Disclaimer

We declare that this material, which we now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. We understand that plagiarism, collusion, and copying is a grave and serious offense in the university and accept the penalties that would be imposed should we engage in plagiarism, collusion, or copying. We have read and understood the assignment regulations set out in the module documentation. We have identified and included the source of all facts, ideas, opinions, viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged, and the source cited are identified in the assignment references. We have not copied or paraphrased an extract of any length from any source without identifying the source and using quotation marks as appropriate. Any images, audio recordings, video or other materials have likewise been originated and produced by me or are fully acknowledged and identified. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study. We have read and understood the referencing guidelines found at DCU Academic Integrity and Plagiarism Policy. We understand that we may be required to discuss with the module lecturer/s the contents of this submission.

Signed by: Kashmira Chawan, Bhargav Anant Athavale
Date: 01 Aug 2022

Airbnb Analysis Using Feature Selection Methods and Sentiment Analysis

Kashmira Chawan
School of Computing
Dublin City University
Dublin, Ireland
kashmira.chawan2@mail.dcu.ie
21261109

Bhargav Anant Athavale
School of Computing
Dublin City University
Dublin, Ireland
bhargav.athavale2@mail.dcu.ie
20210278

***Abstract*—In recent years, our interaction with the world around us has changed. 'Sharing economy' is a new socio-economic model in which resources can be shared with other people, as and when required, often through digital means. Airbnb is a notable example. Airbnb is an online marketplace for short term rentals and tourist experiences. The primary reasons for the popularity of the platform have been its reasonable prices, availability and spread of locations. In this paper we perform an analysis of Airbnb data in three cities- New York, London and Sydney, to study the features affecting the prices of a listing. We use Filter, Wrapper and Embedded Methods for feature selection to understand the relation of different features with respect to prices in different cities. Also, in order to improve customer satisfaction, we perform Sentiment Analysis on customer reviews to gain insights into the different living experience, and how to improve it.**

***Index Terms*—Feature Engineering, Feature Selection, Airbnb, Sentiment Analysis, NLP**

I. Introduction

In the last decade, various digital marketplaces have become prominent where people exchange services. Companies such as Airbnb and Uber[1] are notable examples. These companies can be considered a part of the broader *sharing economy*. Airbnb offers short term rental living spaces and tourist experiences in various cities around the world. The company does not own any of the accommodations. The owners use the platform to advertise their listings and are referred to as *hosts*. As of 2022, there are 2.9 million hosts on the platform [16]. There are also a total of 7 million listings in over 220 countries. It offers a wide variety of listings which are shared or private depending on the customer requirement from single room flats to bungalows to townhouses. The listings also differ based on the services offered such as the number of beds, locations, amenities offered based on prices. Unlike the traditional hotel industry, the prices of the listings depend on the host and vary greatly across different cities and countries. It has emerged as an alternative to the traditional hotel industry.

Airbnb has been extensively studied in recent times. There have been studies analysis of prices using machine learning [1],on price prediction models [2], as well has using multi-modality data[3] to improve the platform. Data mining techniques have been used to extract important information which can drive the market. As the economy and demographics of an area change frequently, it is important to keep the price prediction models, feature selection methods updated in order to predict accurate results and improve services. In this paper, we therefore explore various feature selection techniques and feature selection methods in order to understand the important factors around Airbnb in a city. These can then be subsequently used to improve price prediction models. Each city will have different characteristics and important features.

Also, in the hospitality industry, customer reviews are very important as they can drive the market and affect the popularity of the platform. Happy customers are loyal customers. We perform Sentiment Analysis on the dataset to separate positive and negative reviews. We further extract key terms in negative and positive reviews separately in order to understand what drives their sentiment. This will allow use to improve the platform and develop better NLP models catered to the housing industry. Therefore we are trying to understand what factors influence Airbnb and how to utilize them to improve the entire system.

In Section II, we explore the different methods and different techniques used in the past with respect to Airbnb. It also inspects the methodology used and results obtained. In Section III, we explore the dataset, its source and pre-processing techniques used, which helped prepare the data for further analysis. Section IV explores the methodologies used in our study. Section V discusses the results obtained and Section VI talks about future scope and conclusion of this study.

II. Related Work

In an analysis of Airbnb data presented in [1] on various cities in the United States, they found that a variety of factors such as current availability, potential customers, days of the week affected the price of Airbnb listing. The main objective of the paper was to predict prices of Airbnb based on the geographical location and past reviews. Comparison between the traditional hotel industry and Airbnb is discussed. Airbnb has encouraged the customer to use the Internet to book accommodation unlike earlier hospitality industries. Airbnb economic models needs to be constantly evolving to keep up

[1]Airbnb founded in 2008 and Uber in 2009

with the ever-changing needs to maintain growth and stability [14]. The pricing of an Airbnb listing also depends upon various services and amenities offered by it. They observed that a successful booking is also dependent upon the interaction between host and potential customer [1]. Another finding from the paper was that Airbnb prices also affect the residential property prices in the neighbourhood. For this study, linear regression, logistic regression, and random forest algorithm have been used along with outlier detection. Exploratory analysis was done to visualize the trends over time in Airbnb listings in the USA from the period 2008-2020 [8].From the study they concluded that during the summer vacations the number of bookings increased, and when the reviews started increasing, people began to trust the platform more and this also factored into more bookings.

On a research conducted on New York Airbnb data in [2], various machine learning algorithms were used to create a price prediction model. The research was conducted on the 'New York City Airbnb Open Data' available on Kaggle for year 2019. Sentiment analysis was also performed on a reviews dataset available using the Natural Language Processing algorithm 'sentimentr'. Sentimentr uses four types of valence shifters to calculate the score of each sentence of review. Regression analysis was used for determining relationships between a dependent variable and one or more independent variables. Using ANOVA, a suitable model was selected. Multiple models such as Deep neural network, Random Forest (RF) and XGBoost were checked. In XGBoost new models are added to rectify the errors of earlier models. It also determines the complexity of the model. Bagging is a methodology where; the results of various predictor models are combined to get the result. They concluded that bagging, random forest and XGBoost appear to out-perform the other models as it has a smaller variance compared to other models. Bagging model exhibited the highest performance for the test data.

According to the investigative research paper [3], customer satisfaction is one of top factors for success of online platforms such as Airbnb. Data for the city of London was collected from the investigative website 'Inside Airbnb' [15]. One of the important pillars of increasing customer contentment is understanding the customer demographic. Using syntactic dependency parsing and string processing, customer reviews were divided into three different categories – 'individuals', 'couple', and 'families. This was done on basis on specific keywords in the reviews such as – 'I' appears for an individual review, 'my family' belongs to the family's category. For calculating the customer satisfaction, a total of six indicators were considered based on the review left by the customer namely – accuracy of the review, cleanliness at the property, the check-in experience, communication between the host and customer before and after successful booking, the geographical location and value of the rental unit. Regression algorithms such as Multiple Linear Regression (MLR) and the Gradient Boosting Regression (GBR), Artificial Neural Network were used. From the study it was concluded that the above six indicators are not enough to determine the precise level of satisfaction in the customer for the city of London. It also varied across gender, geographical, cultural, and psychological reasons [3]. Studies also showed that certain amenities such as Wi-Fi, free food and beverages also influenced the customer's perception regarding the booking. They concluded that customer demographic should be included in future results for better results.

As we can conclude for a city, that several factors affect the Airbnb price prediction, this study [4], which uses multi-modality data, gives us more insight . Multi-city, multi-factor data increase the accuracy of the price prediction model. The pricing of a rental unit depends heavily on its geographical location – e.g., its distance to city centres or commercial / financial centre, price increases shorter the distance. Analysis of three types of data was carried out – text data, numeric data, and map data. Also, to find the best prediction model with highest accuracy, it is necessary to work with different machine learning models along with different types of datasets. Around 9 million data points from 10 different cities between 2015 and 2019 was taken into consideration [15]. For the price prediction model, they tested four models:Linear Regression, Support Vector Regression, XGBoost and Deep Neural Network. It was found that XGBoost performed the best among them. Subsequently XGBoost was performed on different types of data. In single modality only listing data was used for price prediction. In dual modality, text data and geographical data was used along with listing data. In multi-modality data all three types of data are used. Using multi-modality data increases the accuracy of our price prediction model [4].

Airbnb dataset for the city of Beijing as of June 2020 was studied in [5]. For the price prediction model various features were selected. The distance from subway stations was also added to increase the accuracy of the model. It was calculated using Euclidean distance. House price data was also considered from Lianjia website [9]. XGBoost regression model was used to find the importance of various features and features of less importance were dropped. Amenities such television, internet connection were also studied as part of a price prediction model and how they influence the booking of an Airbnb rental unit. We can conclude from the above paper that many factors affect the prices such as geographical location, outliers, amenities, distance to transport hubs such as subway stations.

Various feature selection strategies such as filter, wrapper, embedded and hybrid, as well as a full discussion of the methodologies used by these methods was discussed in [6] A comparison of various feature selection approaches was also performed. Their study states the techniques utilised in the filter approach include information gain, chi square, fisher's score, and correlation and coefficient. Wrapper methods include sequential forward selection, sequential backward selection, and a genetic heuristic search algorithm. In embedded, only one technique was used: regularisation. A hybrid technique is basically a blend of techniques. Using both a filter and a wrapper, as well as an embedded and a wrapper.

They listed the advantages as well as drawbacks of all four methods. The filter approach is a quick and easy method that does not rely on the classifier. The disadvantage is that it does not take into account the dependencies between the features. Wrapper methods take into account feature dependencies and have a higher risk of overfitting, but they produce better results than the filter method. The embedded technique is more computationally intensive. Overfitting raises the weight of models. Hybrid approaches provide better accuracy and flexibility in interacting with other methods, which increases the complexity. This paper explains the significance of each method and strategy that can be utilized in feature selection.

III. Dataset Description And Pre-Processing

Our study consists of two datasets sourced from *Inside Airbnb* [15], an investigative website launched in 2016 which scraps data from the Airbnb website. The data is publicly available. The data is updated quarterly on the site. The first dataset consisting of listings information, is used for feature selection and is updated as of December 2021. It contains 125,798 rows combined from three cities around the world: Sydney, London, and New York.It lists around 74 attributes in each city. Some columns of our interest are property_type, baths, bedrooms, accommodates, and neighbourhood.

Our second dataset consists of customer reviews. It consists of around 148,539 rows and 6 columns and is from the same cities as above. Columns of interests contain listing_id and comments by the reviewers. The aim of selecting this dataset is for data mining purpose. To improve customer satisfaction, it is important to understand the things affecting the customers. We are focusing on reviews from latest available quarter i.e. September 2021 - December 2021. This helps us focus on the latest data as that would have greater importance over old data.

A. Data Cleaning

Data cleaning is a fundamental activity to be performed before preparing data for analysis. It helped us to produce better results and focus only on key terms and findings.

1) Listings Data: We performed data cleaning by deleting unnecessary columns such as id, host_id, that were not useful in our analysis and would have taken up space as well as likely slowed down runtime. We also checked for missing values in each column. A histogram was plotted for each city and from the histogram, the number of missing values across columns was learned. Some columns had 10%-15% percent of their data missing, while others had 60% of their values missing. Missing values were found in columns such as host_response_rate and host_response_time. Again, because these columns were not beneficial for our analysis and were resulting in incomplete data, we removed them. We decided to drop the columns with more than 20% of samples missing because we did not want to lose valuable information and considered that 20% would be safe. After deleting columns with 20% of missing values, we did an imputation for the rest of the data. For numerical columns, imputation for the missing value is done with the mean and for categorical columns imputation was performed with the mode.

2) Reviews Data: We performed data wrangling in order to prepare our data for sentiment analysis. The reviews were found to contain many special characters, emoji's and special symbols. We normalization on text and remove all characters expect alphabets and numbers. It has been found that sentiment analysis is performed better on whole sentences, therefore we will be keeping the stop words for now. They would be removed subsequently after analysis. We also remove non-English reviews using the *langdetect* package as we would focusing only on English reviews.

B. Exploratory Data Analysis

We only perform EDA on listings dataset as reviews dataset is text-only. We explore our analysis of price-related questions and get preliminary insights through visualization.We plotted a simple bar graph to understand the number of accommodates in each city. Fig.1 shows the distribution of the accommodates in Sydney city. It can be seen that most of the listings in Sydney accommodates two people. Similarly for the city of London in Fig.2, most of the listings prefer two people and also for New York in Fig.3, most of the listings have two accommodates.

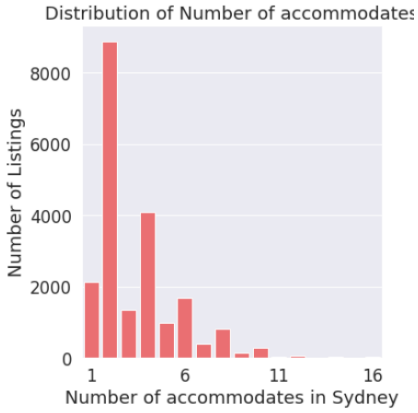

Fig. 1. Sydney Accommodations Capacity

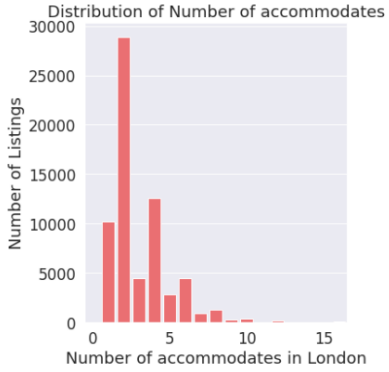

Fig. 2. London Accommodations Capacity

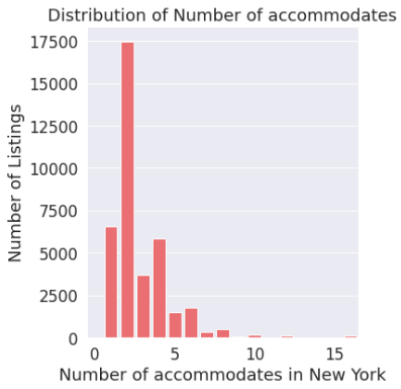

Fig. 3. New York Accommodations Capacity

Fig 4,5,6 shows the number of bedrooms available in each city's listings. According to the bar graph, Sydney, London, and New York all have the same number of bedrooms. The majority of the two-bedroom properties are available in all three cities.

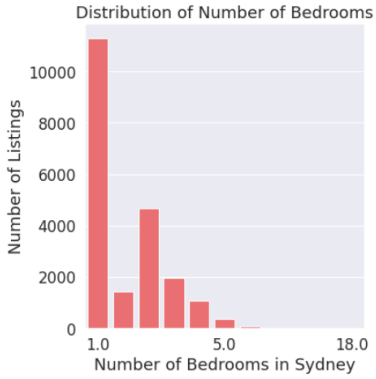

Fig. 4. Distribution of bedrooms in Sydney

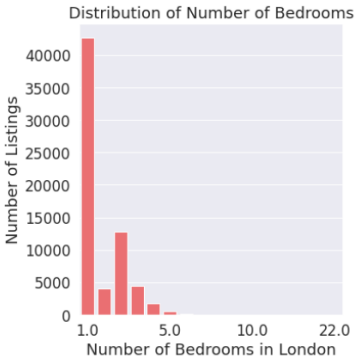

Fig. 5. Distribution of bedrooms in London

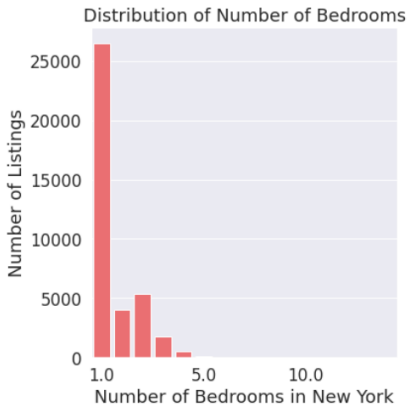

Fig. 6. Distribution of bedrooms in New York

From the insights it was found that every city has similarity, so to better understand we decided to find the distribution of price according to number of accommodates and number of bedrooms.

According to the EDA, all three cities have varied price patterns based on the number of people accommodations. Sydney has an increasing trend until 11 accommodates, after which it begins to decline. It has a rising trend in London till 9 accommodates. Following that, there is an unstable pattern with a dramatic increase in price for 16 rooms. whereas, New York is experiencing an upward trend.

Next, we visualized the pricing in relation to the number of bedrooms and discovered that Sydney offers listings with up to 18 bedrooms. The average price for a single bedroom is $92 and goes up to $4500 for an entire home or apartment with 18 bedrooms. Prices in London start at $60 for one bedroom and go up to $1056 for 22 bedrooms. There was one listing with 11 bedrooms for 6000$ that was an entire serviced apartment and another with 22 bedrooms that was an entire residential home. In New York, the average price for a 1 bedroom apartment is $90, and the average price for a 10 bedrooms is $1,643. An entire residential home is the property type with 10 bedrooms. The maximum amount of bedrooms, which is 16, is for an entire home/apartment, which costs $500 on average in New York.

Therefore, through EDA it is observed that, though the number of accommodates and number of bedrooms are same in number of listings but the prices differ for number of accommodate and number of bedrooms in Sydney, London and New York.

IV. Methodology

A. Feature Engineering

Feature engineering is a machine learning technique that uses data to produce new variables that are not in the training set. It can generate new features, with the goal of simplifying and speeding up data transformations while also improving model accuracy.

1) Categorical feature encoding: We first transformed categorical features to numerical values based on specific criteria before we can use them for modelling. This is referred to as encoding.

2) Encoding for binary categorical variables: Most Machine Learning algorithms are incapable of working with categorical input and must be transformed to numerical data. Some of the variables had t for true and f for false values. This is called binary outcome. Encoding was used to convert category responses to numerical. Above was the simple method to deal with binary outcomes but in order to deal with multiple categorical variable responses we used one hot encoding. One hot encoding is a procedure that converts categorical variables into a form that can be fed into ML algorithms to help them predict better.

In machine learning, one-hot encoding is used to quantify categorical data. In a nutshell, this method generates a vector with the same length as the number of categories in the data set. If a data point belongs to the i^{th} category, the vector's components are assigned the value 0, except for the ith component, which is assigned the value 1. This allows one to keep track of the categories in a numerically meaningful manner. In our case we, columns such as neighbourhood_cleansed, property_type, room_type were converted to numerical response using one hot encoding

B. Feature Selection

Feature selection techniques choose a subset of features while retaining the original data representation. They keep the original meaning of data and hence provide the benefit of interpretation. Furthermore, these strategies seek to prevent over-fitting by allowing for the generation of faster and more cost-effective models, hence improving performance [7]. Techniques used in our study are Filter Methods, Wrapper Method and Embedded Methods [6][9].

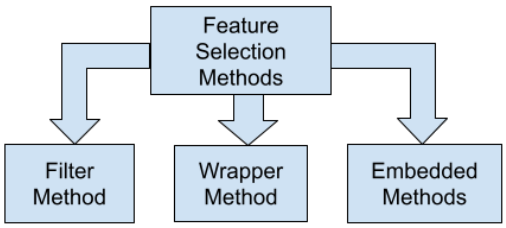

Fig. 7. Feature Selection Methods

1) Filter Methods

This method filters features based on general characteristics of the dataset (some measure such as correlation) such as correlation with the dependent variable.Filters extract features from data without using any learning method [7]. When the number of characteristics is large, it is usually the faster and best approach. Filter methods which we have used are, Removing Constant Features, Removing Quasi-Constant Features, Removing Duplicate Features and Removing Correlated Features [6].

a) Removing Constant Features: Constant features provide no useful information for further investigation. As a result, we removed them from the dataframe.

b) Removing Quasi-Constant Features: The features that are nearly constant are known as quasi-constant features [8]. We may now arbitrarily alter the variance threshold for constant features when we have set a variance threshold of 0. The process is nearly identical to the last one. It is suggested that the quasi-Constant characteristics be examined in the already reduced training and test data sets. Instead of supplying 0 as the threshold parameter value, we now passed 0.01, which implies that if the variance of the values in a column is less than 0.01, that column will be removed.

c) Removing Duplicate Features: Finally, we focus ourselves on the duplicate features. As a result, the approach differs slightly because no relevant function from the Scikit-learn library is available [8]. We got the duplicate features and those features were defined and removed from the dataframe.

d) Removing Correlated features: A dataset can have linked characteristics in addition to duplicate features. Two or more features are associated if they are close to each other in the linear space. Correlation between output observations and input features is crucial, and such features should be kept. However, if two or more features are mutually correlated, the model generates redundant information, and just one of the correlated features should be preserved to reduce the number of features [8]. To eliminate the correlated features, we used the pandas dataframe's corr() method. The corr() method returns a correlation matrix that contains correlation between all of the dataframe's columns. We may then loop through the correlation matrix and add that column to the list of correlated columns if the correlation between two columns is larger than the threshold correlation. That collection of columns can be removed from the actual dataset.

2) Wrapper Methods

Wrappers takes model hypothesis into account by training and testing in the feature space. Wrappers are more effective at selecting features. The wrapper method considers the feature dependencies[9]. We have used three techniques, step forward, step backward and recursive feature elimination [10].

a) Step Forward Feature Selection: Sequential feature selection methods are a type of greedy search tech-

nique that is used to reduce an initial d-dimensional feature space to a k-dimensional feature subspace, where k d is the dimension of the feature [8].It begins by assessing each feature independently and selecting the one that produces the highest performing algorithm based on a set of assessment criteria. The second phase assesses all potential combinations of the selected feature and a second feature and chooses the pair that produces the best performing algorithm based on the same pre-set criteria.

For regression, the r squared is the pre-set criteria, because it assesses every potential single, double, triple, and so on feature combinations, this selection technique is known as greedy. As a result, it is computationally expensive and, in certain cases, impractical if the feature space is large. This form of feature selection is implemented using a special Python module called mlxtend [8]. The halting criteria in the mlxtend implementation of the step forward feature selection is an arbitrarily specified amount of features. So the search will end once we have reached the appropriate number of features. This is fairly arbitrary because we might choose a suboptimal amount of features or a large number of features.

b) Step Backward Feature Selection
In step backward, fitting a model with all characteristics is the first step in feature selection. Then it eliminates one feature. It will eliminate the one that delivers the best algorithm for a given set of assessment criteria. It will then eliminate a second feature, this time the one that provides the best performing algorithm. And so it goes, deleting feature after feature until a particular threshold is reached. This approach is also known as greedy since it examines every conceivable n, then n-1, n-2, and so on feature combinations [8]. As a result, it is computationally expensive and, in certain cases, impractical if the feature space is large. The pre-set criteria for this is also r squared.

c) Recursive feature elimination The steps in this method are as follows:
1) Sort the features by importance as determined by a machine learning algorithm: it might be tree importance, LASSO / Ridge, or linear / logistic regression coefficients. According to our dataset we have used linear regression coefficients
2) Remove one feature (the least important) and construct a machine learning method using the remaining features.
3) Determine a performance metric of your choice, such as roc-auc, mse, rmse, or accuracy.
4) If the measure falls below an arbitrarily specified threshold, that feature is valuable and should be retained. Otherwise, we can do away with that feature.
5) Repeat steps 2-4 until all features have been eliminated (and thus evaluated) and the performance drop has been assessed.
The difference between this method and the step backwards feature selection method is that it does not remove all characteristics before deciding which one to eliminate. It removes the least important one, as determined by the machine learning model [8]. After that, it determines whether or not that feature should be eliminated. As a result, it removes each feature just once throughout the selection process, whereas step backward feature selection removes all features at each step of the selection process[8].

3) Embedded Methods:

Embedded approaches incorporate the feature selection procedure into the model training process. They are usually faster than wrapper approaches and can give an appropriate feature subset for the learning algorithm [9]. Our research employs Lasso regularization and tree-based approaches [11][12].

a) Lasso Regularization Regularisation entails applying a penalty to the various parameters of the machine learning model in order to decrease the model's freedom and therefore avoid overfitting. The penalty is applied over the coefficients that multiply each of the predictors in linear model regularisation. Among the several kinds of regularisation, Lasso or l1 has the ability to reduce some of the coefficients to zero. As a result, that feature can be eliminated from the model [8].
b) Tree Based Method- Random Forest Regressor
One of the most common machine learning algorithms is random forests. So, we decided to go on with this too. They are so effective because they provide strong prediction accuracy in general, little overfitting, and simple interpretability. This interpretability is provided by the ease with which each variable's influence on the tree decision may be determined. In other words, it is simple to determine how much each variable contributes to the decision [11].
We used this selection strategy to get the feature importance. It assigned a score to each aspect of data, with the higher the score indicating that the information is more valuable or relevant to price attribute.

C. Modelling

The data is separated into two groups: x_train, _train, and x_test, y test. Price is the target variable i.e., 'y' and rest

are the independent variables i.e., 'X'. Data is split into 80% training data and 20% testing data. Skleran is used to import the algorithms model. model.fit is used to construct the model. Model.predict is used to make predictions after the model has been built using the above process (x_test). We chose regression techniques because the data is numerical, and the base models used are Random Forest and XGBoost Regressor.

D. Evaluation

Evaluation metrics used in mean squared error and R2 score [1][3].

1) Mean Squared Error (MSE): MSE is the most commonly used statistic for regression problems.It is the squared difference between the predicted and actual value. It is easy to optimise because it is differentiable and has a convex shape. Large mistakes are penalised by MSE. MSE formula is given by:

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}(\hat{yi} - yi)^2 \tag{1}$$

2) R2 Score: The R2 score is a significant indicator for evaluating the performance of a regression-based machine learning model. It is also known as the coefficient of determination. It works by calculating the amount of variance in the predictions explained by the dataset. Simply described, it is the difference between the samples in the dataset and the predictions provided by the model. R2 score is a measure that indicates the performance of your model, rather than the absolute loss of how many wells your model performed. It is the ratio of the sum of squares and the total sum of squares given by:

$$R2 = 1 - \frac{SSE}{SST} \tag{2}$$

where SSE is the square of the difference between the actual and predicted values

$$SSE = \Sigma_{i=1}^{n}(yi - \hat{yi})^2 \tag{3}$$

where SST is the entire sum of the square of the difference between the actual value and its mean.

$$SST = \Sigma_{i=1}^{n}(yi - \bar{yi})^2 \tag{4}$$

The observed target value is yi, the anticipated value is ŷi the mean value is y-bar, and the total number of observations is n.

E. Sentiment Analysis

Sentiment Analysis is use of natural language processing, text analysis to process information such as customer reviews, surveys to extract useful knowledge or facts. In the hospitality industry it is of utmost importance because good/bad reviews can affect the perception of future potential customers. Sentiment Analysis is performed using various techniques such as normalization, bag-of-words. We use NLTK (Natural Language Toolkit) for performing sentiment analysis. It is a suite of libraries dedicated to research in NLP (Natural Language Processing).

We perform analysis Using the 'SentimentIntensityAnalyzer' function. We use *vader_lexicon*[17] through NLTK. It is lexicon and rule-based analysis which tells us whether the words are positive or negative and how strong is the positivity or negativity. Vader internally produces four sentiment scores which initially produces three - positive,negative or neutral. The final score is the compound score which gives us the normalized score between -1 and 1. We divide the results obtained into positive and negative results using the compounded score. In order to extract keywords we use *Rake*[18] through NLTK which stands for Rapid Automatic Keyword Extraction. This library also removes all English stopwords from the text. After that we fetch keywords from reviews to understand the actual problems for negative reviews and good solutions from positive reviews. Subsequently we can fetch listings with most negative review or most positive reviews to take necessary action.

V. Results

A. Feature Selection

According to above results, Recursive Elimination with XGBoost Regressor worked well with R2 score of 61% and less mean squared error of 1193.48 when compared to other feature selection methods.

Model/Methods	Mean Squared Error	R2 Score
Filter method implemented using XGBoost Regressor	1262.09	0.59
Forward selection Using Random Forest Regressor	1674.95	0.58
Backward elimination Using Random Forest Regressor	1732.41	0.58
Recursive Feature elimination using XGBoost Regressor	1193.48	0.61
Lasso Regularization	1469.63	0.53
Tree Based method using Random Forest	1408.75	0.54

TABLE I
Results of Model Parameters

Top 5 features that affect the pricing in Sydney are private room type, shared room type, listing count for shared rooms, entire rental property and number of accommodates. Similarly, top 5 features in London are bedrooms, Westminster neighbourhood, Kensington and Chelsea neighbourhood, accommodates and number of host listings.For New York, top 5 features are private room type, property type such as hotel room, accommodates, room in boutique hotel and bath.

B. Sentiment Analysis

Using the results from the analysis, we can extract the necessary keywords as follows. Above results are shown only for a couple of reviews. It can be replicated on other reviews as well. This results can then be utilized to train new language processing models specially catered to customer reviews in the hospitality industry.

Sentiment Score	Keywords
0.9986	'good safe area within walking distance', 'belongings ilgi provided clean snowwhite linens', '2 tube stations shops cafes', 'friendly sweet super caring hostess', 'wellgroomed fabulous garden squirrels', 'calm friendly caring environment', 'beautiful tastefully decorated house',
0.9985	'computer extra shower items kitchen items although', 'readily available iron gym equipment', 'open late walkable distance', 'cooking etc portobello market', 'many lovely cafes bakeries', 'ate great street food', 'get fresh air'

TABLE II
POSITIVE KEYWORDS

Sentiment Score	Keywords
- 0.9872	'find another host avoid', 'damage something already broken', 'sit outsidethe location', 'rubbish outdoor chairs', 'protentional back charge', 'clutter couldnt access', 'came apart therefore'
- 0.9929	'new chairs toilet doesnt flush unless', 'inside felt really insecure took', 'kill x 5 cockroaches', 'one huge sheet pane', 'felt really uncomfortable sleeping', '3 night stay mattress', 'legs back panel broken'

TABLE III
NEGATIVE KEYWORDS

VI. CONCLUSION

According to the top 5 features from three cities, the number of accommodates attributes is common in three cities which means prices highly depend on the number of people. Short-term lettings have expanded and will be a large industry in the future, according to the Australian Housing and Urban Research Institute, and Sydney's eastern suburbs have been a focus for Airbnb activities [13]. Our findings from Sydney are comparable. The host listing count for shared rooms is one of the top 5 attributes, and if we look further down the graph, we can see neighbourhoods such as Manly, East Village, and Canterbury, which are located in the east of Sydney. Since these are beaches and tourist destinations, there are a lot of short-term rentals, and people prefer shared rooms. Observing the London results, we can see that neighbourhoods are among its top features, which indicates that regions like Westminster, Kensington, and Chelsea have an impact on London prices. If new hosts have property in certain locations, this can enable them to choose their listing price. Referring to a world atlas article, housing is the most expensive basic expense in New York, including both rental apartments and home purchases. As a result, three of the top five attributes are related to room type and property type, as per our findings.As a result, these insights were obtained using a recursive elimination method with the XGBoost Regressor, which proved to be the best approach for our research problem.

From our study, we can conclude that customer experience can be enhanced using NLP methods such sentiment analysis. More complex approaches using additional factors such as linking the customer demographic with the reviews can be developed in future which will help use to cater accordingly. Thus,using a combination of above methods we can try to improve Airbnb. Our code for the study can be found at [21]

REFERENCES

[1] J. Dhillon et al., "Analysis of Airbnb Prices using Machine Learning Techniques," 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 2021, pp. 0297-0303, doi: 10.1109/CCWC51732.2021.9376144.

[2] A. Zhu, R. Li and Z. Xie, "Machine Learning Prediction of New York Airbnb Prices," 2020 Third International Conference on Artificial Intelligence for Industries (AI4I), 2020, pp. 1-5, doi: 10.1109/AI4I49448.2020.00007.

[3] M. Chiny, O. Bencharef and Y. Chihab, "Towards a Machine Learning and Datamining approach to identify customer satisfaction factors on Airbnb," 2021 7th International Conference on Optimization and Applications (ICOA), 2021, pp. 1-5, doi: 10.1109/ICOA51614.2021.9442657.

[4] N. Peng, K. Li and Y. Qin, "Leveraging Multi-Modality Data to Airbnb Price Prediction," 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME), 2020, pp. 1066-1071, doi: 10.1109/ICEMME51517.2020.00215.

[5] S. Yang, "Learning-based Airbnb Price Prediction Model," 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT), 2021, pp. 283-288, doi: 10.1109/ECIT52743.2021.00068.

[6] A. Kaur, K. Guleria and N. Kumar Trivedi, "Feature Selection in Machine Learning: Methods and Comparison," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021, pp. 789-795, doi: 10.1109/ICACITE51222.2021.9404623.

[7] S. D'Souza, P. K. V. and B. S, "Feature Selection and Modeling using Statistical and Machine learning Methods," 2020 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), 2020, pp. 18-22, doi: 10.1109/DISCOVER50404.2020.9278093.

[8] Feature Selection,https://github.com/obaidM/FeatureSelectionPython

[9] Aparna U.R. and S. Paul, "Feature selection and extraction in data mining," 2016 Online International Conference on Green Engineering and Technologies (IC-GET), 2016, pp. 1-3, doi: 10.1109/GET.2016.7916845.

[10] Y. T. Naing, M. Raheem and N. K. Batcha, "Feature Selection for Customer Churn Prediction: A Review on the Methods & Techniques applied in the Telecom Industry," 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), 2022, pp. 1-5, doi: 10.1109/ICDCECE53908.2022.9793315.

[11] J. K. Jaiswal and R. Samikannu, "Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression," 2017 World Congress on Computing and Communication Technologies (WCCCT), 2017, pp. 65-68, doi: 10.1109/WCCCT.2016.25.

[12] R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), 2016, pp. 18-20, doi: 10.1109/ICACA.2016.7887916.

[13] Crommelin, Laura and Troy, Laurence and Martin, Chris and Parkinson, Sharon, Technological Disruption in Private Housing Markets: The Case of Airbnb (October 8, 2018). AHURI Final Report No. 305, Australian Housing and Urban Research Institute Limited, Melbourne, DOI:10.18408/ahuri-7115201, Available at SSRN: https://ssrn.com/abstract=3280620

[14] Moon, Hyoungeun & Miao, Li & Hanks, Lydia & Line, Nathaniel. (2019). Peer-to-peer interactions: Perspectives of Airbnb guests and hosts. International Journal of Hospitality Management. 77. 405-414. 10.1016/j.ijhm.2018.08.004.

[15] "How is Airbnb really being used in and affecting the neighbourhoods of your city?", http://insideairbnb.com (accessed Jan 15, 2022)

[16] Deane, Steve, "2022 Airbnb Statistics: Usage, Demographics, and Revenue Growth", https://www.stratosjets.com/blog/airbnb-statistics/ (accessed Jun 1,2022)

[17] "Natural Language Toolkit", https://www.nltk.org/howto/sentiment.html (accessed May 1,2022)

[18] Rake NLTK, https://github.com/csurfer/rake-nltk

[19] Chinese real-estate brokerage company, http://bj.lianjia.com

[20] https://www.worldatlas.com/articles/the-most-expensive-cities-in-the-world.html

[21] Airbnb Data Analysis, https://gitlab.computing.dcu.ie/athavab2/2022-mcm-TERMPLATE/-/tree/master/src