Data Analysis 3: Assignment 2: Business Report

Mohammad Ahmed

Introduction

The goal of this report is to provide a detailed description for the price prediction model to help a company in setting a price for their new apartments that are yet to be introduced to the public. Our objective is to derive a price for apartments that accommodate 2-6 guests and are small to mid-sized in San Deigo, United States. The dataset is obtained from Inside Airbnb. The first step in carrying out a decent analysis is to clean the dirty dataset, preparing variables that will be used and to understand the domain. We will be using various prediction models and analyze which one helps us the most and what model provides the best results. The major predictor variables include number of accommodates, total number of beds in the accommodation, number of bathrooms, the neighborhood and various amenities present in the data. Eventually, we will be providing a conclusion on what model gives the best prediction measured by relative RMSE values.

Data Preparation

Before initializing building predictive models, it is important to focus on cleaning the data and preparing it for our analysis. The original data set in the Airbnb San Diego consists of 12871 observations and 75 columns. This data refers to rental prices per night on 23rd March 2023. Our target variable is price in USD. The transformation from original data to a clean dataset consists of processing the variables like amenities, dropping columns that are not required for our analysis and seem irrelevant to our models. Creating factor and binary variables and dealing with the missing values. Detailed data cleaning steps can be obtained if required as well. We also filtered the price variable to be below 500 because anything above 500 USD per night seems like extreme values our objective is for small to mid-sized apartment so those were dropped.

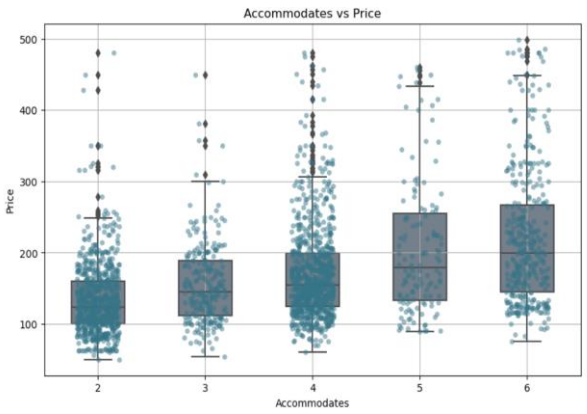
Feature Engineering

Feature engineering includes what type of predictor variables we want to include, what functional forms of predictors we will have and looking at the possible interactions between different variables to give our analysis a good depth. We have basic variables that include predictors like accommodates, type of property, number of beds, etc. Alongside the basic variables, we also include factorized variables like neighborhood groups etc. We also used the amenities dummies that we created above which include multiple variables. After creating the required variables, we dealt with missing values. Dropping these missing values isn't a suitable option so they are imputed. With beds, we decided to impute a few missing values by linking it to the number of accommodates and using half the number of accommodates to replace the value of beds assuming that there are double beds in the listing. One assumption was that every listing will have at least one bathroom.
As the goal of project is to build price prediction model, it is crucial to check price and log of price distribution.

Price distribution shows Airbnb apartment prices is skewed with a long right tail and the log price is close to normally distributed. In this project, log of price is not considered, prediction is carried out with price for all the models.

Exploratory Data Analysis

Some of the important predictor variables are related to the size: for example, the total number of guests an apartment can accommodate and number of beds, number of bathrooms. The below box plots show the average price per number of guests. On the other hand, apartments with many guests have high average prices (referring to figures on next page)



Modelling

The best model gives the best prediction in the live data. Before turning to the modelling part of the project, it is worth mentioning that to avoid over fitting, the original data is split into two random parts by 20% to 80% ratio. Holdout set contains the 20% and the rest is work data set. In addition, the k-fold cross-validation is a good way to find a model which gives the best prediction for the original data. For this project 5-fold cross validation is used. This means splitting the data into five random samples and calculating and deciding based on the average of 5 CV RMSE result.
Further details are provided in the technical report. Our model consists of basic variables, basic additional variables, reviews variables, amenities and interactions.

Models

It is important to run and evaluate different models for a given data set. In order to predict apartment prices, the following models and algorithms were used. Naming them as following:

OLS | GBM | Random Forest | CART | LASSO

	model	CV RMSE
0	OLS	60.733168
1	LASSO	62.167712
2	CART	76.412485
3	random forest	56.880000
4	GBM	59.253066

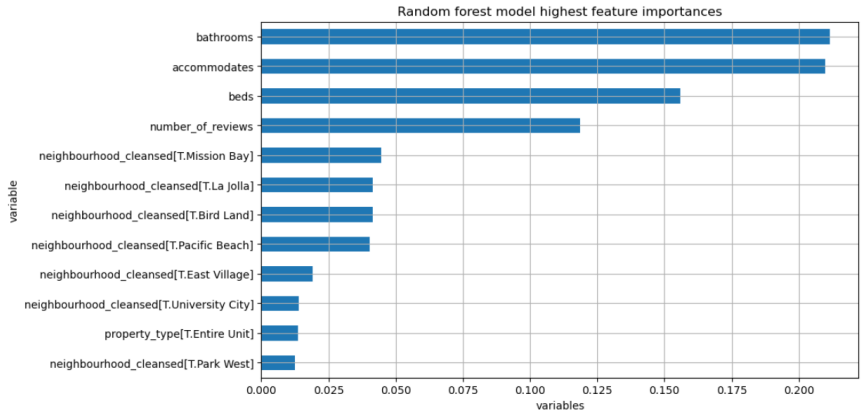
As it can be seen through the chart above, Random Forest gave us the best CV RMSE, a value of 56.88. This shows that our best performing model is Random Forest. However, we can also see that GBM is not too far behind with a CV RMSE value of 59.25. The difference between both the models is barely 2.37 therefore GBM also turns out to be a good model for us in order to forecast the prices for new Airbnb listings in San Diego.

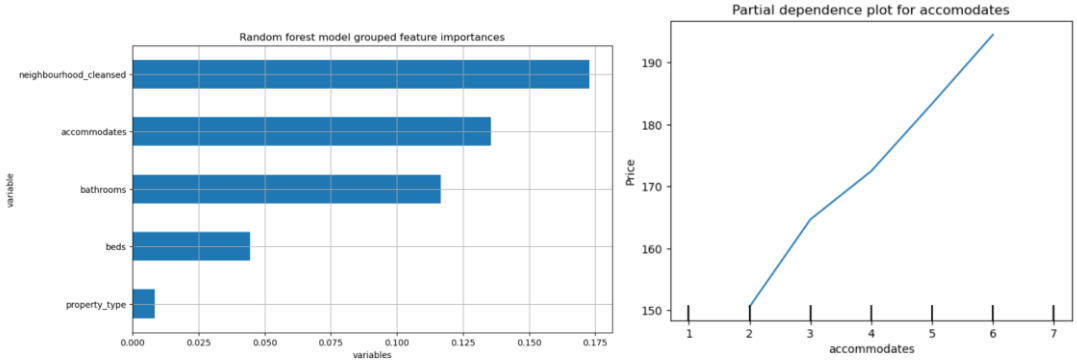
Diagnostics

Our best model turns out to be Random Forest that is an example of ensemble method which uses the results of many predictive models and combines those results to generate a final prediction. We used basic variables, basic additions, reviews and dummies for this model. The RMSE comes out to be 62.7003 in the training set and then we obtained a better value of 56.88 in the holdout set which turns out to be the best when compared to all other models.

Variable Importance Plots

Variable importance plot for the most important variables shows that number of bathrooms, number of accommodates, number of beds and San Diego neighbourhoods are the most important along with others demonstrated in the plots below. The grouped variable importance shows that neighborhood, number of accommodates, bathrooms, beds and property type are the most important variables.





Partial Dependence Plot shows how average y differs for different values of x conditional on all other predictor variables. It is based on predictors for the holdout set. Partial dependence plot for number of accommodates and price shows that price increases as the number of accommodates increase.

Conclusion

The goal of this report was to find a better model to predict Airbnb prices in San Diego for small to mid-size apartments. Five models were illustrated to compare across model performance. Random Forest resulted to be the best model by 56.88 USD RMSE. The second-best model was basic GBM which gives an RMSE value of 59.25 and highlights meaningful characteristics about the nature of Airbnb apartments in San Diego. Key price drivers based on post prediction diagnostics are the number of bathrooms, number of accommodates, number of beds and the neighborhood.