# **New York City Airbnb Open Data**

CU Boulder CSCI 5502 Data Mining Project Proposal

Authors: Andrew Edds, Hari Purnapatre, Karan Juneja, Kexin Yu, Rohan Jayaram, Tittiwat Tonburinthip

Date: Feb 15, 2022

**Abstract**

New York City (NYC) is the most popular city in the United States, and it has been described as the cultural, financial and media capital of the world. Because of NYC’s characteristics, the rents in NYC were the highest in the nation. And since 2021, the rental rate has gone up 21%. NYC is also a place that many people want to have a journey. Except for living in traditional hotels, people, especially young people, are loving to book rooms from Airbnb nowadays. Airbnb is a website that can provide users with a variety of accommodation information, such as the location, reviews from previous residents, price, etc. In this paper, we create several models based on the NYC dataset from Kaggle, building a model that can accurately predict the rent is our objective. Decision Tree,Random Forest and LightGBM are the algorithms we have used to create models. Due to the real world, it’s impractical that the data we use is tidy enough and can be used to create a good fitting model. To improve the accuracy of our predictions, we have done the hyperparameter tuning. Outliers are another thing that can affect the accuracy. We have done some tests about whether a model can become better from deleting some variables’ outliers. To test the effectiveness of our model, GroupKFold and StratifiedKfold were used to split data, and MAE is the metric we use to evaluate our models. The model created with LightGBM and Stratified KFold gets the best result, the MAE of this strategy is 62.4 (train) and 58.8 (test).

**Introduction**

New York is the [most populous city](https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population) in the United States and one of many favorite destinations for many people worldwide. It’s also an essential prime area for international companies, universities, restaurants, etc. New York city composes of five boroughs—[Brooklyn](https://en.wikipedia.org/wiki/Brooklyn) (Kings County), [Queens](https://en.wikipedia.org/wiki/Queens) (Queens County), [Manhattan](https://en.wikipedia.org/wiki/Manhattan) (New York County), [the Bronx](https://en.wikipedia.org/wiki/The_Bronx) (Bronx County), and [Staten Island](https://en.wikipedia.org/wiki/Staten_Island) (Richmond County). Thanks to a good location for any business, New York is now officially the most expensive city to live in, and the rental rate has gone up 21% since 2021. Many tourists choose to find apartments from Airbnb, an online application, instead of luxury hotels to stay for a short period. Airbnb launched its application to help tourists find a good lodging place, primarily [homestays](https://en.wikipedia.org/wiki/Homestay) for [vacation rentals](https://en.wikipedia.org/wiki/Vacation_rental). The company also has their customer reviews to persuade other people to stay. This application has an algorithm to calculate the rental rate per night, which depends on the number of nights to stay, the availability of that place, and room types. This project aims to create a machine learning model to predict the rent per night in the different locations of New York City. Several techniques such as Decision tree, Random forest, and LightGBM.

**Related Work**

There are several projects, both corporate and academic, that have utilized short-term vacation rental datasets to understand tourism markets in the past. As introduced, with an estimated 500 million guests per year, the Airbnb market is massive and growing. Thus, with more listings, stays, and guests, comes more data, and with more data, we can make more insights into this vacation rental industry.

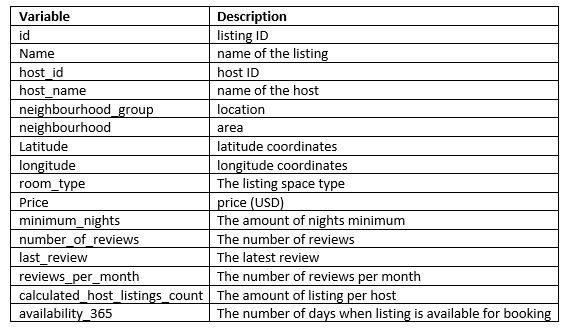
Already in the market, there is a company called AirDNA that aims to amalgamate Airbnb data for investors, renters, market analysts, and more. The service is built to provide users of a clear snapshot of how a specific Airbnb market operates. AirDNA aims to build dynamic price models, however, the variance of the price across markets is broad, and the ability to understand the dynamics of more specific regions, such as the Burroughs of NYC, is not available. This project aims to bridge that gap by making the models more dynamic based on location.

In addition to AirDNA, there are two companies called MashVisor and Guestly that also aim to use large vacation rental datasets to service insights for renters, guests, market analysts, and investors. Firstly, MashVisor is built to provide investors with an understanding of how a specific property can be excepted to perform on Airbnb, and thus, given the cost of the home, how good of an investment that property would be. Guestly works more for the renter by helping to optimize stay schedules, communication, pricing, and more.

There have also been several academic studies published that complete some sort of statistical analysis or modeling using short term vacation rental data, however, the field is relatively new and thus the academic work regarding Airbnb is limited. This project aims to fulfill many of these needs in the industry, and by providing stronger price modeling and vacation analysis, investors, renters, guests, and market analysts can rely on our findings to make informed decisions about how to operate in the market.

**Main method**

The NYC Airbnb data from kaggle has 48,895 rows and 16 columns.



Related-Variable

Visualization

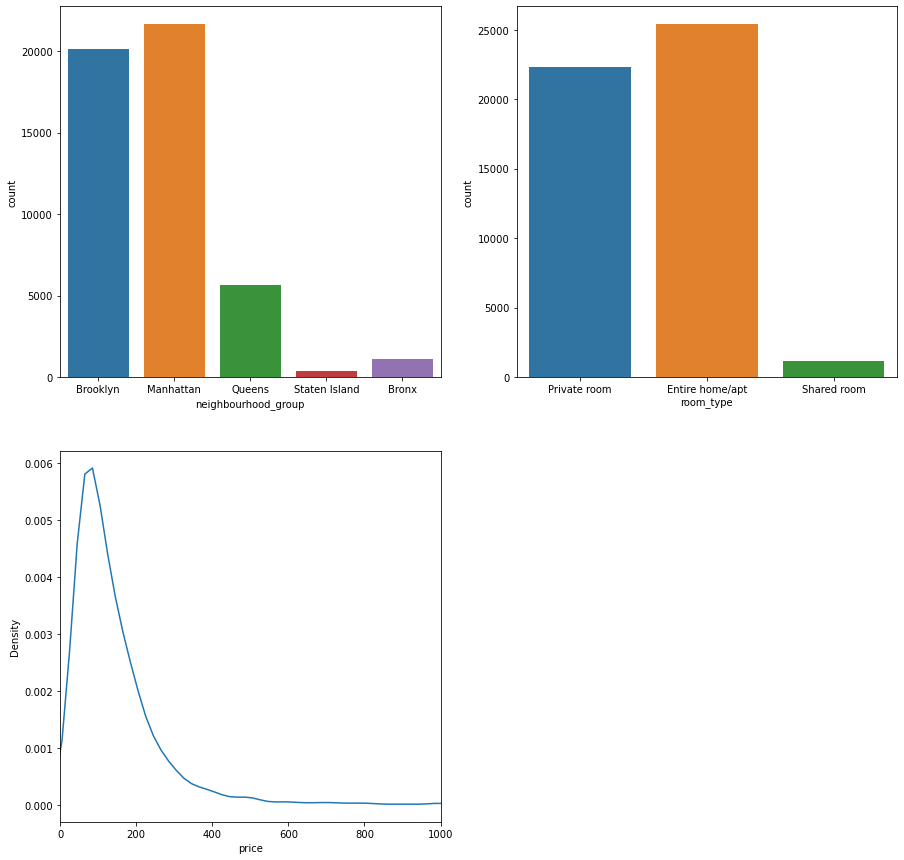


Fig1. general information

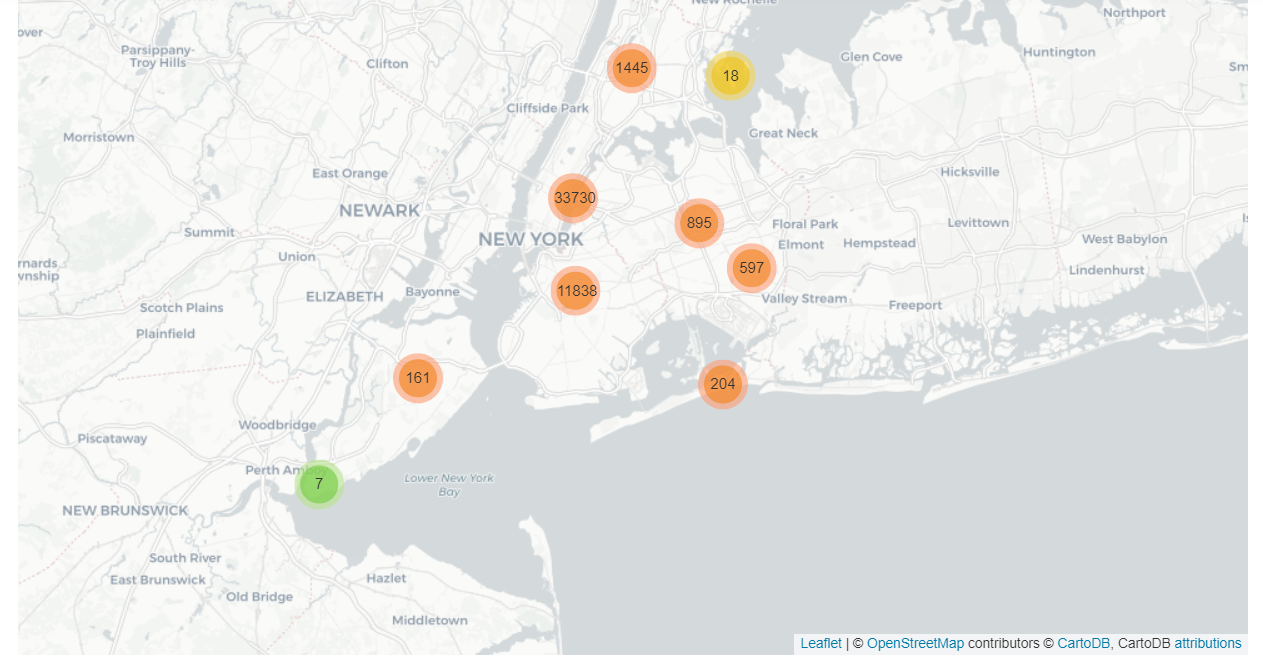


Fig2. Bubble map of each location in NYC

From the above figure, we can see that most of the data are in Brooklyn and Manhattan and the type of the rooms generally are the private rooms and entire homes. The price of the rooms are mostly less than 200 usd/night.

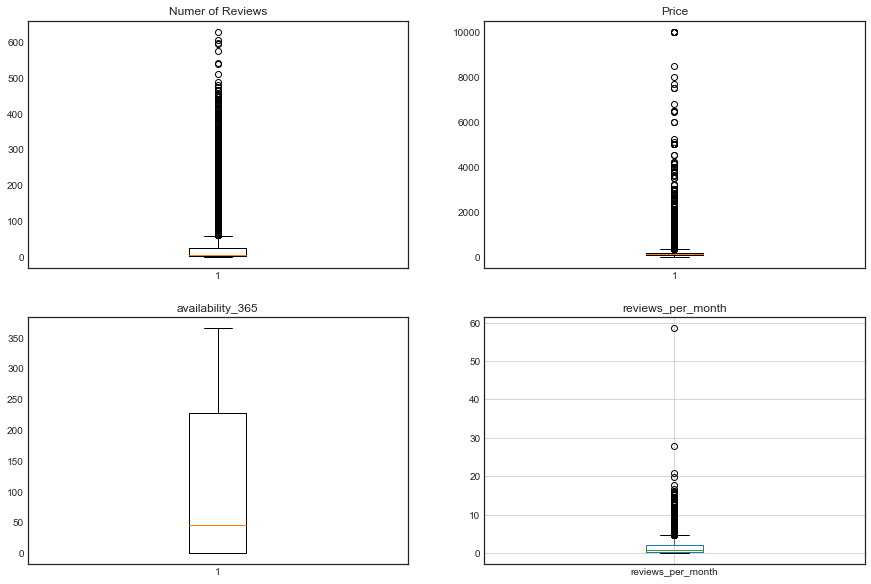


Fig3. boxplot

According to the figure 3, there are some outliers in the number of reviews, price, and reviews\_per\_month

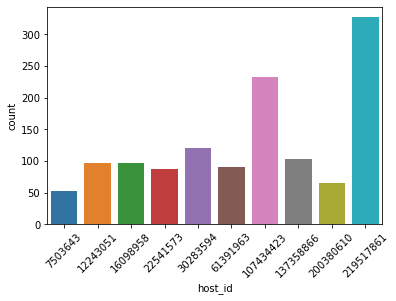


Fig4. The count of residences vs host id

According to figure 4, many hosts have a lot of residences in Airbnb. The number of unique hosts is 76.6 % of total data.

Data preprocessing

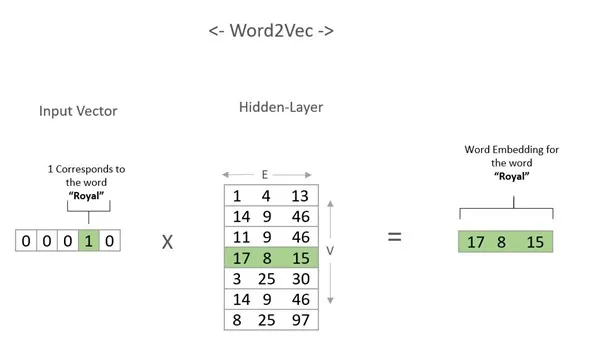
This project will implement three machine learning techniques.

Data preprocessing

1. Remove id, name, host\_id and host\_name columns
2. Fill NA of last\_review and reviews\_per\_month with means
3. Use Word2Vec to tokenize and create vector for the column ‘Name’ of the dataset
4. Using feature scaling such as standard scaler
5. Use GroupKFold and Stratified KFold to separate train and test sets

Word2Vec  
Word Embedding is a word representation type that allows machine learning algorithms to understand words with similar meanings. It is a language modeling and feature learning technique to map words into vectors of real numbers using neural networks, probabilistic models, or dimension reduction on the word co-occurrence matrix. Some word embedding models are Word2vec (Google), Glove (Stanford), and fastest (Facebook).

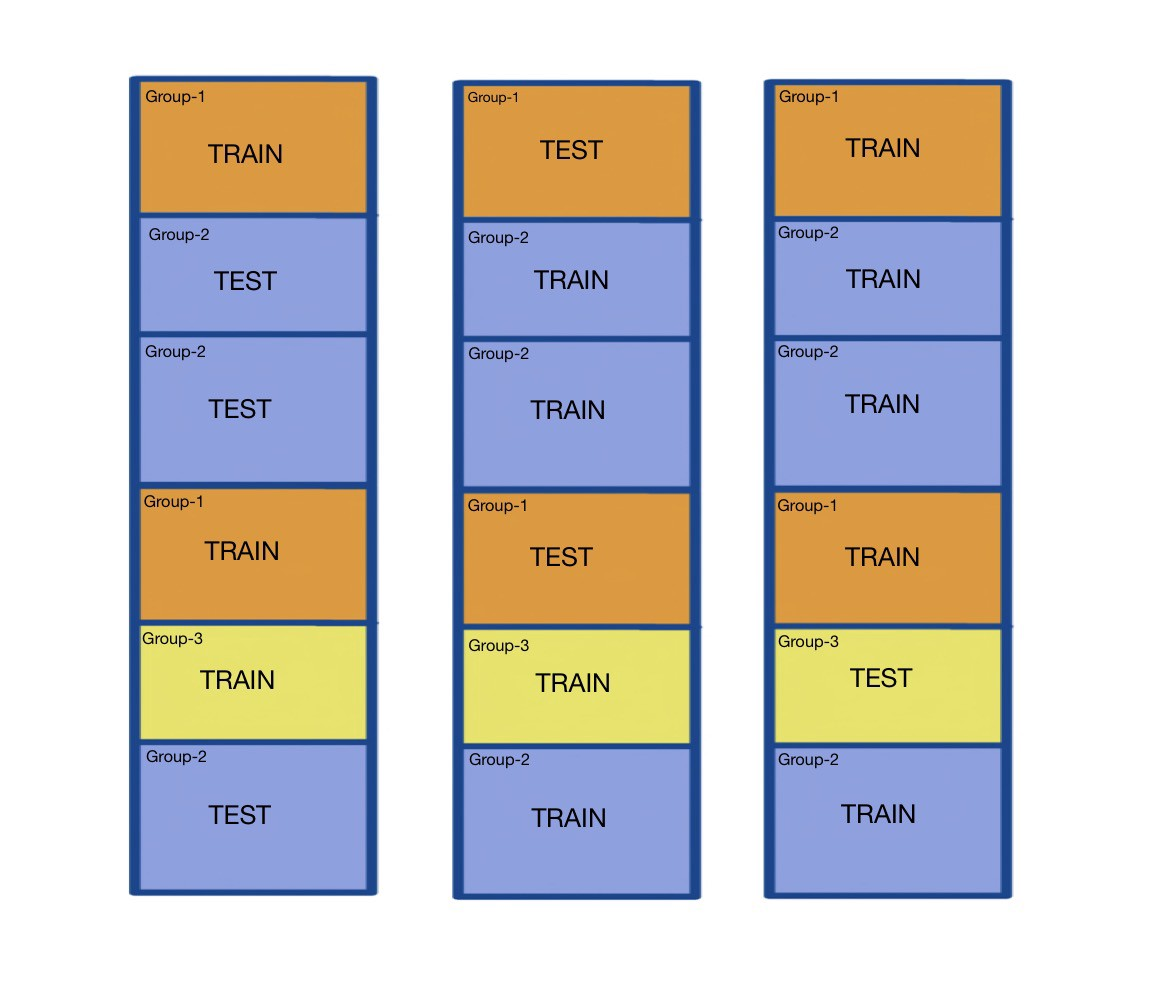
Machine learning methods



**Word2Vec**

GroupKFold

GroupKFold is the K-fold iterator variant with non-overlapping groups. The same group will not appear in two different folds (the number of distinct groups has to be at least equal to the number of folds). The folds are approximately balanced in the sense that the number of distinct groups is approximately the same in each fold.

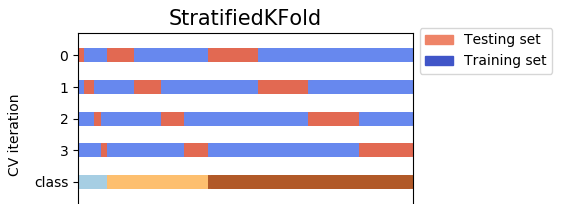
**GroupKFold**

GroupKFold

GroupKFold is the K-fold iterator variant with non-overlapping groups. The same group will not appear in two different folds (the number of distinct groups has to be at least equal to the number of folds). The folds are approximately balanced in the sense that the number of distinct groups is approximately the same in each fold.

StratifiedKFold

The stratified k fold cross-validation is an extension of the cross-validation technique used for classification problems. It maintains the same class ratio throughout the K folds as the ratio in the original dataset. So, for example, you are dealing with diabetes prediction in which you have the class ratio of 70/30; by using stratified K fold, the same class ratio is preserved throughout the K folds.



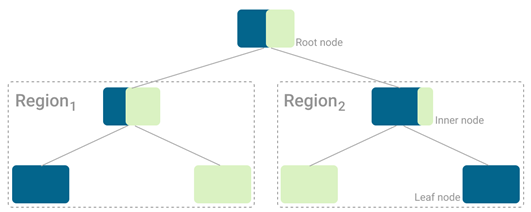
**StratifiedKFold**

Machine learning methods

Decision tree

A decision tree is a Supervised Machine Learning Algorithm that uses a set of rules to make decisions, similar to how humans make decisions. The intuition behind Decision Trees is that you use the dataset features to create yes/no questions and continually split the

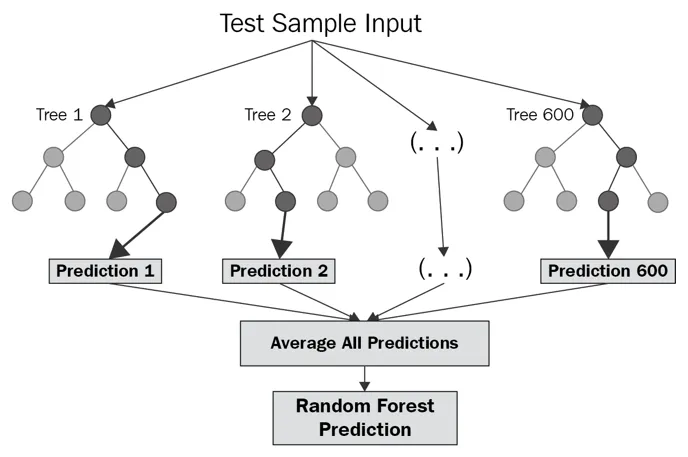
dataset until you isolate all data points belonging to each class. With this process you’re organizing the data in a tree structure.



**Tree structure**

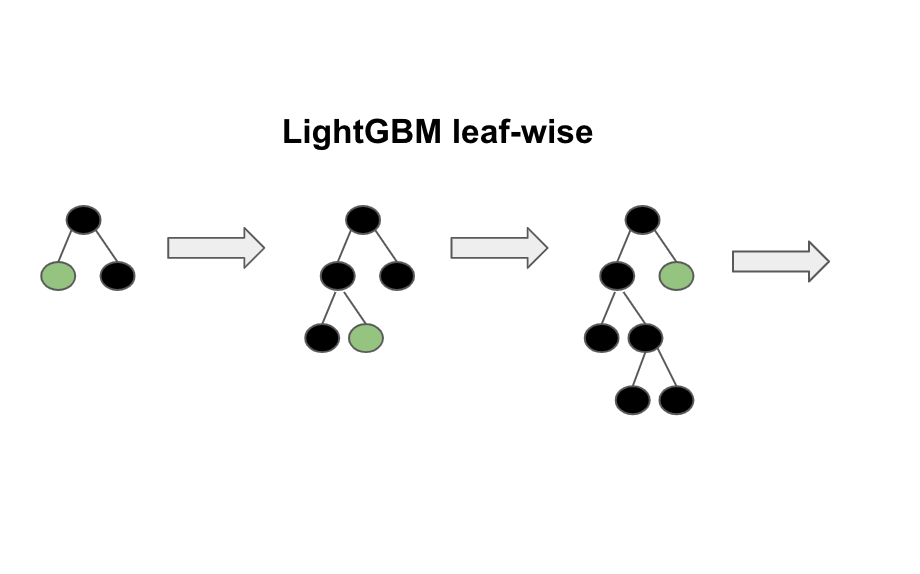
Random Forest

Random forest is a technique used in modeling predictions and behavior analysis and is built on decision trees. It contains many decision trees representing a distinct instance of the classification of data input into the random forest. The random forest technique considers the instances individually, taking the one with the majority of votes as the selected prediction.



LightGBM

LightGBM is an open-source framework for gradient boosted machines. By default LightGBM will train a Gradient Boosted Decision Tree (GBDT), but it also supports random forests, Dropouts meet Multiple Additive Regression Trees (DART), and Gradient Based One-Side Sampling (Goss). The framework is fast and was designed for distributed training. It supports large-scale datasets and training on the GPU. In many cases LightGBM has been found to be more accurate and faster than XGBoost, though this is problem dependent.



**Evaluation**

The following method will be used to evaluate the model:

1. If it is a classification problem, ROC-AUC curve will be used to evaluate the performance of the model. It is a chart that visualizes the tradeoff between true positive rate (TPR) and false positive rate (FPR). Basically, foreverythreshold, we calculate TPR and FPR and plot it on one chart.It can be shown that ROC AUC score is equivalent to calculating the rank correlation between predictions and targets. From an interpretation standpoint, it is more useful because it tells us that this metric shows how good at ranking predictions your model is. It tells you what is the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance.
2. If it is a Regression problem, R square and root mean square error will be used. R square: R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a [regression](https://www.investopedia.com/terms/r/regression.asp) model and root mean square error is the standard deviation of the residuals

**Results**

At first, our dataset was split into training and testing data by using a function train\_test\_split in the Scikit learn package. We found that there are some parts of data that cannot be equally distributed, especially in host\_id. The Venn diagram showed how the data were distributed. After examining the data, we decided to use KFold to solve this problem.

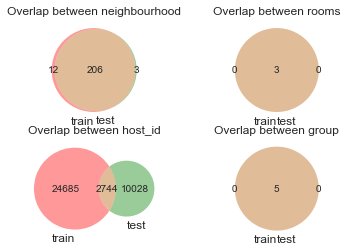


Fig.5 Venn diagram of training and testing data

| Model | Train MAE | Test MAE |
| --- | --- | --- |
| LightGBM Group K | 63.6 | 62.0 |
| Random Forest Group K | 63.5 | 60.2 |
| Decision Tree Group K | 66.8 | 61.0 |
| LightGBM Strat | 62.4 | 58.8 |
| Random Forest Strat | 62.3 | 61.4 |
| Decision Tree Strat | 63.5 | 61.0 |

\*Group K= Group KFold

\*Strat=Stratified KFold

\* MAE= Means absolute error

After we realized the problem of distribution, the dataset was split into two parts which are training dataset and testing dataset. The training dataset was performed in two types of cross-validation, GroupKFold and Stratified KFold. Then, the datasets were used to create the model by using the Decision tree, Random Forest, and Light GBM. The results of the models’ MAE are shown in the above table. The LightGBM of the Stratified KFold dataset is the best model. The gradient boosting model can achieve better results over the decision tree and random forest. Besides, stratified KFold help to arrange various categories of data to be trained in the LightGBM model, which results in the lowest error.

**Milestones**

**Week 6**

Prepare the project proposal report and presentation.

**Week 7-8**

Data Preprocessing

Transforming raw data into an understandable format. It is an important step in data mining as we cannot work with raw data. In the data preprocessing process we clean the missing values in the data, use different encoding methods for nominal and ordinal data, and use feature scaling techniques like standard scaler to scale the features to a similar scale.

Exploratory Data Analysis

Use data visualization and statistical techniques to described dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data using both manual analysis and automated data exploration libraries that visually explore and identify relationships between different data variables, the structure of the dataset, the presence of outliers, and the distribution of data values in order to reveal patterns and points of interest, enabling to gain greater insight into the raw data.

**Week 9-10**

Feature Engineering

Use the insights gathered from Exploratory Analysis and the domain knowledge of the dataset to select and transform the most relevant variables from raw data to create a predictive model using machine learning or statistical modeling. The goal of feature engineering is to improve the performance of machine learning (ML) algorithms.

Model Selection

We will experiment with different models and compare the relative value of different statistical models and determine which one is the best fit for the observed data.

**Week 11-12**

Cross Validation

Experiment with different cross validation strategies which are basically resampling methods that use different portions of the data to test and train a model on different iterations. This will help us generalize our model better and prevent overfitting.

Model Evaluation

Measure the newly trained model performance on a new and independent dataset. This model will compare labeled data with it’s own predictions. We will select an appropriate evaluation metric to quantify the quality of the model's prediction.

Hyperparameter Tuning

A hyperparameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned. The same kind of machine learning model can require different constraints, weights or learning rates to generalize different data patterns. These measures are called hyperparameters, and have to be tuned so that the model can optimally solve the machine learning problem. By using Hyperparameter tuning, the model accuracy can be increased by optimizing the model for the particular data.

**Week 13-15**

Prepare the project final report and presentation.

**Conclusion**

Our models can play an important role in Airbnb price prediction. In addition to a regression problem, the part of natural language processing (NLP) plays an important role in generating the new features for our machine learning models. In this project, our data were split into training and testing data. The training data were performed using various types of cross-validation to confirm that the model was trained with the same distribution of data as in the testing data. Stratified Kfold can perform that task efficiently. When the data from Stratified Kfold is combined with the Lightgbm, the lowest MAE can achieve at 59. In the future, geospatial feature engineering can be a tool to improve model accuracy. It will help generate additional features such as the number of nearby businesses and transportation. This results in lower errors in the future.

**References**

* https://www.avanwyk.com/an-overview-of-lightgbm/
* https://cs229.stanford.edu/
* <https://corporatefinanceinstitute.com/resources/knowledge/other/random-forest/>
* Airdna.co
* Mashvisor.com
* Guestly.com

**Future Work**

In the future, we can try to incorporate the number of businesses around the listings to see if this has any impact on the prices of the apartments. Businesses such as restaurants, shopping malls and cinemas may have an impact on the price listing.

**Author information**

* Tittiwat Tonburinthip: Created ML model (decision tree and random forest) and use reverse geocoding
* Kexin Yu: Performed data visualization
* Rohan Jayaram: Worked on Interpretation of ML model using rfpimp packages.
* Karan Juneja: Worked on NLP and distribution of data.
* Hari Purnapatre: Worked on evaluation metrics, feature engineering and training.
* Andrew Edds: Worked on research by reading research papers and creating features based on domain knowledge and feature importance.