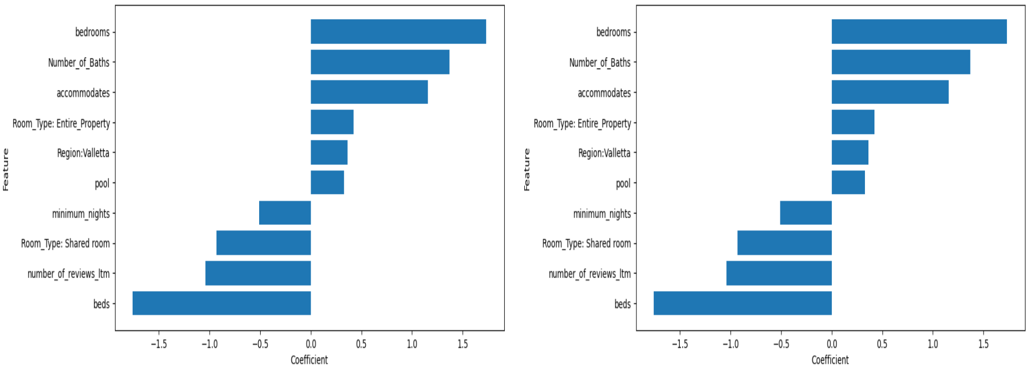


(a) Features with Top 10 absolute coefficient value post Ridge Regularization on data with no additional distance features

(b) Features with Top 10 absolute coefficient value post Ridge Regularization on data with additional distance features

Figure 5.2: Feature Importance for Ridge Regularization Model

Since Lasso regularization favours sparse models implicit feature selection is performed. In Figure 5.3 the top 10 features with the highest absolute coefficient value for both when geospatial data is excluded and included are displayed. Feature selection is almost identical on both data sets and similar coefficient values are assigned. For both datasets, variables capturing information concerning the number of days still available for booking in the next 90 days on the day the data was scraped, the number of days a host has been active and whether a listing is located in the locality 'San Gwann' have been deemed irrelevant and have coefficients set to zero in both model fittings. Additionally, for the model fitted on the data set without additional geospatial features, the variables longitude and whether a listing is located in the locality 'Mellieha' and locality ' Swieqi' are also deemed irrelevant to price prediction, while for the model with additional geospatial features whether the room type is private is additionally considered as irrelevant. None of the additional geospatial features are set to zero by lasso. Having said that none of them have a remarkably high importance to the model with the most important distance feature being walkable distance to the capital with a coefficient value of -0.22.



(a) Model without additional Geospatial Features

(b) Model with additional Geospatial Features

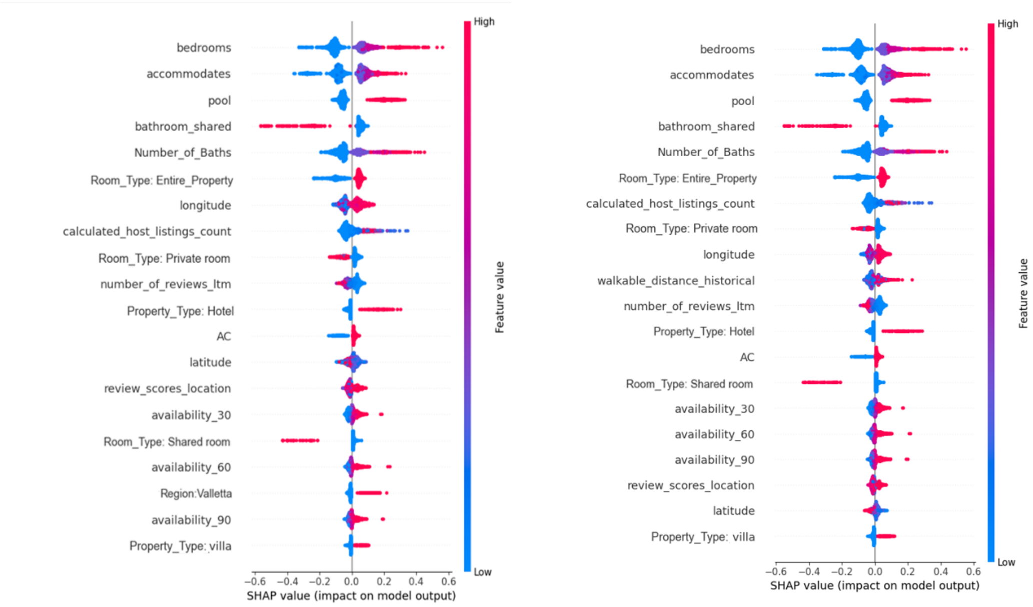
Figure 5.3: Features with top 10 absolute coefficient values post Lasso Penalization

Next we look into insights from the fitted random forest regression model. Initially just like other previously mentioned machine learning models, categorical variables are one hot encoded at first. Once again hyper parameters are determined via a 5 fold cross validation based on a random search followed by a grid search. The optimal parameters following parameter tuning are found in Table 5.2. Further details on the parameter ranges considered during the searches can be found in Appendix B Table B.3.

Parameter	Without Geospatial Data	With Geospatial Data
n_estimators	819	995
max_depth	None	None
max_features	28	29
min_samples_leaf	2	2

Table 5.2: Optimal Parameters for Random Forest

In Figure 5.4, the 20 important features on the basis of Shapley values are attained for the model fitted on the data set without (Left) and with (right) geospatial features. It is clear that high number of bedrooms, high accommodation capacity, presence of a pool in the listing and the number of bathrooms has a high and positive relationship with the price of a listing. On the contrary, listings with a shared bathroom have a remarkable negative relationship on the price of a listing. Analysing further the Shapley values for the extended model with geospatial data it is observed that one of the additional distance features *walkable_distance_historical* also is considered among the top 20 important feature but surprisingly walkable distance to a historical site is positively correlated with price meaning that listings further away from a historical site are more likely to have a higher price. Looking into this further it is found that the variable indicating whether a listing is an entire property or not, *Room Type: Entire property* has an interesting relation with walkable distance to historical sites as shown in Figure 5.5. Firstly, the u-shape is an indication that listings super close to historical sites and listings very far away from historical sites have higher listing prices. Secondly, from this plot we derive an interesting insight that being a listing that is an entire property lowers the impact of proximity to a historical site on price while being a listing that is a hotel room, a shared room or a private room increases the impact of proximity to a historical site on price.



(a) Top 20 important features based on Shapley values for Random Forest fitted on data with no additional distance features

(b) Top 20 important features based on based on the Shapley values for Random Forest fitted on data with additional distance features

Figure 5.4: Shapley values random forest

Being a tree based model, we further examine the performance of random forest by transforming categorical variables *host_response_time*, *property_type* and *room_type* into ordinally encoded variables based on the author's rationale. Order of categories that is deemed suitable is displayed in Table 5.3. No rational order for *region_name* is deemed suitable and instead target encoding is considered whereby each region name category is replaced by the mean of the price of listings in the training data in that particular region. To avoid any data leakage from the test data the mean is only computed on data within the training data set. Unfortunately, no improvement in model performance is observed when performing ordinal and target encoding. Therefore one-hot encoded categorical variables give best performance when fitting a random forest.

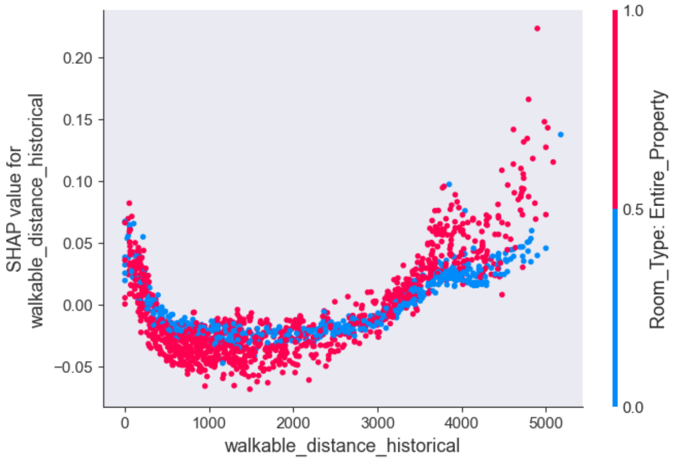


Figure 5.5: Partial Dependence Plot of Shapley values for walkable_distance_historical by Room_Type: Entire Property indicator as color

Categorical Variable	Order of Categories
host_response_time	'within an hour', 'within a few hours', 'within a day', 'a few days or more'
Property_Type	'other','apartment','maisonette','house', 'bed_and_breakfast','guesthouse','hotel','villa'
Room_Type	'Shared room','Private room', 'Hotel room','Entire Property'

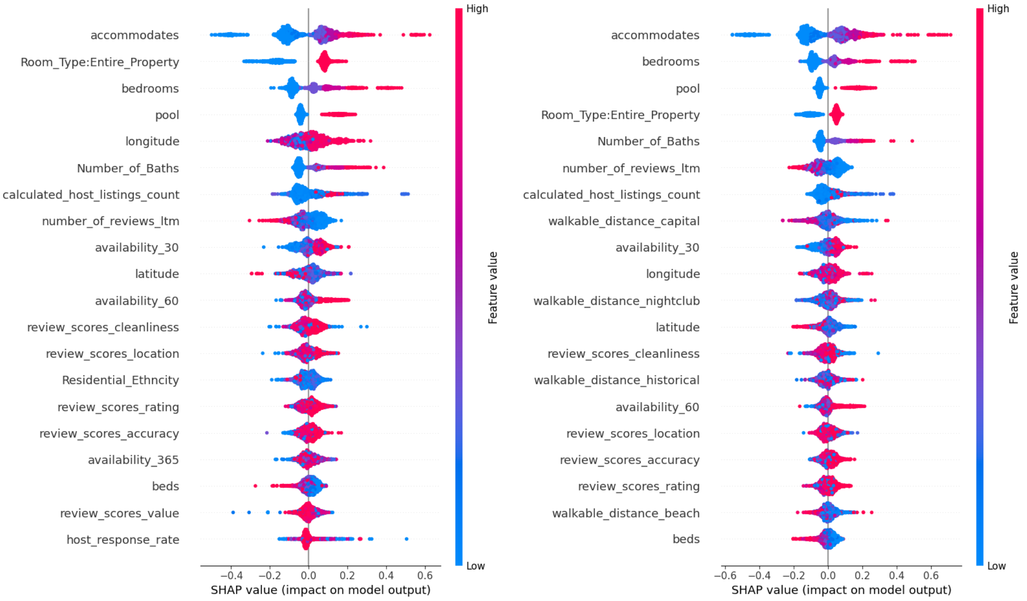
Table 5.3: Order of categories in categorical variables for Ordinal Encoding

We further examined this regression task problem with the use of a gradient boosting algorithms. First we consider fitting an XGBoost model. In Table 5.4 we present a list of optimal parameters that were tuned via random search followed by a grid search based on a 5-fold cross validation. The optimal parameters are listed in the Table 5.4 too. For further details on the range of parameters considered during the search exercise see Table B.4 in Appendix B.

Parameters	Without Geospatial Data	With Geospatial Data
n_estimators	683	967
max_depth	5	5
learning_rate	0.41	0.48
reg_alpha	1	1
reg_lambda	0	0
objective	reg:squarederror	reg:squarederror
subsample	0.88	0.88
colsample_bytree	0.11	0.13

Table 5.4: Optimal Parameters for XGBoost

We present the top 20 important features on the basis of Shapley values for the XGBoost model fit in Figure 5.6. Similarly to random forest, higher capacity of a listing, higher number of bedrooms and the presence of a pool are associated with higher Shapley values and hence higher price of listings for both data sets. In contrast to random forest the feature providing information whether the listing has shared bathrooms or not is not considered as important for XGBoost. XGBoost identifies the walkable distance to the capital to have the highest contribution to the price of a listing among all additional walkable distance features and similar to the linear regression model with applied Ridge and Lasso regularization earlier identifies listings with a short walkable distance to the capital to have higher positive contribution on price (higher prices) although some exceptions are also present as shown below.



(a) Top 20 important features based on Shapley values for XGBoost fitted on data with no additional distance features

(b) Top 20 important features based on the Shapley values for XGBoost fitted on data with additional distance features

Figure 5.6: Shapley values XGBoost

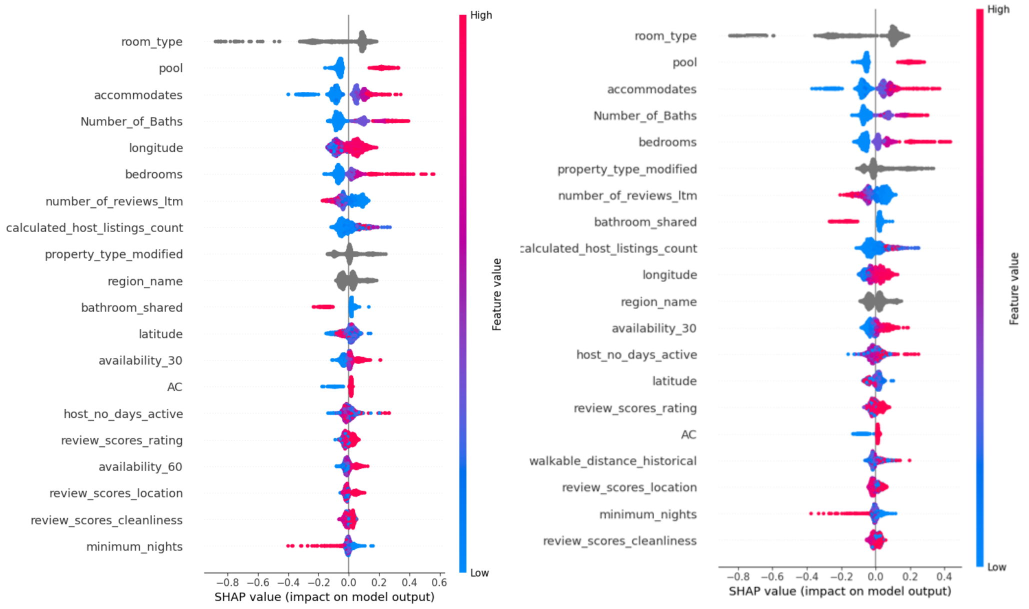
Next we review findings from another boosted tree model, CatBoost. Once again parameter tuning of parameters is carried out via 5 fold cross validation on the basis of a random search followed by a grid search. The optimal parameters are displayed in Table 5.5. For further details on the grids where parameters tuning was performed see Appendix B Table B.5. As a means of regularization, l2 regularization through the parameter $l2\_leaf\_reg$ is considered together with injecting some randomness in the scores of potential splits at each iteration through the parameter $'random\_strength'$. This randomness initialises as noise with mean 0 and variance $1 \times' random\_strength'$ and the set variance then declines with the number of iterations.

In Figure 5.7 we present a plot of the Top 20 important features via the Shapley values in the CatBoost models fitted on the two data sets. Similar to results from random forest fitting and XGBoost, the availability of a pool, the capacity of a listing and the number of baths in a listing are considered as prominent contributors to price listings having a positive relationship with price. However in contrast to previous model fittings here categorical variables are not one hot encoded as explained in the Methodology Section and corresponding Shapley values are shown in

Parameters	Without Geospatial Data	With Geospatial Data
iterations	625	950
depth	9	9
learning_rate	0.1	0.1
l2_leaf_reg	6	3
random_strength	3	3

Table 5.5: Optimal Parameters for Catboost

grey for these variables. Comparing the models fitted on the two different data sets the top 20 important features for Catboost are very similar with minor differences. In fact, only one feature *walkable_distance_historical* from the additional distance features makes it to the top 20 list of most important features in Catboost and is rather at the bottom of the list implying minor importance to the model. Another remark to make is that the *room_type* is considered as the most important feature. This contrasts the results attained in Figure 5.4 for random forest and Figure 5.6 for XGBoost. Recall that in random forest model fitting and XGBoost we stick to one hot encoding unlike in Catboost. This confirms that the importance of one hot encoded variables in tree based models when one hot encoded is underestimated.



(a) Top 20 important features based on Shapley values for Catboost fitted on data with no additional distance features

(b) Top 20 important features based on based on the Shapley values for Catboost fitted on data with additional distance features

Figure 5.7: Shapley values Catboost

Looking into the Shapley values for CatBoost grouped by category for the most important feature *room_type* for the model fitted on Geospatial data in Figure 5.8 we can deduce some expected insights. It is observed that listings for entire places are associated with higher Shapley values and hence higher nightly listing prices while shared rooms are associated with lower Shapley values and hence lower nightly listing prices.

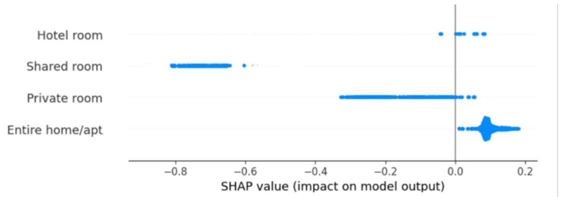
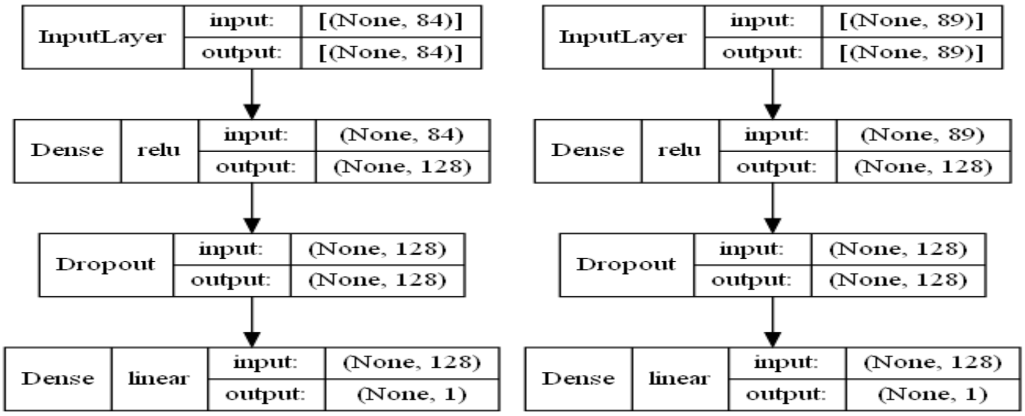


Figure 5.8: Shapley values for room_type grouped by category

As a final model a NN is considered. The architecture of the best performing NN is displayed in Figure 5.9.



(a) Neural Network fitted to data without additional Geospatial features

(b) Neural Network fitted to data with additional Geospatial features

Figure 5.9: Architecture of Best Performing Neural Networks

We deduce that a fully connected NN with 1 hidden layer and ReLU activation function applied to the hidden layer is the best. The ADAM optimizer is found to be best optimizer when compared with RMSprop while mean squared error is the loss opted for. Furthermore, as a means of regularization dropout is experimented with. In Table 5.6 we highlight the optimal parameters derived from tuning. For further details on the ranges of parameters considered during search see Table B.6 in Appendix B.

Parameters	Without Geospatial Data	With Geospatial Data
activation_function (for hidden layer)	ReLU	ReLU
learning_rate	0.0001	0.01
dropout_rate	0.07	0.33
epochs	200	150
batch_size	64	128

Table 5.6: Optimal Parameters for Fully-Connected Feed Forward Neural Network fitted on Data for Malta

As can be seen in the next Section, the NN with optimal parameters performed particularly poorly in comparison to advanced machine learning models such as random forest and boosting based algorithms. We consider modifying the fitted NN by considering a transfer learning approach by first fitting a NN to data from the island of Crete and completely excluding data for listings in Malta. The optimal architecture of the NN fitted is shown in Figure 5.10. Also note that here we drop the variables longitude and latitude and region name variables since this is not valuable transferable knowledge from one domain to the other. Optimal parameters for this NN are identified through an 8 fold cross-validation via random search and grid search. More folds are considered here since the data set is larger for Crete. The optimal parameters attained are shown in Table 5.7. For parameter ranges considered in the search see Table B.7. The best performing model is found to have 2 hidden layers whereby ADAM optimizer is found to be the best optimizer and mean squared error is once again considered as a loss function. The activation function on the output layer is also once again taken to be 'linear'.

Parameters	With Geospatial Data
activation_function (for hidden layer)	ReLU
learning_rate	0.0001
dropout_rate	0.1
epochs	350
batch_size	64

Table 5.7: Optimal Parameters for Fully-Connected Feed Forward Neural Network fitted on Crete Data

Next we consider freezing the first hidden layer and re train the rest of the model with data for Malta and consequently perform predictions on the the data for Malta. Architecture of the NN fitted is displayed in 5.10.

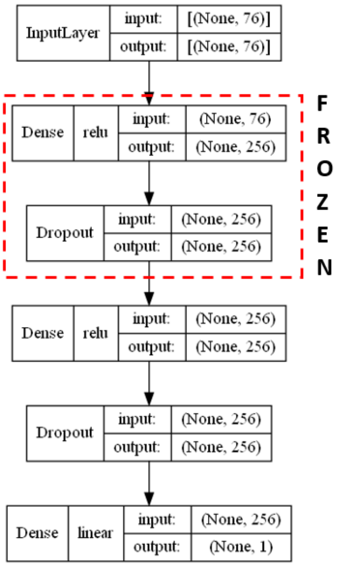


Figure 5.10: Neural Network with Geospatial data for Crete

5.2 Performance Overview

In Table 5.8 we display the R^2 obtained on test data for each of the models considered on both data containing no additional geospatial features and data containing additional geospatial features. CatBoost is seen to be the model with the highest R^2 on test data for both type of data sets achieving an R^2 of 0.7662 for the model fitted on data without distance features and 0.7700 for model with additional features. Compared to the KNN benchmark model an improvement in model performance on test data is observed for both data sets by all models fitted a posteriori. However, only a minor difference in performance can be observed between models fitted to the data set without additional geospatial features and models fitted to the data set with additional geospatial features. All linear regression models fitted reported a minor improvement when model is fitted to data set including additional geospatial features. Recall that none of the additional geospatial features were set to have coefficients zero by the Lasso and very similar performance is observed for all regression models. Similarly, CatBoost and the NN (without transfer learning) report a slight improvement in performance by the model including geospatial data features. Moreover when modifying the neural network to be pre-trained on data for Crete a slight improvement can be observed reporting an R^2 of 0.6662 over the performance of the neural network trained solely on data for Malta reporting an R^2 of 0.6561. Having said that, CatBoost model still remains the best performing model.

Model	Without Geospatial Data R^2	With Geospatial Data R^2
KNN (benchmark)	0.6200	0.6193
Linear Regression(No Regularization)	0.6331	0.6359
Ridge Regression	0.6336	0.6356
Lasso Regression	0.6334	0.6344
Random Forest Regression	0.7358	0.7319
XGBoost	0.7086	0.6957
CatBoost	0.7662	0.7700
Neural Network without Transfer Learning	0.6489	0.6561
Neural Network with Transfer Learning	-	0.6662

Table 5.8: Summary of results comparing R-Squared value on test data

5.3 Discussion

In this thesis we set off our research with the aim to derive a price prediction model for Airbnb listing prices in the Maltese islands through a machine learning approach. As is seen, the boosting algorithm, Catboost, triumph among other machine learning models considered. This result is expected as from existing literature analyzing listings in other cities through a machine learning approach boosting algorithms mainly XGBoost are declared as the best performing models such as Liu (2021) and Yang (2021) as reported earlier in the Literature Review. Having said that, by experimenting and considering a modification of XGBoost, namley CatBoost great improvements in model performance are observed over XGBoost. Furthermore, with regards to studies carried out for Airbnb listings in the Maltese islands a drastic improvement is achieved in model performance when compared to other literature analysing Airbnb price listings in Malta through different approaches as discussed earlier in the Literature Review reporting a best R^2 of 0.51 (Fearne, 2022)) and 0.38 (Ellul, 2019). Our best model achieved an R^2 of 0.77.

Looking into the feature importances identified in our research by the best performing algorithm, CatBoost, features of importance such as whether a property is an entire place or not are in line with those outlined by Fearne (2022) where the latter identified this feature as having the highest absolute regression coefficient in an OLS regression claiming that being an entire property contributes to a raise of 70% in nightly price over listings corresponding to a room within a property. However, Fearne (2022) also claims that for each additional bathroom the price per night drops. This is in contrast to our findings and a potential reason for this is that in our research an additional variable is engineered and considered to account for whether the bathrooms available are shared or not. Furthermore, Fearne (2022) also identifies pool availability as the amenity that leads to the highest price rise which is in line with our findings claiming that pool availability is the most important feature contributing to price.

Consequently, when analysing the feature importance for each model considered, we conclude none of the additional distance features in any model considered are identified as relatively important features. Having said that, all linear regression models reported a slight improvement for when the model is fitted with additional geospatial features whereby both linear regression models with applied regularization (ridge and lasso) are at a common consensus that that the walkable distance to the capital is the most important feature out of all additional walkable distance features reporting a negative coefficient of -0.27 and -0.22 respectively indicating longer walkable distances to the capital are associated with lower prices. Similar to linear regression models, Catboost also reported a slight improvement in model performance when adding geospatial distance features, but also did not identify any of the additional distance features of remarkable importance based on our analysis of the Shapley values. However, the walkable distance to historical sites is considered as the variable with the greatest contribution to nightly price of Airbnb listings out of all geospatial distance variables added. To our surprise the walkable distance to the beach is considered of very little importance to explain the price of Airbnb listings across all models fitted. This contradicts the findings by Chica-Olmo et al.

(2020) which identified distance to the beach as the most important feature for explaining Airbnb prices when compared with distance to the centre and distance to the nearest place of interest for the coastal city Malaga. Similarly, in our research walkable distance to bus stops are not relatively important to any of the models fitted. A similar finding is outlined by Schwarzová (2020) who claims that walkable distance to train station is not relatively important to explain variation in Airbnb price listings in Prague. Proximity to nightclubs is also not relatively important to any of the models fitted in our research.

Another interesting finding drawn from this research is that by considering listings data as well as geospatial data for Crete which is a also a touristic island with similar natural characteristics like Malta, the performance of the deep learning model fitted is improved over the model solely trained on data for Malta. This improvement highlights that transferring knowledge from one island to the other proved to be beneficial and domains are similar enough for knowledge transfer to have a positive impact on model performance.

6

Conclusion

In conclusion an extensive experimentation exercise with various machine learning models have been considered to predict Airbnb listing prices for the Maltese islands. The CatBoost algorithm is shown to have the best performance by being the most generalisable model beyond training data. Further to this we identified a number of characteristics of a listing that are bound to more likely push the nightly price of a listing. Based on the feature importance of the best performing model, whether a listing corresponds to an entire property or a room within a property is the most influential feature on the price of a listing. Furthermore the presence of a pool is identified as the most influential amenity within a listing on price. Additionally, in this research we concluded that including additional geospatial distance features of most popular attractions in Malta as well as distance to public transportation amenities does not lead to a remarkable improvement in model performance and hence come to a conclusion that proximity of an Airbnb listing to places of interest and public transportation access does not have remarkable implications on price. Finally, from this research we conclude that pre-training a deep learning model (NN) on data for a similar touristic island to Malta such as Crete did slightly improve model performance on prediction for our market of interest however this improvement is not enough to beat the performance of the boosting algorithm considered.

6.1 Limitations

A number of limitations revolve around this research. The first is that data for Malta is very limited in terms of number of observations due to the nature of the size of the islands as well the fact that data available is scraped at one point in time. Further to this, since data is only available at one point in time this eliminates any potential to investigate seasonality aspect within the data. Further to this, time constraints to further refine and experiment with the transfer learning approach is also considered a main limitation in this study.

6.2 Future Research

This research serves as a good starting point for further research studies of Airbnb listings within the Maltese islands through a machine learning approach. A further extension to this research would be to undertake an ensemble stacking approach whereby multiple machine learning models fitted in this research are taken as base models of a final model. Further to this, another addition to this research would be to consider a density index of commercial establishments such as restaurants within say 500m radius of a listing. Moreover, future potential research could also be done by considering an instance based transfer learning approach whereby a weighting procedure can be considered to weight observations in the source domain such that data in the source domain that is considered more valuable for the target domain is given more weight than other data.

Bibliography

C. Adamiak. Mapping airbnb supply in european cities. *Annals of Tourism Research*, 71:67–71, 2018. ISSN 0160-7383. doi: https://doi.org/10.1016/j.annals.2018.02.008. URL https://www.sciencedirect.com/science/article/pii/S0160738318300148.

C. Adamiak. Current state and development of airbnb accommodation offer in 167 countries. *Current Issues in Tourism*, 25(19):3131–3149, 2019.

R. Ambolkar, A. Bhagat, B. Buga, and S. Gharat. Hotel recommendation system using advanced efficiency and accuracy with modified bert technique. In *Proceedings of the 2nd International Conference on Artificial Intelligence and Smart Energy, ICAIS 2022*, pages 1747–1752, 2022. URL www.scopus.com.

K. Ayouba, M.-L. Breuillé, C. Grivault, and J. Le Gallo. Does airbnb disrupt the private rental market? an empirical analysis for french cities. *International Regional Science Review*, 43(1-2):76–104, 2020.

D. Bennett. How can i deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25:464 – 469, 10 2001. doi: 10.1111/j.1467-842X.2001.tb00294.x.

M. Bernardi and M. Guidolin. The determinants of Airbnb prices in New York City: a spatial quantile regression approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(1):104–143, 02 2023. ISSN 0035-9254. doi: 10.1093/jrsssc/qlad001. URL https://doi.org/10.1093/jrsssc/qlad001.

G. Boeing. Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65: 126–139, 2017. ISSN 0198-9715. doi: https://doi.org/10.1016/j.compenvurbsys.2017.05.004. URL https://www.sciencedirect.com/science/article/pii/S0198971516303970.

R. Cellmer, M. Bełej, and J. Konowalczuk. Impact of a vicinity of airport on the prices of single-family houses with the use of geospatial analysis. *ISPRS International Journal of Geo-Information*, 8(11), 2019. ISSN 2220-9964. doi: 10.3390/ijgi8110471. URL https://www.mdpi.com/2220-9964/8/11/471.

T. Chen and C. Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145%2F2939672.2939785.

Y. Chen and K. Xie. Consumer valuation of airbnb listings: a hedonic pricing approach. *International Journal of Contemporary Hospitality Management*, 29: 2405–2424, 09 2017. doi: 10.1108/IJCHM-10-2016-0606.

J. Chica-Olmo, J. G. González-Morales, and J. L. Zafra-Gómez. Effects of location on airbnb apartment pricing in málaga. *Tourism Management*, 77: 103981, 2020. ISSN 0261-5177. doi: https://doi.org/10.1016/j.tourman.2019.103981. URL https://www.sciencedirect.com/science/article/pii/S0261517719301797.

M. Crisci, F. Benassi, H. Rabiei-Dastjerdi, and G. McArdle. Spatio-temporal variations and contextual factors of the supply of airbnb in rome. an initial investigation. *Letters in Spatial and Resource Sciences*, 15(2):237–253, 2022.

R. Deboosere, D. Kerrigan, D. Wachsmuth, and A. El-Geneidy. Location, location and professionalization: a multilevel hedonic analysis of airbnb listing prices and revenue. *Regional Studies, Regional Science*, 6(1):143–156, 2019. doi: 10.1080/21681376.2019.1592699. URL https://doi.org/10.1080/21681376.2019.1592699.

Deloitte. Mhra hotel survey by deloitte. key highlights: Q4 2022 and year to date. Webpage, 2023. URL https://www2.deloitte.com/mt/en/pages/finance/articles/mt-hotel-performance-survey.html. Accessed 2023/03/29.

G. Dudás, L. Boros, T. Kovalcsik, and B. Kovalcsik. The visualization of the spatiality of airbnb in budapest using 3-band raster representation. *Geographia Technica*, 12:23–30, 05 2017. doi: 10.21163/GT_2017.121.03.

R. Ellul. Short-term rentals in malta: A look at airbnb listings, 2019. URL https://books.google.se/books?id=nv7nzQEACAAJ.

European_Commission. Economic forecast for malta. Webpage, 2023. URL https://economy-finance.ec.europa.eu/economic-surveillance-eu-economies/malta/economic-forecast-malta_en. Accessed 2023/03/27.

R. Fearne. An analysis of the distribution and price determinants of airbnb rentals in malta. *International Journal of Housing Markets and Analysis*, 15(1):231–246, 2022. doi: 10.1108/IJHMA-12-2020-0147. URL h.

C. Gibbs, D. Guttentag, U. Gretzel, J. Morton, and A. Goodwill. Pricing in the sharing economy: a hedonic pricing model applied to airbnb listings. *Journal of Travel and Tourism Marketing*, 35(1):46–56, 2017.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2528823.

S. Gunduz, U. Ugurlu, and I. Oksuz. Transfer learning for electricity price forecasting. *Sustainable Energy, Grids and Networks*, 34:100996, 2023. ISSN 2352-4677. doi: https://doi.org/10.1016/j.segan.2023.100996. URL https://www.sciencedirect.com/science/article/pii/S2352467723000048.

U. Gunter and I. Önder. Determinants of airbnb demand in vienna and their implications for the traditional accommodation industry. *Tourism Economics*, 24 (3):270–293, 2018. doi: 10.1177/1354816617731196. URL https://doi.org/10.1177/1354816617731196.

D. Gutentag. Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector. *Current Issues in Tourism*, 18(12):1–26, 2013.

K. Gyódi and Nawaro. Determinants of airbnb prices in european cities: A spatial econometrics approach. *Tourism Management*, 86:104319, 2021. ISSN 0261-5177. doi: https://doi.org/10.1016/j.tourman.2021.104319. URL https://www.sciencedirect.com/science/article/pii/S0261517721000388.

J. Hamari, M. Sjöklint, and A. Ukkonen. The sharing economy: Why people participate in collaborative consumption. *Journal of the Association for Information Science and Technology*, 67(9):2047–2059, 2016. doi: https://doi.org/10.1002/asi.23552. URL https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23552.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, 2009. ISBN 9780387848587. URL https://books.google.se/books?id=tVIjmNS3Ob8C.

R. Hill and M. Scholz. Can geospatial data improve house price indexes? a hedonic imputation approach with splines. *Review of Income and Wealth*, 64, 07 2017. doi: 10.1111/roiw.12303.

B. Ibragimov and G. Gusev. Minimal variance sampling in stochastic gradient boosting, 2019.

J. Jiao and S. Bai. An empirical analysis of airbnb listings in forty american cities. *Cities*, 99:102618, 2020. ISSN 0264-2751. doi: https://doi.org/10.1016/j.cities.2020.102618. URL https://www.sciencedirect.com/science/article/pii/S0264275119306559.

N. Kanakaris and N. Karacapilidis. Predicting prices of airbnb listings via graph neural networks and document embeddings: The case of the island of santorini. *Procedia Computer Science*, 219:705–712, 2023.

M. Kuhn and K. Johnson. *Applied Predictive Modeling*. Springer, 2013. ISBN 978-1-4614-6848-6.

R. Liang, K. Jaewook, and W. Xi. Estimating spatial effects on peer-to-peer accommodation prices: Towards an innovative hedonic model approach. *International Journal of Hospitality Management*, 81:43–53, 08 2019. doi: 10.1016/j.ijhm.2019.03.012.

A. Liaw and M. Wiener. Package 'randomforest': Breiman and cutler's random forests for classification and regression. *R Development Core Team*, 4:6–10, 01 2014.

A. Lindholm, N. Wahlström, F. Lindsten, and T. B. Schön. *Machine Learning - A First Course for Engineers and Scientists*. Cambridge University Press, 2022. URL https://smlbook.org.

Y. Liu. Airbnb pricing based on statistical machine learning models. In *2021 International Conference on Signal Processing and Machine Learning (CONF-SPML)*, pages 175–185, 2021. doi: 10.1109/CONF-SPML54095.2021.00042.

T. Lorde, J. Jacob, and Q. Weekes. Tourism management perspectives price-setting behavior in a tourism sharing economy accommodation market: A hedonic price analysis of airbnb hosts in the caribbean . *Tourism Management Perspectives*, 30: 251–261, 04 2019. doi: 10.1016/j.tmp.2019.03.006.

S. Mathur. Impact of transit stations on house prices across entire price spectrum: A quantile regression approach. *Land Use Policy*, 99:104828, 2020. ISSN 0264-8377. doi: https://doi.org/10.1016/j.landusepol.2020.104828. URL https://www.sciencedirect.com/science/article/pii/S0264837719323579.

D. Micci-Barreca. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *SIGKDD Explorations*, 3:27–32, 07 2001. doi: 10.1145/507533.507538.

R. Mousavi and K. Zhao. Examining the impacts of airbnb review policy change on listing reviews. *Journal of the Association for Information Systems*, 23(1): 303–328, 2022.

L. S. Nguyen, S. Ruiz-Correa, M. S. Mast, and D. Gatica-Perez. Check out this place: Inferring ambiance from airbnb photos. *IEEE Transactions on Multimedia*, 20(6): 1499–1511, 2018. doi: 10.1109/TMM.2017.2769444.

NSOMalta. Inbound tourism:december 2022. Webpage, 2023. URL https://nsocms.gov.mt/en/News_Releases/Documents/2023/02/News2023_020.pdfn. Accessed 2023/03/28.

OECD. Oecd tourism trends and policies 2022. OECD Publishing,Paris, 2022. URL https://doi.org/10.1787/a8dd3019-en. Accessed 2023/03/25.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. Catboost: unbiased boosting with categorical features. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf.

P. Rezazadeh, L. Nikolenko, and H. Rezaei. Airbnb price prediction using machine learning and sentiment analysis. In A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, editors, *Machine Learning and Knowledge Extraction*, pages 173–184, Cham, 2021. Springer International Publishing. ISBN 978-3-030-84060-0.

R. Sainaghi. Determinants of price and revenue for peer-to-peer hosts. the state of the art. *International Journal of Hospitality Management*, 33:557–586, 03 2021. doi: 10.1108/IJCHM-08-2020-0884.

L. Schwarzová. Predicting airbnb prices with neighborhood characteristics: Machine learning approach. Master's thesis, Tilburg University, 2020.

S. Sun, X. Wang, and M. Hu. Spatial distribution of airbnb and its influencing factors: A case study of suzhou, china. *Applied Geography*, 139:102641, 2022. ISSN 0143-6228. doi: https://doi.org/10.1016/j.apgeog.2022.102641. URL https://www.sciencedirect.com/science/article/pii/S0143622822000121.

R. Suárez-Vega and J. Hernández. Selecting prices determinants and including spatial effects in peer-to-peer accommodation. *ISPRS International Journal of Geo-Information*, 9(4), 2020. ISSN 2220-9964. doi: 10.3390/ijgi9040259. URL https://www.mdpi.com/2220-9964/9/4/259.

C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. A survey on deep transfer learning. In V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis, editors, *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 270–279, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01424-7.

F. Tang and H. Ishwaran. Random forest missing data algorithms. 01 2017.

T. Teubner, F. Hawlitschek, and D. Dann. Price determinants on airbnb: How reputation pays off in the sharing economy. *Journal of Self-Governance and Management Economics*, 5:53–80, 04 2017. doi: 10.22381/JSME5420173.

V. Toader, A. Negrușa, O. R. Bode, and R. Rus. Analysis of price determinants in the case of airbnb listings. *Economic Research-Ekonomska Istraživanja*, 35(1): 2493–2509, 2022. doi: 10.1080/1331677X.2021.1962380. URL https://doi.org/10.1080/1331677X.2021.1962380.

J. Van den Hurk. Transfer learning for price prediction in real estate. Master's thesis, Tilburg University, 2021.

M. Vladimirova. The brand image of malta as a tourism destination: a case study in public relations and corporate communication practice. 2011.

A. K. Waljee, A. Mukherjee, A. G. Singal, Y. Zhang, J. Warren, U. Balis, J. Marrero, J. Zhu, and P. D. Higgins. Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, 3(8), 2013. ISSN 2044-6055. doi: 10.1136/bmjopen-2013-002847. URL https://bmjopen.bmj.com/content/3/8/e002847.

J. Wallin. Lecture 3 slides on boosting, February 2022.

Q. Yang, Y. Zhang, W. Dai, and S. J. Pan. *Transfer learning*. Cambridge University Press, 2020.

S. Yang. Learning-based airbnb price prediction model. In *2021 2nd International Conference on E-Commerce and Internet Technology (ECIT)*, pages 283–288. IEEE, 2021.

Y. Yang, Z. Mao, and J. Tang. Understanding guest satisfaction with urban hotel location. *Journal of Travel Research*, 57(2):243–259, 2018. doi: 10.1177/0047287517691153. URL https://doi.org/10.1177/0047287517691153.

M. Yazdani. Machine learning, deep learning, and hedonic methods for real estate price prediction, 2021.

G. Zervas, D. Proserpio, and J. Byers. The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. *Journal of Marketing Research*, 54 (5):687–705, 2017. doi: 10.1509/jmr.15.0204. URL https://doi.org/10.1509/jmr.15.0204.

S. Zhang, D. Lee, P. V. Singh, and K. Srinivasan. What makes a good image? airbnb demand analytics leveraging interpretable image features. *Management Science*, 68(8):5644–5666, 2022.

I. Önder, C. Weismayer, and U. Gunter. Spatial price dependencies between the traditional accommodation sector and the sharing economy. *Tourism Economics*, 25(8):1150–1166, 2019. doi: 10.1177/1354816618805860. URL https://doi.org/10.1177/1354816618805860.

Appendix A

Further Data Details

Table A.1: Listing Data variables post data cleaning and pre-processing

Variable Name	Description of Variable	Data Type
Accommodates	The capacity of the listing	Numerical
Bedrooms	Count of Bedrooms in the Listing	Numerical
Beds	Count of Beds in the Listing	Numerical
Number_of_Baths	Count of Baths in the Listing	Numerical
bathroom_shared	Are the bathrooms available shared ?	Boolean
room_type	Type of listing	Categorical:Entire home, Hotel Room, Private Room, Shared Room
property_type	Type of property listing is located in	Categorical:Apartment, maisonette, house, villa,bed_and_breakfast, hotel, guesthouse, other
region_name	Region or Locality listing is located in	Categorical:Sliema, San Giljan, San Pawl il_Bahar,Valletta, Gzira, Mellieha, Swieqi, Msida, Marsascala, Xaghra, Zebbug(Ghawdex), San Gwann, North_Western_Other, South_Eastern_Other, Gozo_Comino_Other
Latitude	Latitude of Listing	Numerical
Longitude	Longitude of Listing	Numerical

Continued on next page

Table A.1 – continued from previous page

Variable Name	Description of Variable	Data Type
minimum_nights	The lowest number of nights the listing can be booked for	Numerical
maximum_nights	The highest number of nights the listing can be booked for	Numerical
availability_30	The number of days listing is available in the next 30 days	Numerical
availability_60	The number of days listing is available in the next 60 days	Numerical
availability_90	The number of days listing is available in the next 90 days	Numerical
availability_365	The number of days listing is available in the next 365 days	Numerical
number_of_reviews_ltm	The number of guest reviews of listing in the past 12 months	Numerical
review_scores_rating	Average of guest review rating on the whole experience during stay	Numerical
review_scores_accuracy	Average of guest review rating on accuracy of listing details on Airbnb site	Numerical
review_scores_cleanliness	Average of guest review rating on hygiene of listing	Numerical
review_scores_checkin	Average of guest review rating on check-in experience	Numerical
review_scores_communication	Average of guest review rating on host responsiveness	Numerical
review_scores_location	Average of guest review rating on listing location	Numerical
review_scores_value	Average of guest review rating on experience worthiness	Numerical

Continued on next page

Table A.1 – continued from previous page

Variable Name	Description of Variable	Data Type
instant_bookable	Can a potential guest make an instant booking or a prior request needs to be made to host?	Boolean
host_no_days_active	The count of days a host has been active on the day data is scraped	Numerical
host_is_superhost	Is the host of the listing a superhost[0]?	Boolean
host_response_time	The response time on average of a host	Categorical: Within an hour, Within a few hours, Within a day, A few days or more;
host_response_rate	The rate at which a host replies to booking requests	Numerical
host_acceptance_rate	The rate a host accepts a booking when a request is made for non instant bookable listings	Numerical
host_has_profile_pic	Does the host have a profile picture?	Boolean
host_identity_verified	Did the host verify its identity?	Boolean
host_email	Did the host provide email address?	Boolean
calculated_host_listings_count	The number of listings a host runs	Numerical
beachfront	Is the listing at the beachfront?	Boolean
pool	Is a pool available at the listing?	Boolean
wifi	Is wifi available at the listing?	Boolean
kitchen	Is a kitchen available at the listing?	Boolean
AC	Is an Air Conditioning available at the listing?	Boolean
washer	Is a washer available at the listing?	Boolean
dryer	Is a dryer available at the listing?	Boolean

Continued on next page

[0] A superhost is a badge given by Airbnb after meeting a set of criteria based on guest rating, experience, reliability and responsiveness. Specific details on criteria can be found here: https://www.airbnb.com/help/article/829

Table A.1 – continued from previous page

Variable Name	Description of Variable	Data Type
heating	Is heating available at the listing?	Boolean
dedicated_workspace	Is a workspace available at the listing?	Boolean
TV	Is a TV available at the listing?	Boolean
hair_dryer	Is a hair dryer available at the listing?	Boolean
iron	Is a clothing iron available at the listing?	Boolean
hot_tub	Is a hot tub available at the listing?	Boolean
parking	Is parking available at the listing?	Boolean
ev_charger	Is an electric vehicle charger available at the listing?	Boolean
crib	Is a baby crib available at the listing?	Boolean
gym	Is a gym available at the listing?	Boolean
BBQ_Grill	Is a BBQ available at the listing?	Boolean
breakfast	Is a breakfast provided at the listing?	Boolean
indoor_fireplace	Is fireplace provided at the listing?	Boolean
waterfront	Is listing located at the waterfront?	Boolean
safety	Is a safety alarm or carbon monoxide alarm or both available at the listing?	Boolean
Price	The nightly price in Euro as advertised on the Airbnb site when the data is scraped	Numerical

Table A.2: Correlation of features with Price in Crete and Malta Data

Variable Name	Crete	Malta
Number_of_Baths	0.53	0.43
bedrooms	0.53	0.53
accommodates	0.54	0.54
onehotencoder_x1_villa	0.50	0.27

Continued on next page

Table A.2 – continued from previous page

Variable Name	Crete	Malta
pool	0.47	0.38
beds	0.42	0.37
calculated_host_listings_count	0.28	0.19
BBQ_Grill	0.26	0.19
hot_tub	0.24	0.11
indoor_fireplace	0.23	0.17
gym	0.19	0.07
crib	0.18	0.27
washer	0.17	0.06
TV	0.16	0.24
walkable_distance_historical	0.12	-0.02
iron	0.11	0.06
AC	0.11	0.29
breakfast	0.10	-0.02
latitude	0.10	0.14
host_identity_verified	0.09	-0.05
host_email	0.098	0.05
onehotencoder_x1_hotel	0.08	0.07
dryer	0.08	0.09
waterfront	0.08	0.05
walkable_distance_capital	0.07	-0.03
hair_dryer	0.07	0.11
safety	0.06	0.08
host_is_superhost	0.05	-0.03
wifi	0.05	0.02
dedicated_workspace	0.05	-0.00
heating	0.04	0.14
host_no_days_active	0.04	0.02
onehotencoder_x2_Hotel room	0.04	0.006
onehotencoder_x0_within a day	0.04	0.07
minimum_nights	0.03	0.00
onehotencoder_x0_within a few hours	0.03	0.05
ev_charger	0.01	0.003
host_has_profile_pic	0.01	-0.02
beachfront	0.01	0.04
maximum_nights	0.01	0.06
onehotencoder_x2_Private room	0.01	-0.37
instant_bookable	-0.00	0.04
kitchen	-0.011	0.04
onehotencoder_x0_a few days or more	-0.01	0.01
onehotencoder_x1 _bed_and_breakfast	-0.01	-0.03
walkable_distance_beach	-0.01	-0.04
walkable_distance_bus_stop	-0.01	0.06
onehotencoder_x2_Entire home/apt	-0.02	0.44
onehotencoder_x1_maisonette	-0.02	-0.02

Continued on next page

Table A.2 – continued from previous page

Variable Name	Crete	Malta
host_response_rate	-0.02	-0.05
host_acceptance_rate	-0.02	-0.01
bathroom_shared	-0.03	-0.45
review_scores_cleanliness	-0.04	-0.09
onehotencoder_x0_within an hour	-0.04	-0.09
onehotencoder_x1_guesthouse	-0.05	-0.01
onehotencoder_x2_Shared room	-0.06	-0.29
review_scores_rating	-0.07	-0.12
availability_60	-0.07	0.16
availability_30	-0.07	0.17
longitude	-0.08	-0.12
availability_90	-0.08	0.16
availability_365	-0.08	0.12
review_scores_value	-0.08	-0.14
onehotencoder_x1_other	-0.09	-0.13
review_scores_accuracy	-0.09	-0.12
review_scores_checkin	-0.09	-0.12
onehotencoder_x1_house	-0.09	-0.08
review_scores_location	-0.10	-0.09
review_scores_communication	-0.12	-0.13
number_of_reviews_ltm	-0.17	-0.15
onehotencoder_x1_apartment	-0.37	-0.05

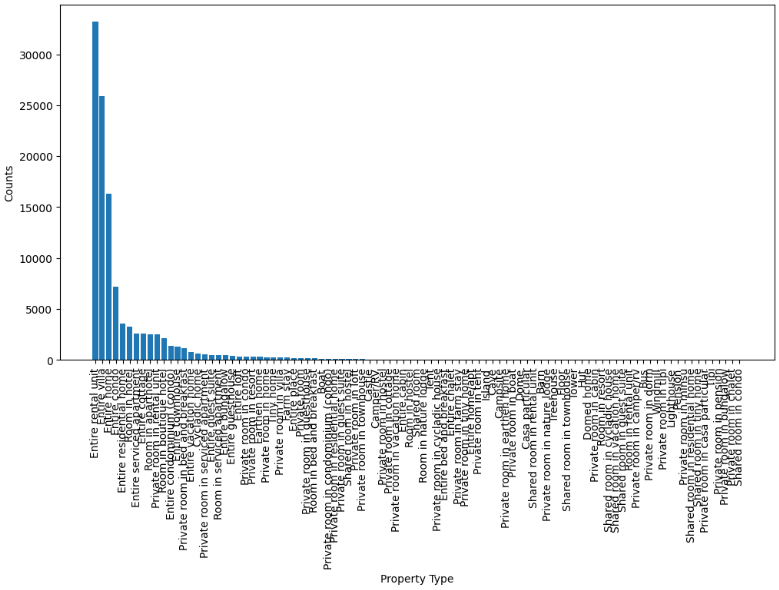


Figure A.1: Count of listings by property type in raw data set for Crete

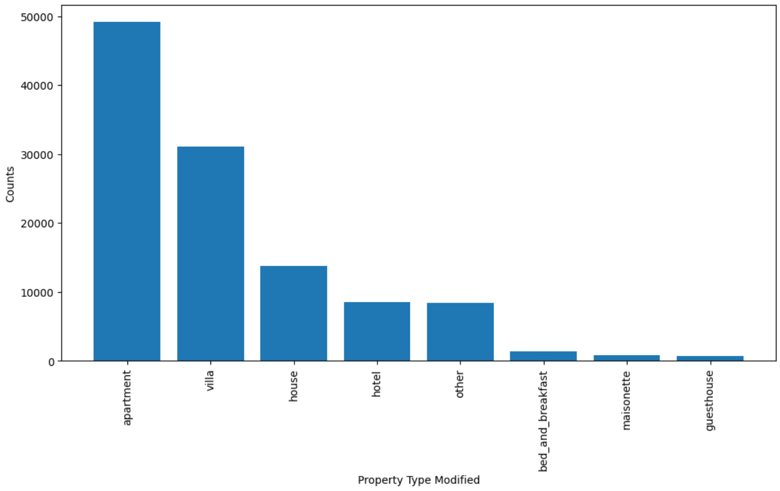


Figure A.2: Count of listings by property type in modified data set for Crete

Appendix B

Further Modelling Details

Model	Python Package	Module
KNN	scikit learn	KNeighborsRegressor
Linear Regression (No Regularization)	scikitlearn	linear_model
Ridge Regression	scikitlearn	linear_model
Lasso Regression	scikitlearn	linear_model
Random Forest Regression	scikitlearn	RandomForestRegressor
XGBoost	scikitlearn	XGBRegressor
Catboost	catboost	CatBoostRegressor
Neural Networks	tensorflow	keras

Table B.1: Python packages used for Modelling

Parameter	Grid Search Range
Ridge Regression Regularization Parameter	[0.0001,10] stepsize:0.1
Lasso Regression Regularization Parameter	[0.0001,10] stepsize:0.1

Table B.2: Grid of parameters for Grid Search for Linear Regression Models

Parameter	Description	Range of Parameters (Random Search)	Range of Parameters (Grid Search) without Geo data	Range of Parameters (Grid Search) with Geo data
n_estimators	The maximum number of trees in the forest	50 random samples from the range [100,1000]	[805,820]:stepsize 1	[980,1000]:stepsize 1
max_depth	The maximum depth of a base model tree	[5,10,20,30,None]	None	None
max_features	The number of features sampled at each split	['sqrt','log2',total_count/3]	total_count/3=28	total_count=29
min_samples_leaf	The smallest number of observations for a further split	[2,5,10,15]	[2,3,4]	[2,3,4]

Table B.3: Random Forest tuned parameters

Parameter	Description	Range of Parameters (Random Search)	Range of Parameters (Grid Search) without Geo data	Range of Parameters (Grid Search) with Geo data
n_estimators	Count of boosted trees	50 random samples from the range [100,1000]	[680,681,682,683,684]	[965,967,968,969,970]
max_depth	The maximum depth of base models	[2,6,10,None]	[5,6,7]	[5,6,7]
learning_rate	Weights step size shrinkage	[0.1,1]:stepsize:0.1	[0.4,0.5]:stepsize:0.01	[0.48,0.52]:stepsize:0.01
reg_alpha	L1 penalization term on weights	[0,0.5,1]	1	1
reg_lambda	L2 penalization term on weights	[0,0.5,1]	0	0
objective	Problem specification and objective function of learning	reg:squarederror	reg:squarederror	reg:squarederror
subsample	fraction of observations that are randomly sampled prior training per tree	[0.1,1]: stepsize:0.1	[0.85,0.9]: stepsize:0.01	[0.85,0.9]: stepsize:0.01
colsample_bytree	fraction of features that are randomly sampled per tree	[0.1,1]: stepsize:0.1	[0.1,0.15]:stepsize:0.01	[0.1,0.15]:stepsize:0.01

Table B.4: XGBoost tuned parameters

Parameter	Description	Range of Parameters (Random Search)	Range of Parameters (Grid Search) without Geo data	Range of Parameters (Grid Search) with Geo data
iterations	The maximum number of boosted trees that can be sequentially constructed	50 random samples from the range [100,1000]	[621,622,623,624,625]	[948,949,950,951,951]
learning_rate	Controls the size of the step in the optimization process	[0.1,1]: stepsize 0.1	[0.01,0.1] stepsize:0.01	[0.01,0.1] stepsize:0.01
l2_leaf_reg	Coefficient of L2 regularization applied to the loss function	[1,3,5,7,9]	[4,5,6]	[2,3,4]
depth	The maximum depth of weak learners	[2,10]:stepsize 1	9	9
random_strength	Adds randomness to the score value of potential splits	[1,8]:stepsize 1	[2,3,4]	[2,3,4]

Table B.5: Catboost tuned parameters

Parameter	Description	Range of Parameters (Random Search)	Range of Parameters (Grid Search) without Geo Data	Range of Parameters (Grid Search) with Geo Data
Activation Function (for hidden layer)	Non-Linear Function applied to the first hidden layer	ReLU, tanh	ReLU	ReLU
learning_rate	Controls the size of the step in the optimization process	30 equally spaced values in the range [0.0001,0.3]	[0.0001,0.1] stepsize:0.01	[0.01,0.03] stepsize:0.001
dropout_rate	The proportion of neurons to be dropped out in the hidden layer at each training iteration	20 equally spaced values in the range [0.01-0.5]	[0.05,0.15] stepsize:0.01	[0.3,0.4] stepsize:0.01
batch_size	The number of observations in each batch where the number of batches is equivalent to the number of gradient updates	[64,128,256]	64	128
Epochs	Count of instances when the neural network is trained on the entire training data set	[50,100,150,200,250]	200	150

Table B.6: Neural Network tuned parameters for Malta Data

Parameter	Range of Parameters (Random Search)	Range of Parameters (Grid Search)
Activation Function (for hidden layer)	ReLU, tanh	ReLU
learning_rate	30 equally spaced values between [0.0001,0.3]	[0.0001,0.1]:stepsize 0.01
dropout_rate	20 equally spaced values in the range [0.01,0.5]	[0.1,0.2]:stepsize 0.02
batch_size	[64,128,256,512]	64
Epochs	[100,500]:stepsize 50	350

Table B.7: Neural Network tuned parameters for Crete Data