Airbnb Price prediction Case Study

Yaswanth seela

Contents

- Introduction
- Problem and Objective
- Data Pre-Processing
- Data Exploration and key Insights
- Model Building
- Conclusion
- Recommendations
- Potential directions for future work

Introduction

- Airbnb is a home-sharing platform that allows home-owners and renters ('hosts') to put their properties ('listings') online, so that guests can pay to stay in them. Hosts are expected to set their own prices for their listings.

Problem and Objective

- The business problem that we are dealing with is to get insights of what variables impact the house prices and be able to predict the house prices as accurately as possible, which leads to drive many business decisions.
- So the objective is to build a machine learning model that predicts the house prices

Data Pre-processing

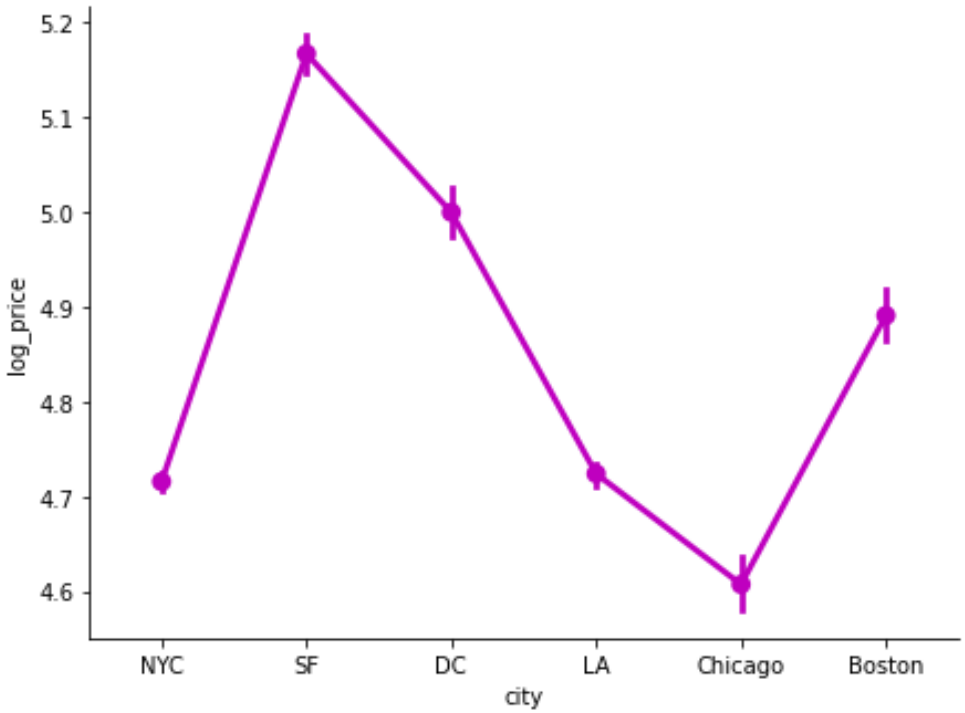
- A machine learning model could only process numeric data. It cannot understand text data.
- So In this phase the data is cleaned and converted to a suitable format that a machine learning model could be fed in with
- A new feature is created using the latitude and longitude feature.

Data Pre-Processing

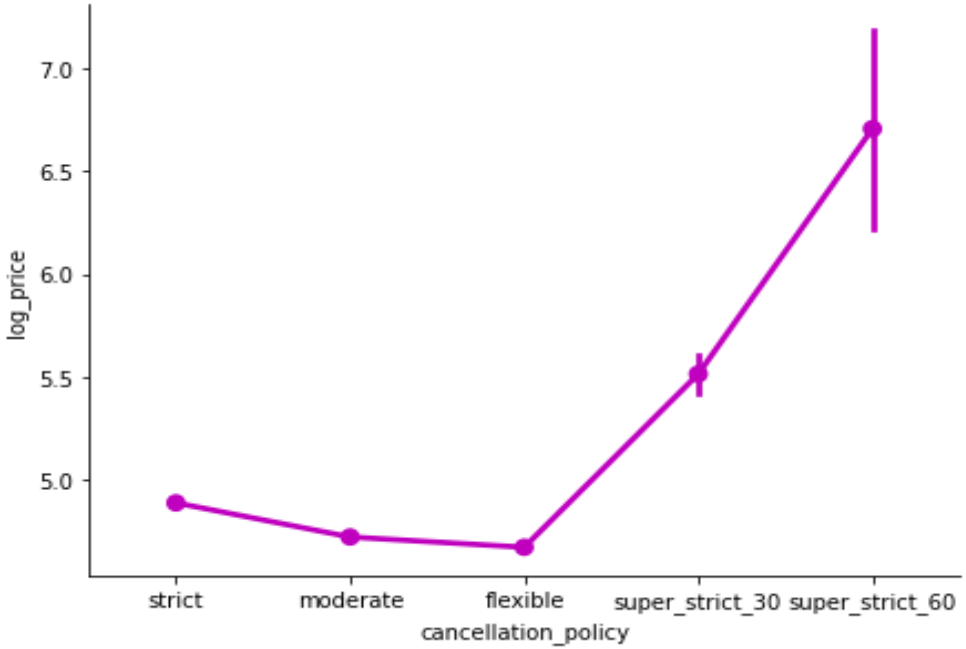
- NLP is not used in the creation of an initial model . Therefore, free text columns are dropped for now, as well as other columns which are not useful for predicting price (e.g. url , host name and other host-related features that are unrelated to the property)

Data Exploration and key insights

- we see that the highest mean log_price is for San-Fransisco whereas lowest is for Chicago

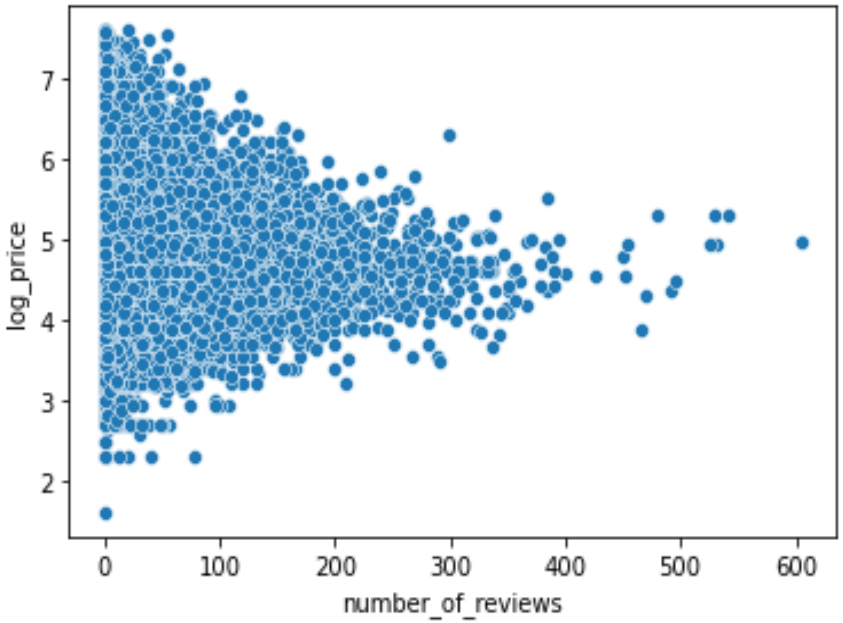


Data Exploration and key insights



- We observe that flexible cancellation policy has lower mean log _price and cancellation policy which is super strict has highest mean log_price.

Data Exploration and key insights



We can infer that the houses with higher reviews are centered around the mean of log_price. Also the houses with less number of reviews have all the ranges of log_price

Data Exploration and key insights

- Amenities feature has huge number of categories. Hence this is not included in model building process.
- A list of majority of the amenities is created using data cleaning techniques

Model Building

- Xgboost model, being a superior performer among most of the machine learning algorithms, is built to get the best possible performance. And many models are further built to see whether their performance can surpass the Xgboost's performance.
- In this process a simple feed forward neural network is built and on examining it's validation performance its is found out that it was overfitting.
- To avoid this neural networks with regularization techniques are implemented.

Conclusion

- Although feed forward neural networks are powerful algorithms it is quiet surprising to notice that Xgboost surpassed feed forward neural network.
- However, even in the best performing model, the model was only able to explain 67% of the variation in price. The remaining 33% is probably made up of features that were not present in the data or due to text data.

Recommendations

- It is most likely that a significant proportion of this unexplained variance is due to variations in the listing photos or blurred images. The photos of properties on Airbnb are very important in encouraging guests to book, and so can also be expected to have a significant impact on price - better photos (primarily better quality properties and furnishings, but also better quality photography) equal higher prices.

Potential directions for future work

- Augment the model with natural language processing (NLP) of listing reviews, e.g. for sentiment analysis or looking for keywords
- Trying to get better quality images
- In addition to predicting base prices, a sequence model could be created to calculate daily rates using data on seasonality and occupancy
- Use geo-json data of the cities to find out where the property quantity is dense and if that effects the prediction.