AirBnb Price Prediction

Deepak Kumar

Agenda

- Introduction
- Problem Overview
- **Data Preprocessing**
- **Data Exploration**
- **Feature Engineering**
- **Modeling Approaches**
- **Results**
- Algorithm Comparisons
- Conclusion

Introduction

- This project analyzes Airbnb listings in the city of US to better understand how different attributes such as bedrooms, location, house type amongst others can be used to accurately predict the price of a new listing that is optimal in terms of the host’s profitability yet affordable to their guests.

- This model is intended to be helpful to the internal pricing tools that Airbnb provides to its hosts.

- Furthermore, additional analysis is performed to ascertain the likelihood of a listing’s availability for potential guests to consider while making a booking.

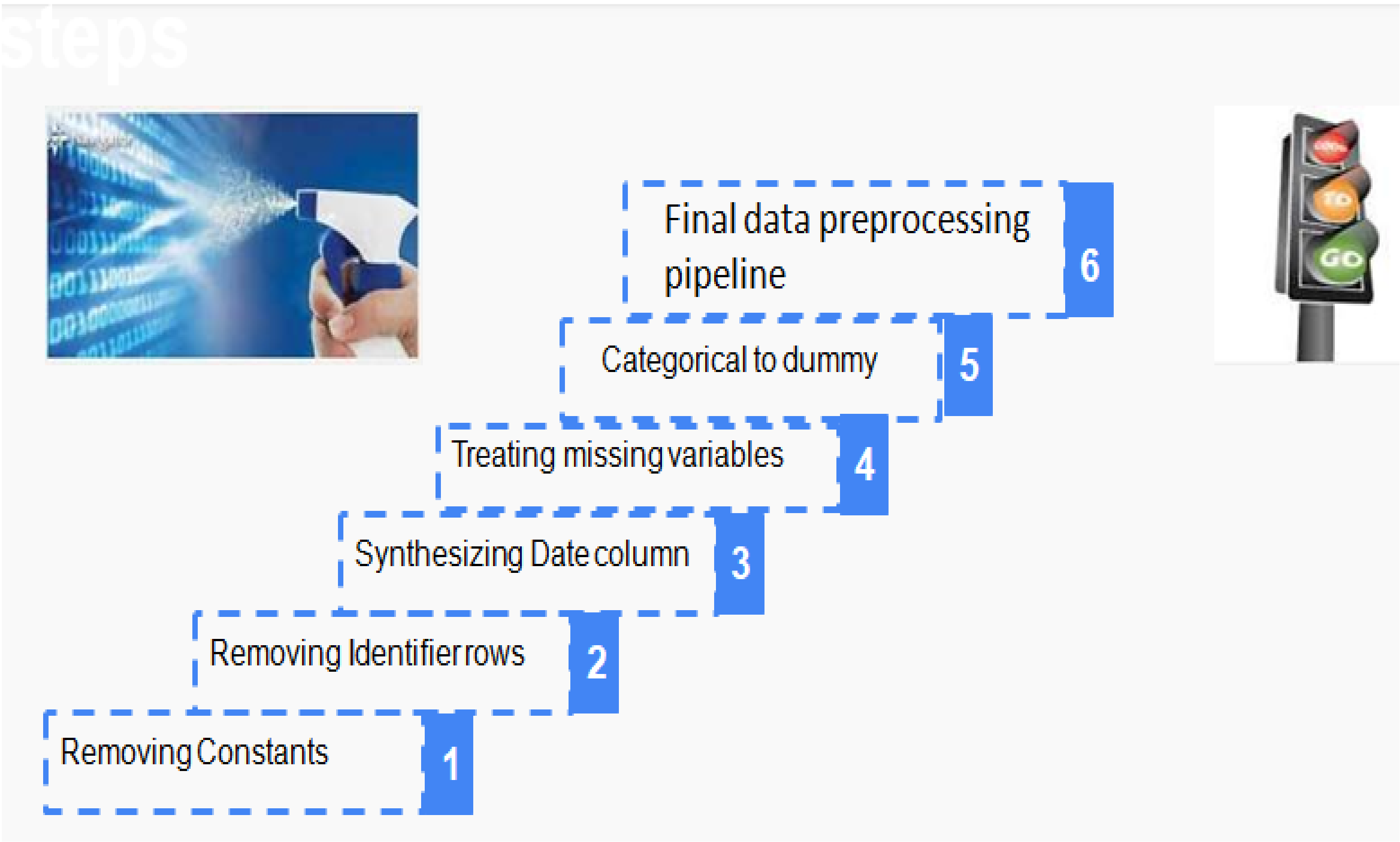
Problem Overview

- Given a dataset with 28 variables such as number of bedrooms and a log-price indicator (greater than 0) for each observation in the training data
- The objective is to suggest the log-price of a particular listing using the 28 features provided for the test observations.
- There are 49999 observations and 29 variables in the training dataset.
- There are 24111 observations and 28 features provided in the test dataset.
- Variable named 'log_price' is the dependent/Response variable

Hypothesis

- ❑ The hosts on Airbnb experiment and charge an optimal price. So, for a new listing can we analyze similar listings in the past to recommend an optimal the host should charge for the new listing?
- ❑ Goal: Hosts get a decent idea how much to charge for a new listing that has no reviews
- ❑ It is uncertain when a listing is hosted or when it becomes unavailable. Hosts/Guests change plans. So, having the information about availability in the past can we recommend guests with a likelihood of a listing being available?
- ❑ Goal: Guests can decide whether to move on or wait for a listing to become available depending on the likelihood.

Data Cleaning Steps



steps

1 Removing Constants

2 Removing Identifier rows

3 Synthesizing Date column

4 Treating missing variables

5 Categorical to dummy

6 Final data preprocessing pipeline

Data Preprocessing –Variable Creation

- Continuous features

  →We identified highly correlated continuous features and eliminated them from our input. We filled the null with 0 for price related features (security deposit and cleaning fee), and median value for the rest of the features. We then performed standardization transformation with normalized feature.

- Categorical features

  →For most of the categorical features, we directly performed one-hot encoding, while for a small fraction of list features, like amenities and host verifications, we encoded them into vectors via dictionary building and mapping

- Date features

  →Extracted date components and created cyclic features
  →Also created date difference factors

Method Of Analysis

❑ Importance of variable were extracted using Random Forest

❑ Compared performances and consistency between single Ridge and Lasso model

❑ Performance Comparison across multiple algorithms : Linear Model ,RF ,GBM , XGB

❑ Model Accuracy
  - RMSE

Method of Analysis – Multiple Linear Regression

❑ Iteration

- Select one variable at a time from the variable importance table created using random forest

❑ Significance

- Check significance of new variable along with existing variables by its t-value and probability of t- statistics.
- If RMSE is reduced, keep the variable, else drop it

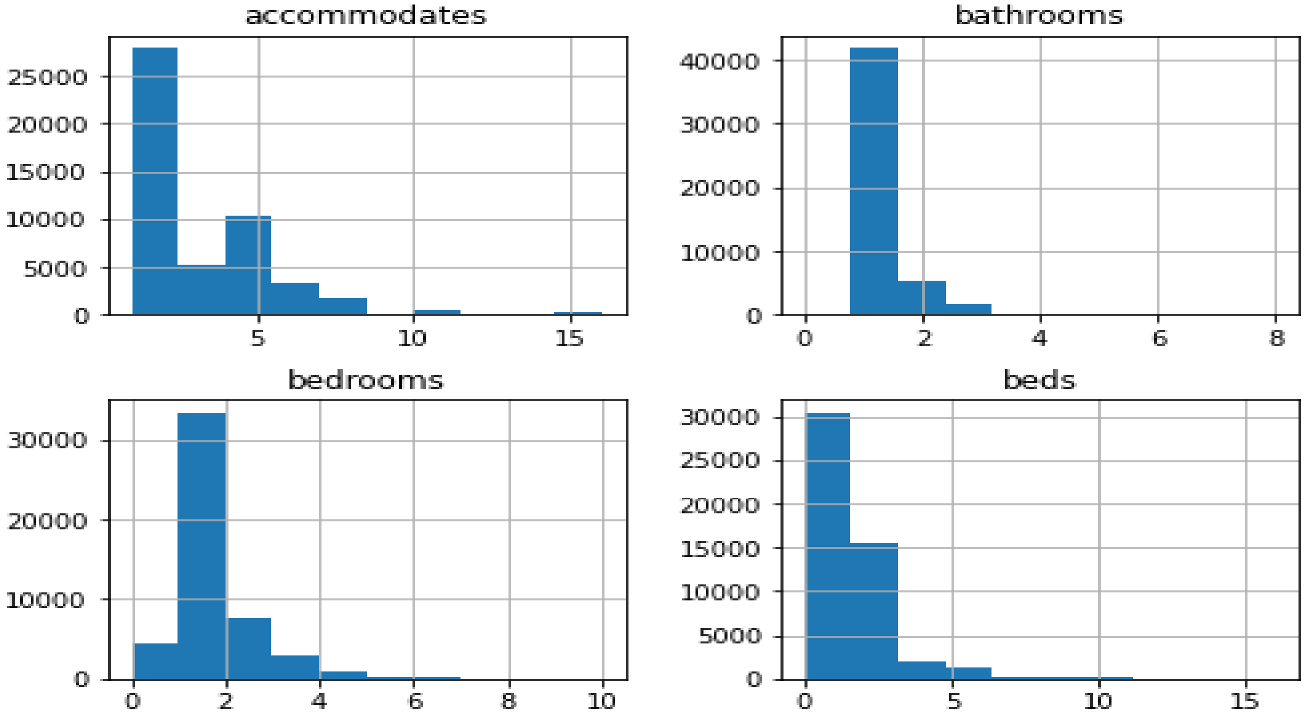
❑ Model Accuracy

- After adding the new variable, the model accuracy on train and test data using RMSE is checked.

.

Data Exploration

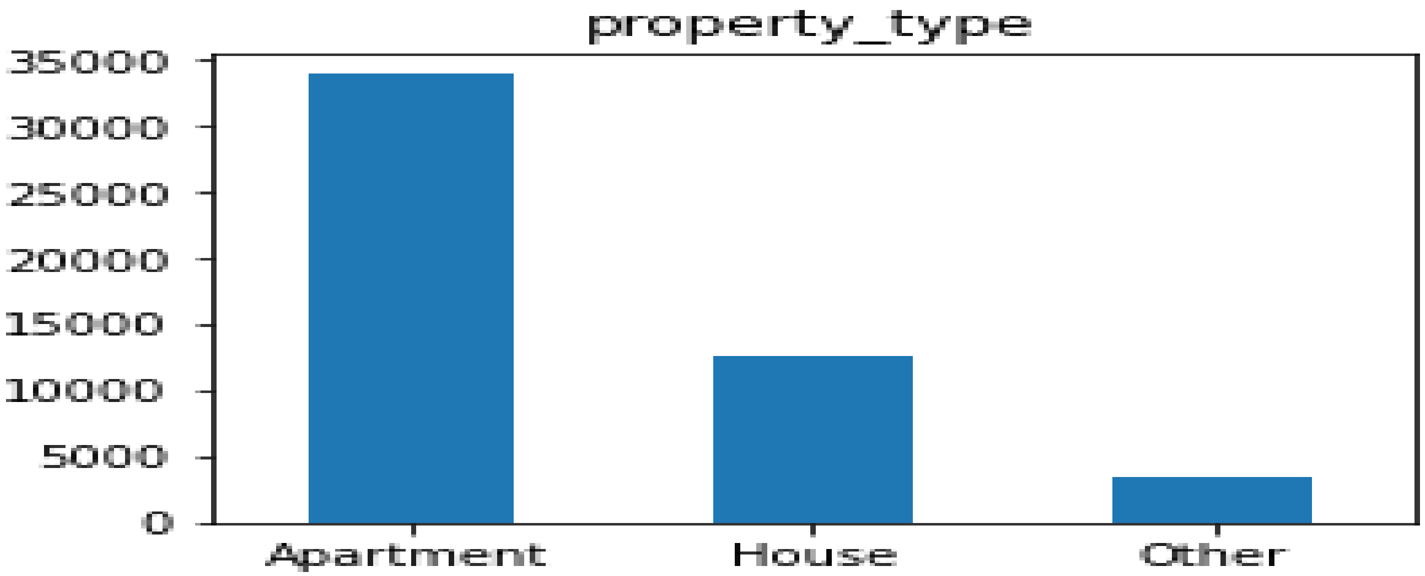
Distribution different variables:



- ❑ The most common property setup sleeps two people in one bed in one bedroom, with one bathroom.
- ❑ That accommodate more people achieve noticeably higher nightly rates, with diminishing returns coming after about 10 people.

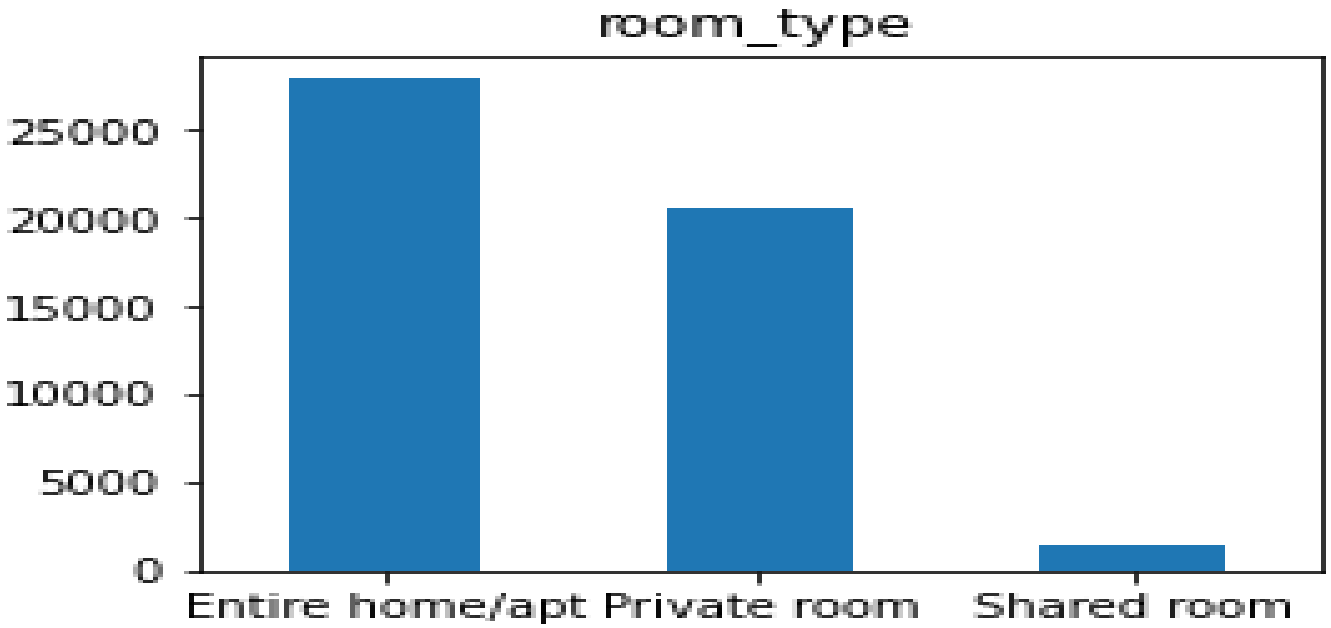
Data Exploration

Common Type of property



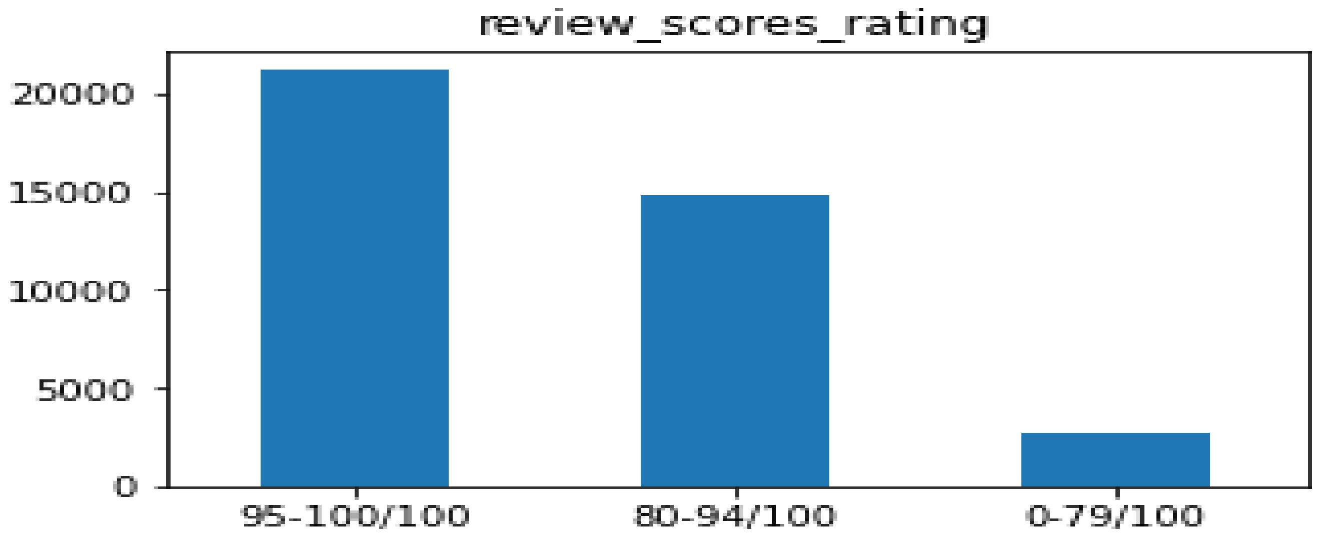
About 70 % of Property are apartment ,remainder are houses or more uncommon property types.

Data Exploration



- ❑ About 55% of listings are entire homes (i.e. you are renting the entire property on your own).
- ❑ Most of the remainder are private rooms
- ❑ Fewer than 1% are shared rooms (i.e. you are sharing a room with either the property owner or other guests).

Data Exploration



- ❑ For every review category, the majority of listings that have had a review have received a 10/10 rating for that category (or 95-100/100 overall) .
- ❑ Clearly people love their Airbnbs
- ❑ Ratings or 8 or below are rare
- ❑ Guests seem to be most positive about communication, check-ins and accuracy

Feature engineering

- Text feature extraction
- All Airbnb listings are provided with the detailed description of the accommodations in a text field “Description”. This field is very useful for hosts to highlight important and attractive features of their properties. We found out how this field impacts the price as well as user review ratings by creating 2 new features out of the description field:
- Sentiment Intensity score – VADER sentiment intensity analyzer was used to identify the sentiment of each description. The range of values are from -1 to 1. VADER gives a cumulative score by adding the sentiment score of each word in the description. Strong positive words get a +ve score and similar –ve score is given to negative words. Negation is also incorporated while giving the score. Capitalized words and punctuations affect the overall score too.
- Description Length - The length of the description was obtained by performing a word count on this field. This feature was engineered to find out if longer descriptions yielded in a higher rating and were people willing to pay the extra money just because the host was very friendly and enthusiastic.

Feature engineering

- Topic modeling
- All the description records were combined to form the corpus and Latent Dirichlet Allocation was employed on the corpus to find out the top 5 topics. After we got the top words in the 5 topics, 4 were meaningful. These 4 topics were now labeled after looking at the topic word distribution carefully. The topics were –

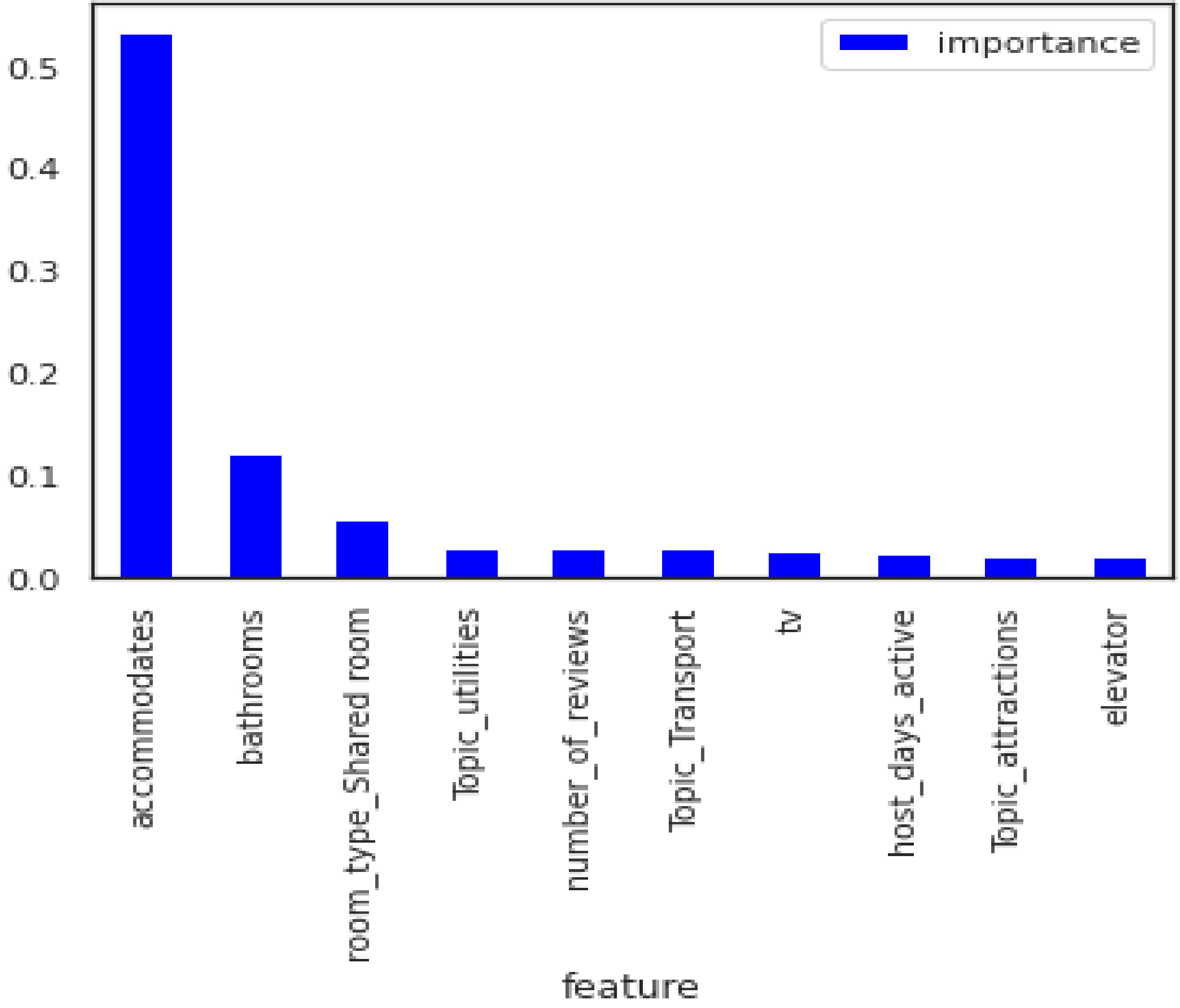
  1. Description about the listing – like high floored, furnished, etc
  2. Transport – nearest transport facilities from the listing
  3. Attractions – Nearby places of interest from the listing
  4. Amenities – facilities provided by the host The probability of each of the topics in every description was calculated and added as feature in the dataset.

Modeling Approaches

- ❑ Linear and Ridge Regression model
- ❑ Random Forest
- ❑ XgBoost
- ❑ Neural Network

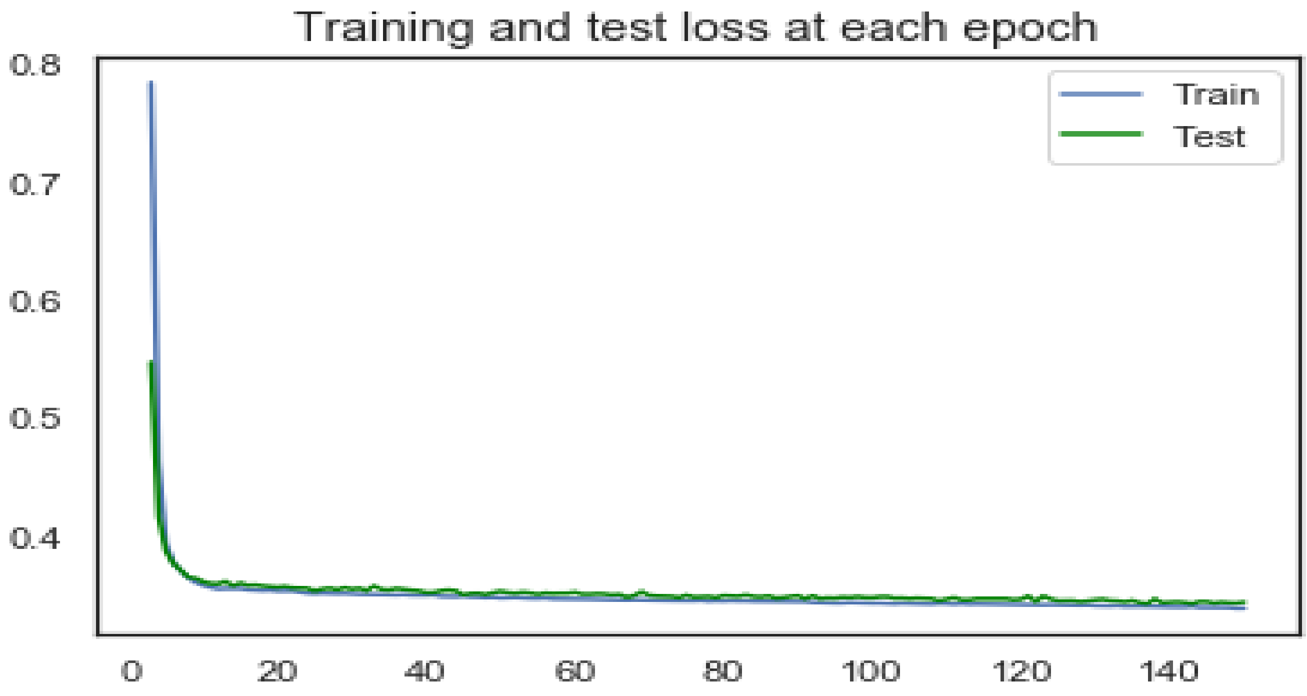
Modeling Interpretation

Top 10 features selected Random Forest



Model Inferences

Neural Network Model



Neural Network has also removed the overfitting issue, as train and test sets are performing the same.

Results

	Training RMSE	Validation RMSE
Linear Model	.4911	.5102
Random Forest	0.4527	0.4933
XgBoost	0.3867	0.4201
Neural Network	0.5132	0.5133

- We can see from above results that Neural Network has removed the over fitting issue, as train and test sets are performing the same

Conclusion

❑ For price prediction, we started with a linear regression model with independent features available in the original dataset (which is our base model).

❑ Overall, the XGBoost model is the preferred model, which performs ever so slightly better than the best neural network and is less computationally expensive.

❑ It could possibly be improved further with hyper-parameter tuning.

Thank You