Singapore Airbnb rental price Capstone project

Anitha Veeramani

Aug 10th, 2020

Machine Learning Engineer Nanodegree Capstone Proposal

Project Background

Built a model to predict rental price in the Singapore market using ML model. Identified the best price that a client can rent their house based on the market price and previous listings

Airbnb is an internet marketplace for short-term home and apartment rentals. It allows the owner to rent out home for a week, or rent out empty bedroom. One challenge that Airbnb hosts face is determining the optimal nightly rent price. Hosts with multiple listings are more likely to be running a business, are unlikely to be living in the property.

Why Singapore?

As in **SINGAPORE** — Short-term rentals offered by platforms such as **Airbnb** will remain **illegal**, and later the rule was implemented were rental is allowed only minimum stay of three months to private residential properties, hence personally would like to explore Airbnb Singapore data to check the violation and hotspot rental areas in Singapore after analyse the dataset.

Machine Learning Workflow - Singapore AirBnB price prediction Modelling

Problem Statement

The purpose of this task is to predict the rental price of Singapore Airbnb based on the data provided in the kaggle and inside Airbnb dataset and analyse what are the key features that affect the price of the listing. We use machine learning models to predict the optimal prices that the hosts can set for their properties. This is done by comparing the property with other listings on a variety of parameters like location, property type and other features. Also, based other data, the hosts would like to know what review rating they can expect or if they have a price with respect to other similar listings or are losing out on rating because of their pricing.

Datasets and Inputs

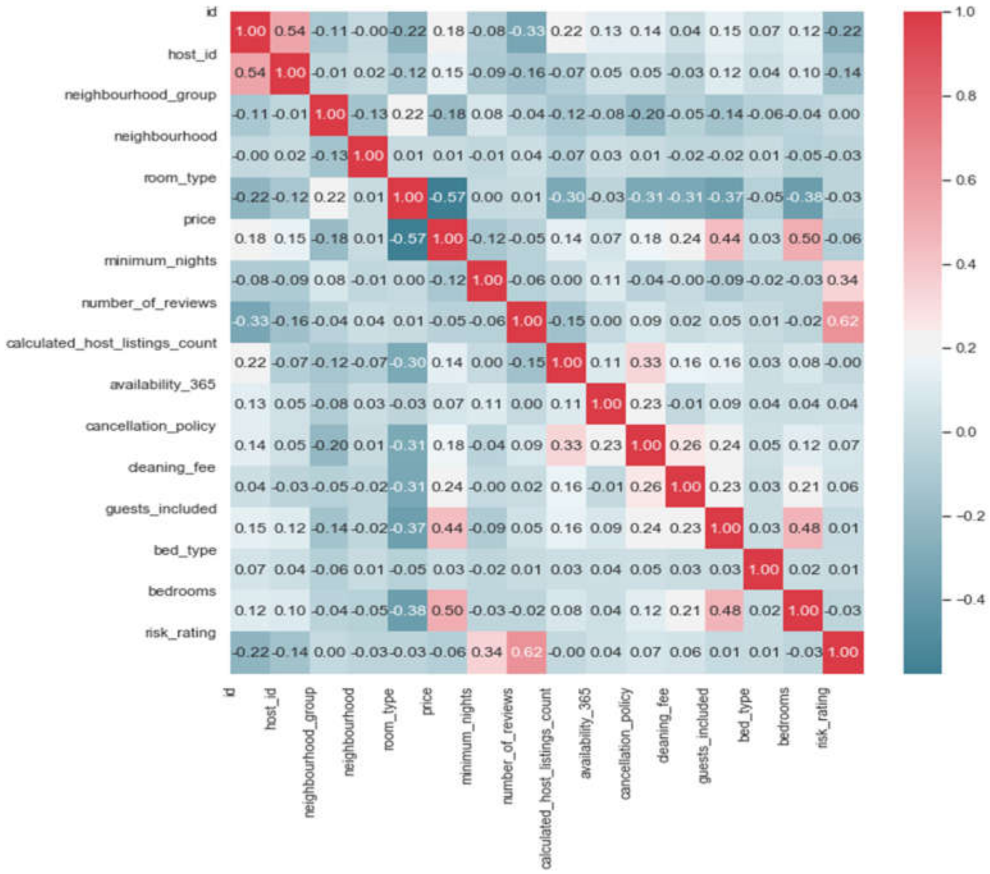
The Singapore Airbnb dataset used for this project comes from kaggle.com (listings.csv) and additional information extracted from detailed Singapore listings from insideairbnb.com (listings.csv.gz) which was collected on 28 August 2019 according to the website.

The data set as scraped by Inside AirBnb is extremely comprehensive, containing 7320 rows and 106 columns. However, many of the data collected are unnecessary metadata (e.g. urls, space, experience, summary) has low relevance to the yield model, hence extracted only relevant data which are useful for this project as below:

- guest-included- the number of guests the rental can accommodate

- bedrooms- number of bedrooms included in the rental
- cancellation_policy- listings cancelation policy like flexible or strict
- cleaning_fee- additional fee to clean after use the place
- property_type- Type of property like hotel, apartment, condo etc.
- bed_type- types of bed like sofa, real bed or air bed etc..

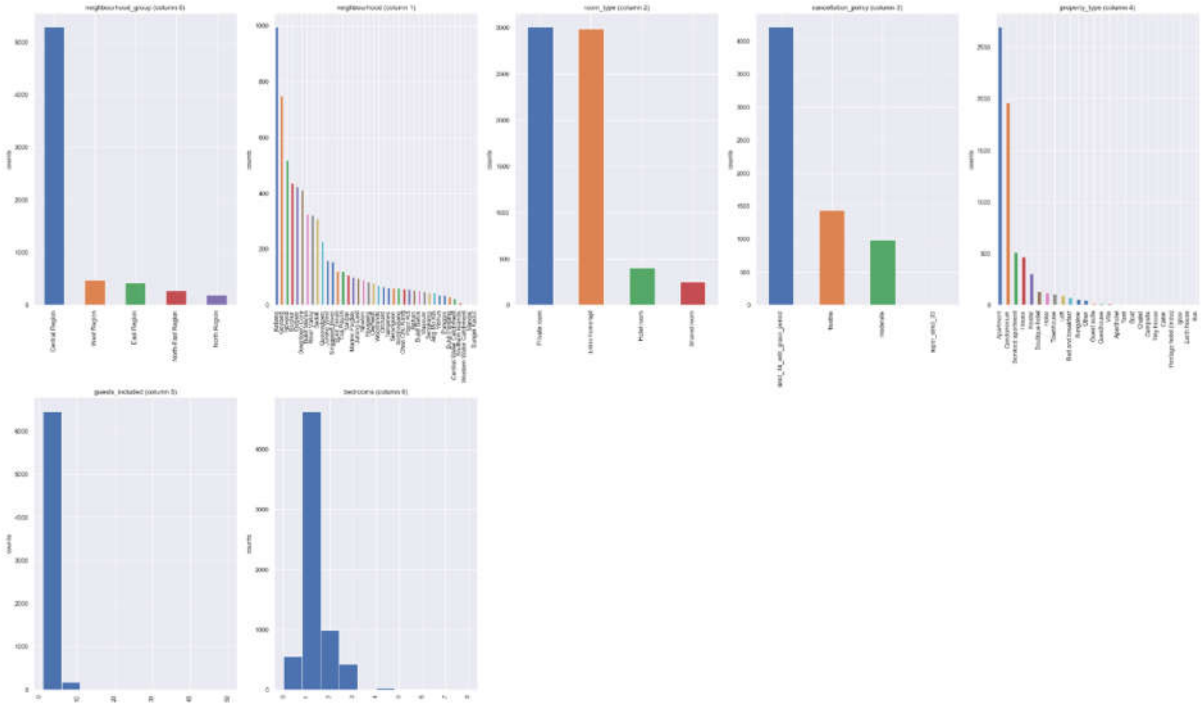
To predict the rental price, first find high level trends and correlation between other features with property prices are analysed using heat-map. From the heat-map we can choose the features whose values are in red like bedrooms & guests_included as key features for price prediction.



Exploratory Data Analysis

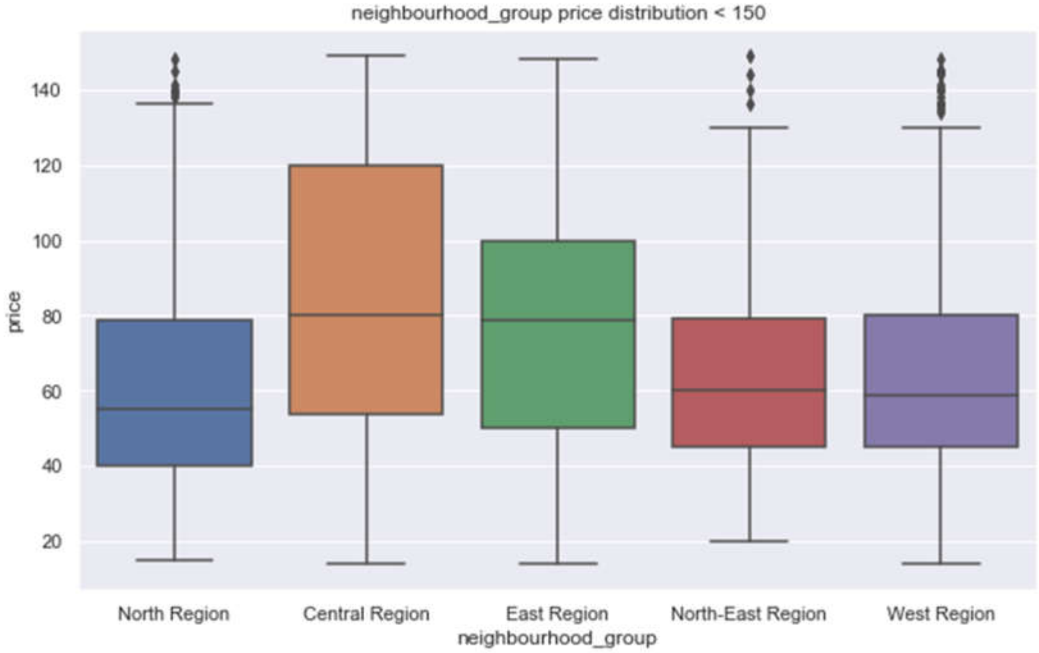
In the beginning of our analysis, it is essential to visualize the data and especially the variables we are interested in. For the output rental price variable, the histogram provides us with information regarding its distributional properties. According to the plot, it is apparent that the distribution of price is not normal as most of the listing is below 500. Hence we filter records above 500 as threshold and train the model.

For further analysis, histogram of all variables is plotted to analyse data format and their distribution

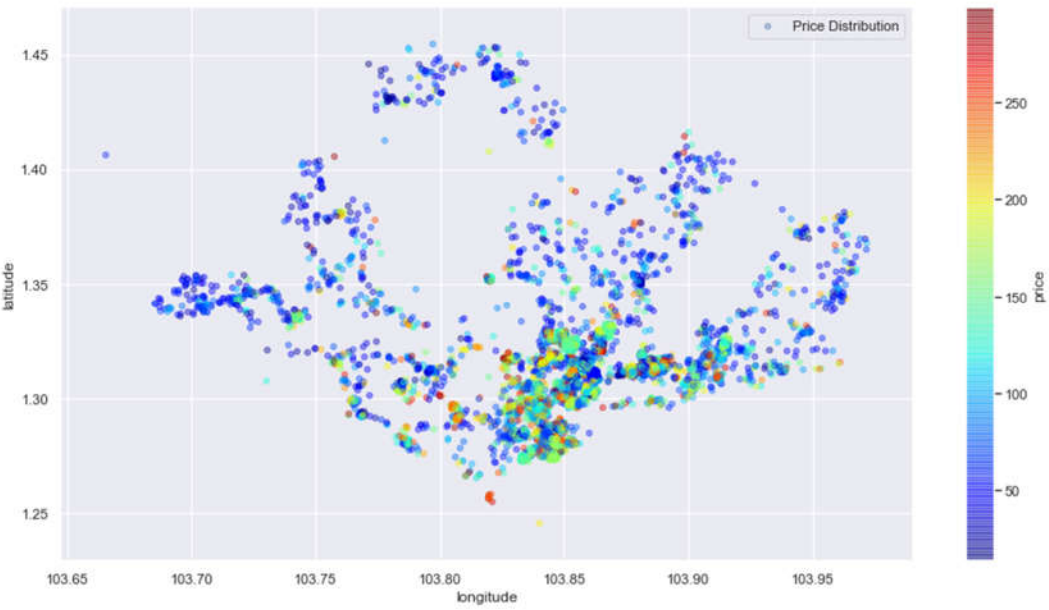


Further explore the impacts of all the other features against the price, in order to provide more informative based on their distribution's summary and find answers for certain questions as below

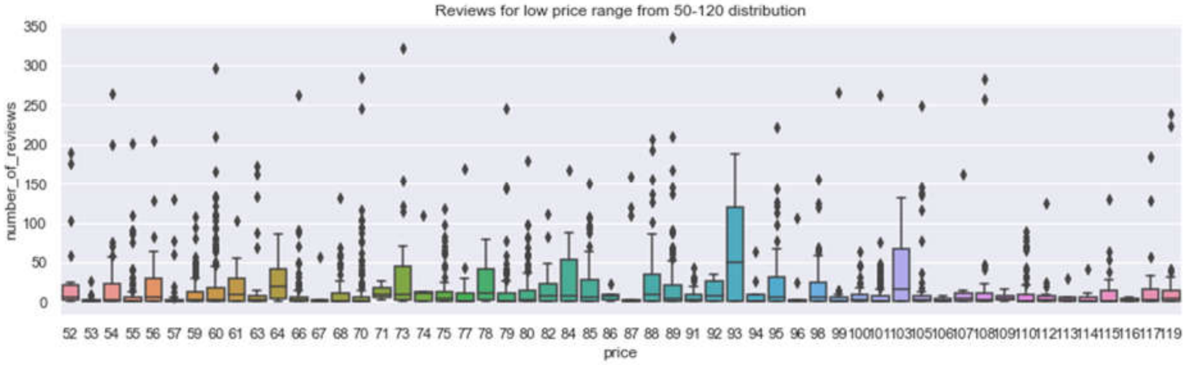
1. Hot area where most of the listings present in the neighbourhood. This conclude that more number of listing are present in the Central region than in the North-East region.



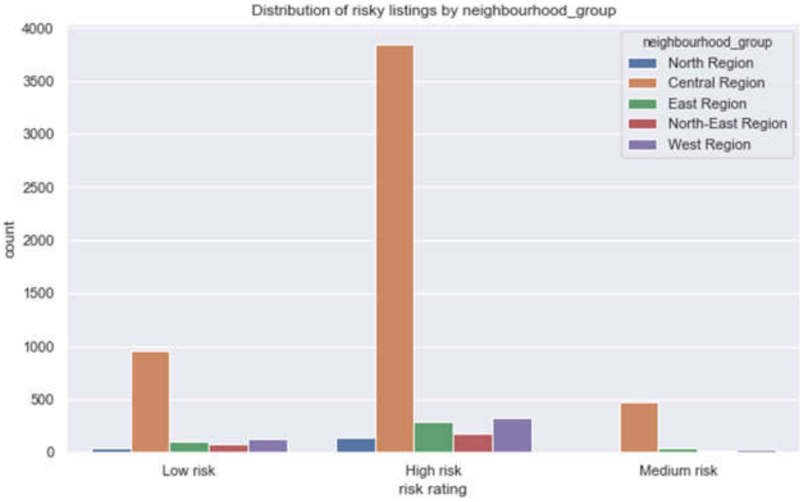
2. Higher/ Expensive price listings available in the neighbourhood regions are explained by price against the geographical location of the Singapore using lat-long features as below. Generally, people prefer to choose places in the central part of Singapore. Hence places in the Central region such as orchard etc.. is highlighted in this map



3. Impact of number of reviews against the price. As my assumption cheap places always get most review. But after analysis of plot below, it clearly states that irrespective of price the number of reviews vary. Hence we can understand there is no correlation between price and review. Hence they are independent features

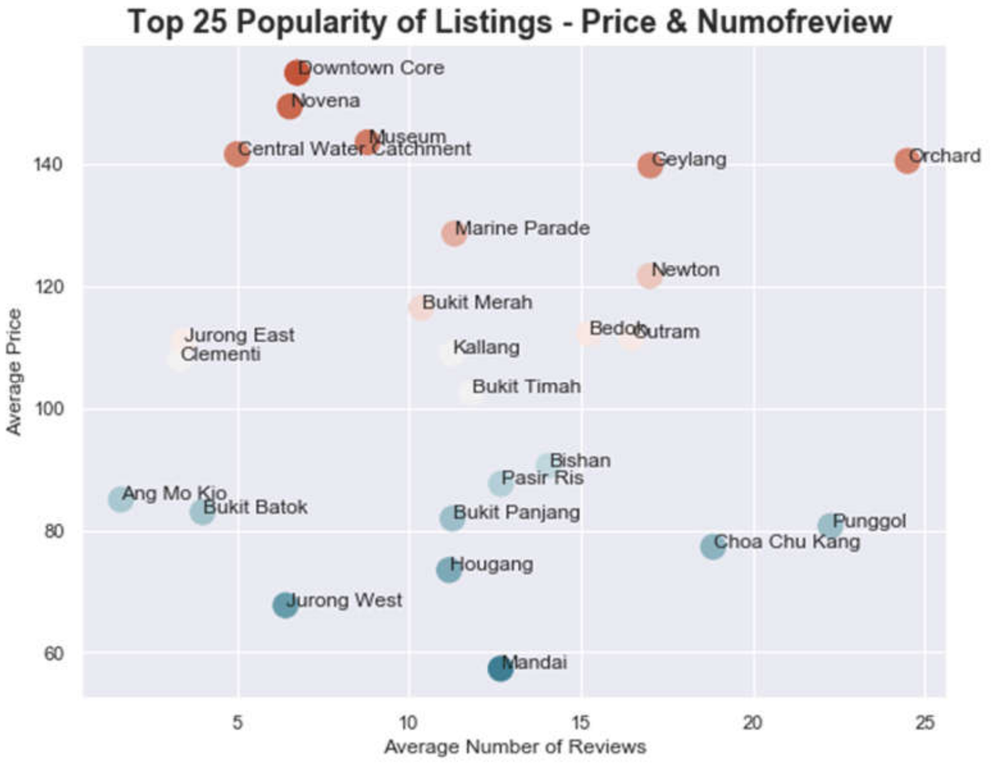


4. Risky listings are identified when the number of reviews are high for the particular place and minimum_night stay is low we create new feature as Risky_rating. This new feature defined as High, low, medium based on number of nights < 30 and more reviews as high. Then later compared plot against risky listing against neighbourhood as below



As more listing in the Central region, there is more high risk listing also identified in Central region is clearly explained in the plot

5. Top 25 popularity listings based on number of reviews against price (Multiple variable correlation)



Feature Engineering and feature selection:

We often need to perform additional data processing after we load the dataset, and also we need to understand the dataset by using various techniques before we can fit a model. Define relevant and important features and normalise them to train the model

Fill missing values: We need to check if any data is missing and investigate why and understand the root cause of missing data If found, we should replace the missing values either with default or with meaningful values. Example replace bedrooms as 1 for the missing value of the room type is private.

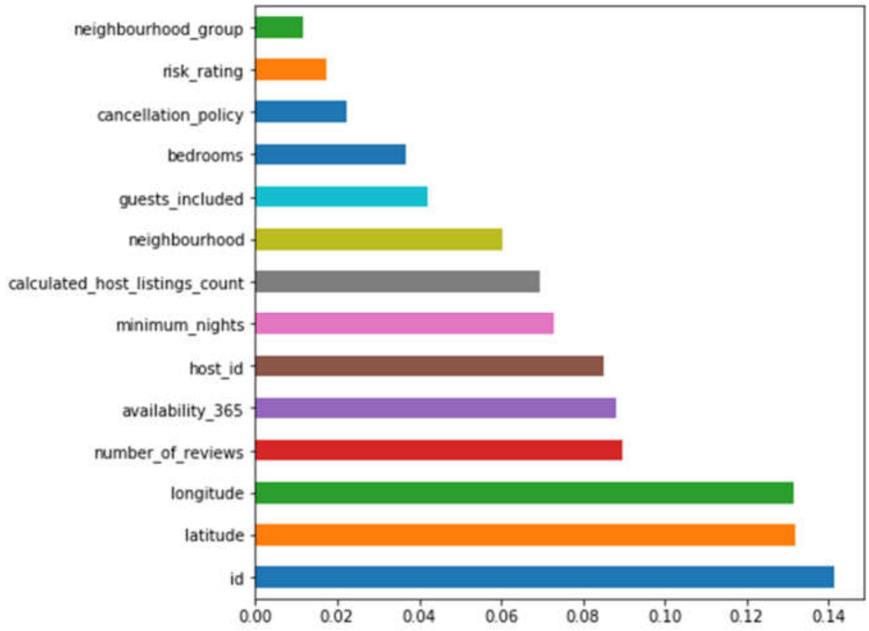
Encoding Categorical Variables: Since model cannot handle string type variable, we need to encode the raw categorical features into int . When the number of categories is low, we use label encoder than one hot encoding which is for binary label data. And, when the cardinality is high, use ordinal encoding which encode based on frequency count of each category. As most of the features like "neighbourhood_group, neighbourhood, room_type, bed_type, risk_rating, cancellation_policy" are low cardinality feature we use label encoding to convert categories to int.

Remove irrelevant data: To make the data set less cumbersome to work with, we removed many of the columns in the original data set like reviews_per_month, hostname, name etc..

Removal of outliers: Removed prices more than $500, as these prices were not having enough support.

Feature extraction: Risk_rating is the only extra features generated based on the minimum night stay is <30 and get more number of reviews as High, Medium and low

Feature Selection: To better understand what is driving the predictions of price we examine how each feature is ranked by our model. From the feature ranking it is clearly displayed that location, number of reviews and room availability within 365 days are primary indicators of price.



Cleaned and pre-processed dataset is stored in data folder listings_clean.csv. For further to train the model, the cleaned dataset is splitted into 70% training set and 30% testing set using sklearn modules. After model selection evaluate the best model on the testing set.

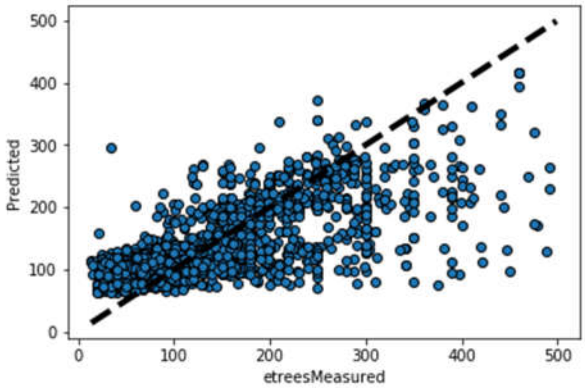
Model Selection & Validation:

In this section we need to train different model prototype and then do model selection and tuning based on evaluation metrics. Since the listing price is continuous in nature, we used regression algorithms like linear regression as based model, tree based repressors (RF, evaluation tree) and XGboost as we believe always XGboost give best performance for regression problem. Finally, the models are evaluated using evaluation metrics. Interpret the model results by predicting price for the test dataset and calculate accuracy based on prediction and ground truth value.

Benchmark Model: For this problem, the benchmark model will be XG-Boost and Linear regression model. Using XG-Boost, I was able to predict price based on a cleaned dataset, with a R^2 score of ~0.70 on test datasets. A model with a larger R-squared value means that the independent variables explain a larger percentage of the variation in the independent variable.

Evaluation Metrics: Prediction results are evaluated on the log loss between the predicted values and the ground truth. Accuracy based on RMSE and R^2 is calculated. RMSE is an important indicator of a model's accuracy, XG-Boost with low root-mean-squared error (RMSE) and high R^2 based on cross-validation that can selected as best fit model.

Model	RMSE	R^2
Linear Regressor	72.74	0.334
Extra Trees Regressor	61.51	0.524
Random Forest Regressor	51.16	0.671
XGBoost Regressor	48.398	0.705

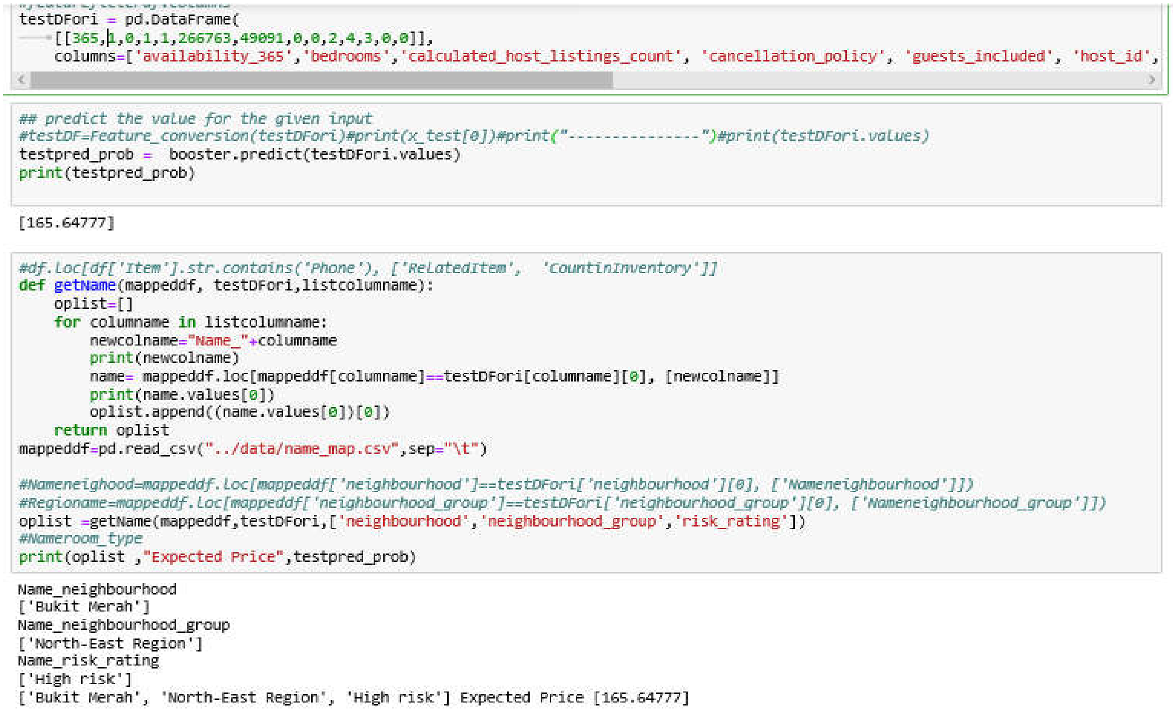


A model with a larger R-squared value means that the independent variables explain a larger percentage of the variation in the features and more variable are scattered.

Model-evaluation tools using cross-validation:
Such as `model_selection.cross_val_score` and `model_selection.GridSearchCV` rely on an internal *scoring* strategy. In the case of the SgAirbnb dataset, the samples are balanced across target classes hence the accuracy score is >98% .

Prediction:

Mean absolute error (MAE), mean squared error (MSE) and R2 score were used to evaluate the trained models. Use the selected model prototype, predict rental price for the listing based on user input as shown below



```
testDFori = pd.DataFrame(
    [[365,1,0,1,1,266763,49091,0,0,2,4,3,0,0]],
    columns=['availability_365','bedrooms','calculated_host_listings_count', 'cancellation_policy', 'guests_included', 'host_id',
```

```
## predict the value for the given input
#testDF=Feature_conversion(testDFori)#print(x_test[0])#print("---------------")#print(testDFori.values)
testpred_prob =  booster.predict(testDFori.values)
print(testpred_prob)
```

```
[165.64777]
```

```
#df.loc[df['Item'].str.contains('Phone'), ['RelatedItem',  'CountinInventory']]
def getName(mappeddf, testDFori,listcolumnname):
    oplist=[]
    for columnname in listcolumnname:
        newcolname="Name_"+columnname
        print(newcolname)
        name= mappeddf.loc[mappeddf[columnname]==testDFori[columnname][0], [newcolname]]
        print(name.values[0])
        oplist.append((name.values[0])[0])
    return oplist
mappeddf=pd.read_csv("../data/name_map.csv",sep="\t")

#Nameneighood=mappeddf.loc[mappeddf['neighbourhood']==testDFori['neighbourhood'][0], ['Nameneighbourhood']])
#Regioname=mappeddf.loc[mappeddf['neighbourhood_group']==testDFori['neighbourhood_group'][0], ['Nameneighbourhood_group']])
oplist =getName(mappeddf,testDFori,['neighbourhood','neighbourhood_group','risk_rating'])
#Nameroom_type
print(oplist ,"Expected Price",testpred_prob)
```

```
Name_neighbourhood
['Bukit Merah']
Name_neighbourhood_group
['North-East Region']
Name_risk_rating
['High risk']
['Bukit Merah', 'North-East Region', 'High risk'] Expected Price [165.64777]
```

Conclusion:

In this study, we used different machine learning algorithms to predict Singapore Airbnb's listing price. First data is cleaned, normalised and important features were extracted after detailed data analysis. Then split the data into train and test dataset. For training models train more than one regression algorithm and finally select the best model for this problem which give best accuracy based on R^2 square. Removing features like minimum_stay, number_of_reviews and guest_included affect the R^2 value of the XGboost model from 70% to 67%

Hence from this study we can define the features such as minimum stay, price and number of reviews have been used to estimate the occupancy rate and the income per month for each listing than other features. We learned that RMSE can be used to calculate the error of our models, which we can then use to iterate and try and improve our predictions.

Reference:

1. https://towardsdatascience.com/predicting-airbnb-prices-using-machine-learning-in-vancouver-1b42ca52eece
2. https://github.com/priyadarsanshankar/AirBnbPricePrediction/blob/master/README.md
3. https://www.kaggle.com/kerneler/starter-airbnb-singapore-listing-d82a5416-b
4. http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647491.pdf
5. Regression Algorithm : https://www.dataquest.io/blog/machine-learning-tutorial/
6. Data Exploration: https://github.com/Modingwa/Data-Engineering-Capstone-Project