Machine Learning Engineer Nanodegree

Capstone Proposal - Udacity

Thiago T. S. Gavioli

November 2nd, 2021

Predicting prices for Airbnb in Rio de Janeiro - Brazil

1 - Domain Background

Booking an entire apartment, home, etc. during a trip is becoming increasingly common over time. Rather than book a room in a hotel, people have preferred entire places, whether by privacy or saving some money. In such context, Airbnb is nowadays one of the most important service for booking apartments, home, or any kind of place for staying during a trip or even a place to rent for living.

Whatever the reason, one important factor in this kind of transaction is the price. A host looks for aiming the best price which gives a good financial return and keeps the place booked most of the time, and the guests look for the best price considering the benefits.

That said, having a way of predicting the prices according to historical data is a valuable tool. For this purpose, evaluating the best options is the main goal of this project.

My personal motivation for this is because I am a heavy user of Airbnb service.

2 - Problem Statement

Having some information about historical prices and features available, this project aims to evaluate the best ways of predicting prices for listings in Airbnb service located in Rio de Janeiro, one of the most important cities from Brazil. This way, we would be able to help either hosts or guests to optimize their decisions when it comes for price.

3 - Datasets and Inputs

For this project we are going to use two datasets provided by Airbnb [1] with information from Rio de Janeiro.

The datasets are:

- calendar.csv
- listings.csv

Calendar contains mainly information about price and if the listing is available or not in a specific date. It is possible to relate this dataset to listings since it includes the listing ids.

It has 8.545.976 of rows and 7 columns. The date ranges from 2021-09-28 to 2022-09-29.

Listings contains the detailed information about each place available in Airbnb. Details about the host, price, neighborhood, review, amenities, etc., can be found. Here are the features which will be used to predict prices for each place/listing.

The original dataset has 23.414 rows and 74 columns (features + price). The range of date is the same as presented in calendar dataset.

The raw dataset contains different types of data (int., float, object) which must be handled before being analyzed and used for modelling. Missing data, outliers and other problems will properly handle in the context of the problem.

4 - Solution Statement

The proposed solution to this problem is applying machine learning regressor algorithms to predict price according to selected features.

First, we are going to handle the raw features in order to find the best inputs to each model evaluated. We will clean the dataset according to the necessity, analyze all the information we have through an exploratory data analysis, select the best features to fit the models and evaluate which model performs best.

The evaluation metrics mentioned later in this proposal will be used to compare the models with the benchmark and between each of themselves.

5 - Benchmark Model

For the benchmark we are going to use a simple linear regression model from scikit-learn library.

6 - Evaluation Metrics

The evaluation metric for the project is the R2 score, which tells us how well a prediction approximates the real data points [2].

7 - Project Design

Data preprocessing

At first we are going to work on our datasets to cleaning what is necessary and identify different data types to preprocess in order to make them uniform as demanded.

- Remove outliers
- Dealing with missing data
- Change data type whenever necessary

- Drop unnecessary columns
- Get dummy variables from categorical features
- Filter data in order to achieve better results in our models

Exploratory Data Analysis

Following on the project, we are going to check visually (through some charts) the distribution of our features and the target variable. The relationship between them will also be analyzed.

Data Splitting

Split the data into training and validation set with a 70/30 split.

Model Training and evaluation

I will start with a simple linear regression model, then trying different architectures, models and hyper-parameters to reach the best performance according to the metric picked out.

8 – References

[1] Datasets provided by Airbnb,

http://insideairbnb.com/get-the-data.html

[2] Interpretation of R2,

https://en.wikipedia.org/wiki/Coefficient_of_determination#:~:text=R2%20is%20a%20statistic,predictions%20perfectly%20fit%20the%20data.
