Istanbul: Price Predictions

Business Report

INTRODUCTION TO THE PROJECT

This prediction analysis aims to predict prices for certain Airbnb listings in Istanbul, Turkey & can accommodate 2 to 6 persons. Furthermore, the analysis, uses various constraints and/or filters to incorporate only certain types of properties. This allows for a more comprehensive data input to the various prediction models. After important data cleaning & transformation, the models run at efficiency and provide valuable insights. The aim is to predict the price which is the target variable depending on various predictors. The predictors impact the target variable with different intensity. For a predictive analysis, I run 4 predictive models which include: OLS linear regression, LASSO, Random Forest & CART. However, before running the prediction models, this report aims to shed some light on the data cleaning and transformation of certain variables before running the machine learning predictive models.

DATA SET OVERVIEW

The data used is based on a specific date which is *30th December 2022* & can be found on the official website (URL at end of report). The specific date is an important aspect of the analysis as it constraints the time dimension of the analysis. The choice of *30th December 2022* is in mainly since it is the most recently available data. When the raw data is imported, there are 36,717 observations for a total of 75 variables. This project must ideally include 10,000 plus observations. The 36,000 plus observations seem to be too high for the analysis. However, after various cleaning and transformations, the observations reach close to the required 10,000 plus observations. It is important to note that each observation represents a single rental property listed at Airbnb.

DATA WRANGLING

One of the most critical parts of this predictive analysis is data cleaning by removing errors, renaming columns, creating new variables, dropping irrelevant variables, and creating a data set which is more accessible and efficient for the predictive models. In this regard, I begin skimming the data set. I then proceed to limit by analysis by accommodates from 2 to 6 people. Furthermore, the most important variable for my analysis is the price per day which is the target variable. I remove the dollar signs and convert it to numeric type. Once the target variable is cleaned, I proceed to clean the predictor variables. These variables will be used to predict the value of the target variable. I drop columns that are not relevant for the analysis such as listing URL, scraping ID and many others. I then drop observations that have no value for target variable, price. Furthermore, I impute values for certain potential predictors including number of beds, accommodates and reviews. I then proceed to convert property types and neighborhood to factor variables & filter and keep only apartments and rental units as property types. Lastly, I use a function to extract the amenities and convert certain amenities as binary variables along with creating these amenities as dummy variables. Once, the data is cleaned, I have 16,187 observations and 32 variables. The predictive models will have this clean data set as input data frame.

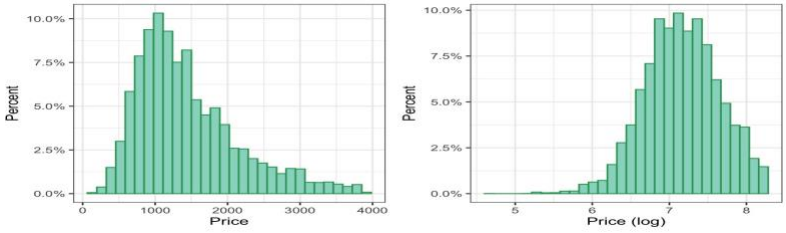
EXPLORATORY DATA ANALYSIS

After completing data wrangling, my aim is to conduct exploratory data analysis. In other words, I want to know my data better. My goal is to understand the descriptive summary statistics of my data by visualizing its distribution. Apart from their distribution, I aim to identify the relationship between my predictor variables and target variable, price. The important distinction to make is between the target variable: Price & the Predictors: accommodates, amenities and property type. The EDA process is divided into two parts. First, label engineering and then feature engineering. These two steps define my predictions and impact the predictive models.

LABEL ENGINEERING

The target variable price is Turkish lira is the variable of interest in this analysis. Before I predict, I want to understand the distribution of this integral variable. It is imperative to note that the price is kept in Turkish Lira for better understanding for the marketing team here in Istanbul (1 USD is approx. 19 Turkish Lira).

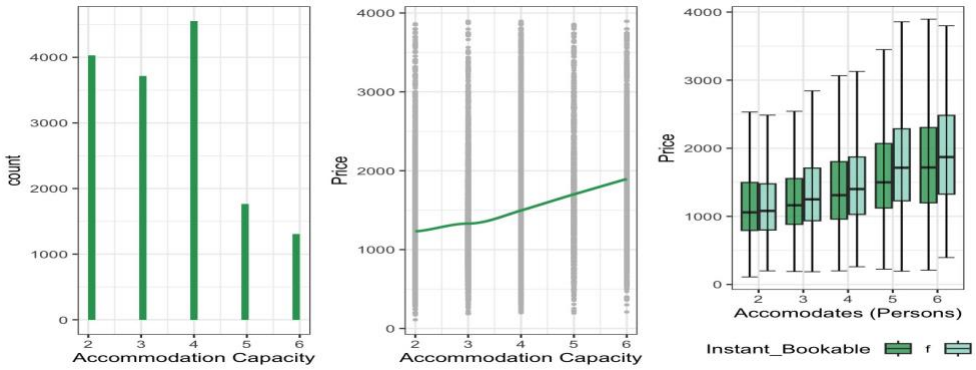
After narrowing the scope of analysis, the mean is 2,535 Lira (USD 133.42) and the median is 1,316 Lira (USD 69.26). As seen below, the target variable price has a relatively longer right tail. This is primarily because the mean is larger than the median. Hence the price is slightly skewed. I conduct a log transformation & visualize its distribution which shows that it results in a slightly left side tail rather than normal distribution. Therefore, since log is not achieving a near perfect normal distribution, I decide to keep price as the main target variable rather than the transformation or log of price. Price seems relatively normal to a great extent keeping in mind the inclusivity represented by price variable.



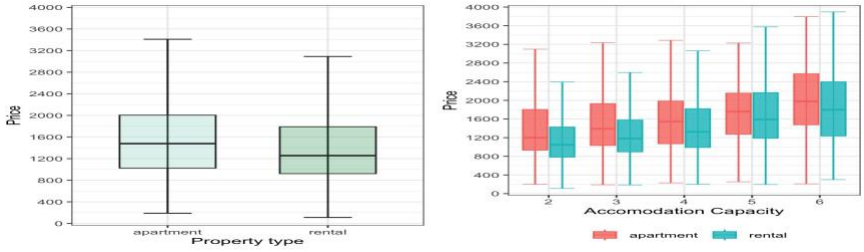
FEATURE ENGINEERING

Feature engineering is considered one of the most important considerations for a data analyst while conducting prediction models. Hence, great emphasis is placed on descriptive statistics, distributions, and interactivity of predictor variables. To begin with, accommodation capacity or accommodates, is considered. The data is restricted from 2 to 6 accommodates and its distribution can be viewed in the histogram below. As seen, maximum accommodations are where accommodates is 2 people. This shows that the most common listing at Airbnb for Istanbul have an accommodation capacity for only two people. Then moving on to larger values, the count significantly drops for accommodations with 3 people. A significant proportion of accommodation listings have the accommodation capacity of four persons. Lastly, very few listings have 5-6 people as accommodates.

I then visualize the distribution of accommodates and price. A clear trend can be identified, that as accommodation capacity increases, the price distribution will increase also. However, to be more specific, a significant increase is visualized when a property has 4 accommodates instead of 3. For accommodation capacity, I also consider whether the property is instantly bookable or not. This is a binary variable and is color coded as seen below. From 2 to 6 accommodates, the box plots reveal that as accommodates increase, the number of listings get more costly on average.



Another crucial predictor is the accommodation type and its distribution as well as impact on price. For this purpose, a box plot can be visualized as seen below. Apartments listings are relatively higher in price compared to rental units. This may be attributed to a multitude of factors for example apartments generally include those with higher accommodation capacity or great amenities which lead to a higher price distribution for apartments. A box plot is also shown on the bottom right with accommodation capacity and type incorporated. The box plots show that for all accommodation capacities the apartments are higher in price compared to rental units. A minor exception is at accommodation capacity of 5 persons, where apartments or rentals are similarly priced.



A key predictor in my predictive modelling is the amenities and their impact on price. For this I created dummy variables for amenities as part of my data cleaning. Below, I create 4 visualizations that show each a specific amenity and its relationship with accommodation type and mean price. The key amenities visualized below include facility of coffee maker, air conditioning, gym and whether the listing is baby friendly or not. The visualizations show that on average, for these 4 amenities, the mean price is higher in rentals as compared to apartments when the facility or amenity is provided labeled by the binary variable 1.

As part of Exploratory Data analysis, including label and feature engineering, I had the opportunity to know my data better and perform relevant data transformations.

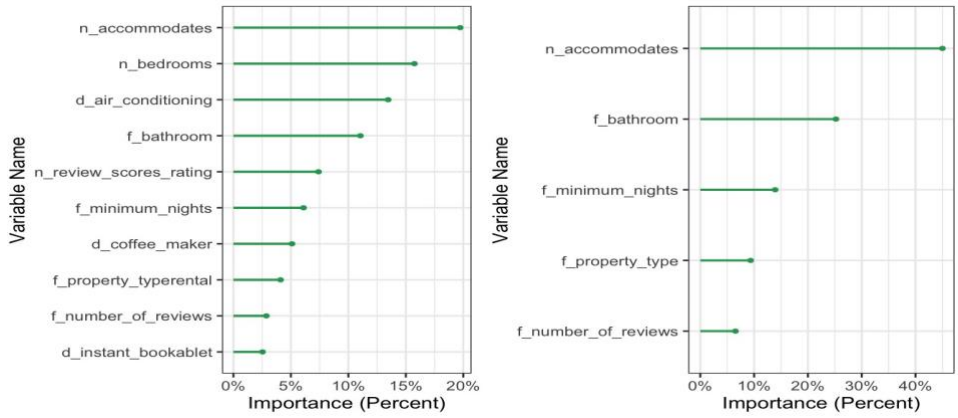
PREDICTION Model Comparison

To predict I run 4 different prediction models as outlined in the introduction. These include, first, the Ordinary Least Squared or OLS model which contains specific models. I then proceed to run LASSO as my second model. My third prediction model is the Random Forest model which is significantly more complex compared to OLS and LASSO. Finally, to conclude, I run a Classification & Regression tree also known as CART. RMSE for the models are summarized as follows:

	CV RMSE	Holdout RMSE
OLS Model 2	640.19	632.69
LASSO (model with interactions)	639.73	632.29
CART	666.18	663.30
Random forest (with amenities)	635.12	627.97

PREDICTOR VARIABLES AND THEIR IMPORTANCE (RF)

Various predictors in RF prove to have different importance in terms of their impact on the target variable. As shown in the bottom right graph, number of accommodates has a significantly higher impact compared to other predictors such as number of reviews or property type. As seen on the bottom left graph, number of accommodates and beds is significantly higher in importance even according to amenities provided such as coffee maker, baby friendly accommodation or even air conditioning.



CONCLUSION

As a business analyst, to predict the price for an accommodation with 2-6 persons in Istanbul, I recommend the Random Forest Model as the best predictor. This is primarily because it has the least RMSE amongst the four models. Comparing the four, CART is not recommended due to an exceptionally high RMSE. However, amongst the simpler models, OLS is simplest and must be taken into consideration. The choice of model also depends on the needs of the marketing team. In case a simple model is required, the OLS based on 20 predictors only is a feasible model also. However, this depends on the level of accuracy the marketing team needs. Hence, if an error of 640 Liras (USD 33,68) is acceptable, OLS is recommended. However, the Random Forest method will be most accurate with an error of only 635 Liras (USD 33.4) .

This prediction analysis was performed with a lot of consideration and in case the marketing team needs to understand more about the models, a technical report and about the data can be viewed using the below URLs.

My Github Repository

Click here for Airbnb Data