**Sentiment analysis and Price prediction**

**For Airbnb**

**Name- Shivani Sharma**

**Student Id – 0781657**

**Table Of Content**

[1. Abstract 3](#_Toc109333890)

[2. Literature Review 4](#_Toc109333891)

[3.1 Data Collection 7](#_Toc109333892)

[3.2 Data Cleaning 9](#_Toc109333893)

[3.2.1 Cleaning Null Values 9](#_Toc109333894)

[3.2.2 Summary Statistics 10](#_Toc109333895)

[3.2.3 Checking and treatment of Outliers 11](#_Toc109333896)

[3.2.4 Exploratory Data Analysis 16](#_Toc109333897)

[3.3 Text Pre-Processing 17](#_Toc109333898)

[3.3.1 Removing Contractions 17](#_Toc109333899)

[3.3.2 Removing Noise 18](#_Toc109333900)

[3.3.3 Converting to Lower 18](#_Toc109333901)

[3.3.4 Extracting English and emoji reviews 18](#_Toc109333902)

[3.3.5 Calculating Polarity score 18](#_Toc109333903)

[3.4 Word cloud for different Emotions 18](#_Toc109333904)

[3.5 Feature Engineering 21](#_Toc109333905)

[3.5.1 Correlation of independent variable with dependent variables 21](#_Toc109333906)

[3.5.3 Data Normalization 22](#_Toc109333907)

[3.5.4 Feature Selection 22](#_Toc109333908)

[3.6 Data Modelling (Price Prediction) 23](#_Toc109333909)

[3.6.1 Linear Regression 23](#_Toc109333910)

[3.6.2 Decision Tree Regressor 24](#_Toc109333911)

[3.6.4 Random Forest regressor 24](#_Toc109333912)

[3.6.5 Price Prediction with Sentiment Score 25](#_Toc109333913)

[3.6.6 Feature Importance 27](#_Toc109333914)

[3.6.7 Price Prediction without Sentiment Score 29](#_Toc109333915)

[3.7 Result Evaluation 31](#_Toc109333916)

[4.Conclusion 32](#_Toc109333917)

[5 References 33](#_Toc109333918)

**List Of Figures**

[Figure 1 - Methodology 8](#_Toc109662887)

[Figure 2 - Outlier Detection in Price 13](#_Toc109662888)

[Figure 3 - Outlier Treatment in Price 13](#_Toc109662889)

[Figure 4 - Outlier Detection in Minimum Nights 14](#_Toc109662890)

[Figure 5 - Outlier Treatment in Minimum Nights 14](#_Toc109662891)

[Figure 6 - Outlier Detection in Number of reviews 14](#_Toc109662892)

[Figure 7 - Outlier Treatment in Number of reviews 15](#_Toc109662893)

[Figure 8 - Outlier Detection in Reviews per month 15](#_Toc109662894)

[Figure 9 - Outlier Treatment in Reviews per month 15](#_Toc109662895)

[Figure 10 - Outlier Detection in Calculated host listings count 16](#_Toc109662896)

[Figure 11 - Outlier Treatment in Calculated host listings count 16](#_Toc109662897)

[Figure 12 - Outlier Detection in Number of reviews ltm 16](#_Toc109662898)

[Figure 13 - Outlier Treatment in Number of reviews ltm 17](#_Toc109662899)

[Figure 14 - Price Distribution based on room type 17](#_Toc109662900)

[Figure 15 - Top neighbourhood and their average price 18](#_Toc109662901)

[Figure 16 - Reviews and Average Price for different hotels 19](#_Toc109662902)

[Figure 17 - Emoji Word Cloud 21](#_Toc109662903)

[Figure 18 - Positive Word Cloud 21](#_Toc109662904)

[Figure 19 - Negative Word Cloud 22](#_Toc109662905)

[Figure 20 - Negative Word Cloud 22](#_Toc109662906)

[Figure 21 - Correlation graph 23](#_Toc109662907)

[Figure 22 - Linear Regression Prediction 27](#_Toc109662908)

[Figure 23 - KNN Regressor Prediction 28](#_Toc109662909)

[Figure 24 -Decision Tree Regressor Prediction 28](#_Toc109662910)

[Figure 25 - Random Forest Regressor Prediction 29](#_Toc109662911)

[Figure 26 - Linear Regression Feature Importance 30](#_Toc109662912)

[Figure 27 - Decision Tree Regression Feature Importance 30](#_Toc109662913)

[Figure 28 - Random Forest Regression Feature Importance 30](#_Toc109662914)

[Figure 29 - Linear Regression Prediction 31](#_Toc109662915)

[Figure 30 - KNN Regressor Prediction 32](#_Toc109662916)

[Figure 31 - Decision Tree Regressor Prediction 32](#_Toc109662917)

[Figure 32 - Random Forest Regressor Prediction 33](#_Toc109662918)

[Figure 33 - Distribution of Sentiments 34](#_Toc109662919)

[Figure 34 - Distribution of sentiments over years 34](#_Toc109662920)

**List Of Tables**

[Table 1- Details for Listing Dataset 9](#_Toc109662921)

[Table 2 - Description for Review Dataset 11](#_Toc109662922)

[Table 3 - Null Value Count Before Cleaning 11](#_Toc109662923)

[Table 4 - Summary of numerical attributes 12](#_Toc109662924)

[Table 5 - Summary of Categorical attributes 13](#_Toc109662925)

[Table 6 - Feature Selection 24](#_Toc109662926)

[Table 7 - RMSE value calculation 27](#_Toc109662927)

[Table 8 - Calculation of RMSE 31](#_Toc109662928)

# **Abstract**

Sentiment analysis is an analytical technique that determines the emotional meaning of communications by combining statistics, natural language processing, and machine learning. Customer messages, call Centre encounters, online reviews, social network posts, and other information is all evaluated using sentiment analysis. Changes in opinions regarding organizations, products, or services, as well as individual elements of such items or services,

can be tracked using sentiment analysis.

For my Capstone Project I have decided to work on sentimental analysis using NLP (Natural Language Process). Also, I will be visualizing the data with the help of Power BI to understand the data more. I would be doing exploratory analysis on the listing dataset and sentimental analysis on the review’s dataset. After that I will do classification by using sentimental dataset output on the listing dataset. I will calculate that the customers reviews about the hotel that whether its positive, negative, or neutral. Also, I will check that whether the hotel reviews increased or decreased over time. And in future the hotel will be selected by customers or not based on reviews which they got from other customers. Once the reviews are analysed as positive, negative, or neutral next step would be to predict the price for different Airbnb hotels based on the different features.

**Keywords:** - *Machine learning (Linear Regression, Decision Tree Regressor, Random Forest regressor, KNN regressor), Libraries (NumPy, Pandas, Matplotlib, Seaborn, nltk, word cloud, plotly, sklearn), Regex, Sentiment Analysis*

**Tools: -** Python, Power BI.

**GitHub Source**: - <https://github.com/Shivani2101/Capstone-Project.git>

**Business Questions:**

1. What are the customers reviews about the hotel that whether its positive, negative, or neutral?
2. To check whether positive, negative reviews are increased or decreased over time?
3. Predicting price for different Airbnb hotels based on different parameters?

# **Literature Review**

Lot of research has been made earlier on the Airbnb dataset to predict the price of the hotels based on different scenarios using machine learning algorithms. A price prediction model using algorithms (Yuanhang Luo, Xuanyu, Zhou, and Yulian Zhou (2019)) Random Forest, XGBoost, and Neural Network was made and features like host-id and name were ignored to remove noise and features like number of rooms were taken to build the model and make predictions. XGBoost and Neural Network outperform other models after substantial feature engineering and extraction on the New York and Paris datasets. The outcome is examined using R-Squared and Median Squared Error (MSE). XGBoost has an R-Squared error of 0.722, while Neural Networks has an R-Squared error of 0.769.

Another prominent implementation of Airbnb price prediction included photos as well as textual data to construct the model. (Emily Tang and Kunal Sangani (2015)) used an image-rich dataset of San Francisco real estate listings. Select listing information, multinomial bag of words features, text sentiment features, and visual features are among the features extracted. They used a bag of words approach to extract visual elements from listing photographs, which was a novel notion. Descriptors are extracted from photos using SURF (Speeded Up Robust Features). They constructed a visual word dictionary and feature vectors for each item based on the image descriptors. They only used a linear kernel Support Vector Machine (SVM) to examine the performance and accuracy of each.

Additional features utilising external resources were studied in another similar research. Longitude and zip code are typically left out of most research because they have minor impact on forecast accuracy. To increase efficiency, (QuangTrung Nguyen) computes the number of nearby tourist attractions using location information. Other than using simply numerical data from the dataset, all three studies stated above employed supplementary elements such as textual data, photographs, and location.

Customer reviews, we believe, are another significant feature to consider when developing a pricing predictor. Sentiment analysis has been always handled as a Natural Language Processing tasks on many different levels, starting from the classification task. It has been handled at different levels starting with sentence level (Hu and Liu, Kim & Hovy, (2004)) and at phrase level (Agarwal et al., 2009). Twitter has humongous data and which data analysis and specific topic – sentiment analysis is getting performed on the real data which is based to check the emotions of people and in the very beginning sentiment analysis on the twitter data was challenging and gathering the data and making insights was difficult, whereas some researchers used the emoticons data for analysis. The emoticon at the end of the tweet was positive than the tweet was said to be positive, otherwise negative. They use Naive Bayes, MaxEnt, and Support Vector Machines (SVM) to create models, and they claim that SVM outperforms other classifiers. They use a Unigram, Bigram model in conjunction with parts-of-speech (POS) features in terms of feature space. The unigram model surpasses all other models, they say. Bigrams and POS characteristics are ineffective.

(Barbosa and Feng (2010)) create two distinct classifiers, one for Subjective versus Objective categories and the other for Positive versus Negative categories. They provide independent evaluations for both models, but they do not investigate integrating them or comparing them to one another. Three-way classification system More recently, (Jiang et al., (2011)) report on the development of a three-way network. Twitter classifier that distinguishes between Objective, Positive, and Negative messages. They do not, however, investigate the cascaded effects. Neutral tweets are not detected by the design. The trade-off between making fewer

predictions and F1-measure have been studied in the literature. Like Machine predictions, like human annotations, have confidence levels.

The accuracy of the classifier is improved by pre-processing textual data to reduce noise. Nave Bayes with SVM, according to (Bo Pang and Lillian Lee (2004)), is the most efficient method for offering the highest accuracy. (Xiangjie Kong, Huuizhen Jiang, Zhuo Yang, Zhhenzhen Xu, feng Xia, and Amr Tobla (2016)) adopted the Random Walk model for academic domain correctness. When compared to other techniques, (Huakang Li, Yixiong Bian, Xiuying Xu, and Guozi Sun (2017)) proposed the Monte Carlo decision tree algorithm for mining interest similarity. NB, according to (G. Vinodhini (2012)), is the best technique for determining document quality. (Faruk and Arnab (2016)) developed a methodology for high-accuracy trust management. SVM and KNN were proposed by (Soudamini Hota and Sudhir Pathak) as the best methods for addressing noisy data I textual information.

**3.Methodology:**

For performing sentiment analysis and price prediction on the Airbnb dataset, there are many steps that are to be taken in sequence to reach to a proper conclusion and create an algorithm which can be best suitable to answer the business needs. Below are the steps which are to be performed-

Diagram

Description automatically generated

**Figure 1 - Methodology**

## **3.1 Data Collection**

Data has been taken from Airbnb open source (Inside Airbnb). The dataset corresponds to review and listings dataset for Airbnb listed in Canada Toronto Region. The listing dataset contains nearly 18 features and 15418 records (The details of listing dataset is mentioned below in Table 1). On the other hand, review dataset contains 6 features and 400313 records (The details of listing dataset are mentioned below in Table 2).

**Table 1- Details for Listing Dataset**

|  |  |  |
| --- | --- | --- |
| Field | Data Type | Description |
| Id | int64 | Airbnb's unique identifier for the listing |
| name | object | Name of the listing |
| host\_id | int64 | Airbnb's unique identifier for the host/user |
| host\_name | object | Name of the host. Usually just the first name(s). |
| neighbourhood\_group | float64 | The neighbourhood group as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefile |
| Neighbourhood | object | The neighbourhood group as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles. |
| latitude | float64 | Uses the World Geodetic System (WGS84) projection for latitude and longitude. |
| longitude | float64 | Uses the World Geodetic System (WGS84) projection for latitude and longitude. |
| room\_type | object | "[Entire home/apt|Private room|Sharedroom|Hotel]  All homes are grouped into the following three-room types:  Entire place Private room Shared room Entire place |
| price | int64 | daily price in local currency |
| minimum\_nights | int64 | minimum number of nights stay for the listing (calendar rules may be different) |
| number\_of\_reviews | int64 | The number of reviews the listing has |
| last\_review | object | The date of the last/newest review |
| reviews\_per\_month | float64 | The number of reviews the listing has over the lifetime of the listing |
| calculated\_host\_listings\_count | int64 | The number of listings the host has in the current scrape, in the city/region geography. |
| availability\_365 | int64 | avaliability\_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host. |
| number\_of\_reviews\_ltm | int64 | The number of reviews the listing has (in the last 12 months) |
| license | object | The licence/permit/registration number |

**Table 2 -Description for Review Dataset**

|  |  |  |
| --- | --- | --- |
| Field | Data Type | Description |
| listing\_id | int64 | Id related to each listing present in the listing dataset |
| reviewer\_id | int64 | Id related to each reviewer |
| date | datetime | Date of the day when the review is given |
| reviewer name | object | Name of the reviewer |
| Comments | object | Comments given to each hotel by reviewers |

## **3.2 Data Cleaning**

### **3.2.1 Cleaning Null Values**

**Table 3 - Null Value Count Before Cleaning**

|  |  |
| --- | --- |
| Field Name | Null Value Count |
| Name | 2 |
| host\_name | 6 |
| neighbourhood\_group | 15418 |
| last\_review | 3437 |
| reviews\_per\_month | 3437 |
| license | 10075 |

In data cleaning the data is analysed statistically to understand all the attributes better to get better predictions. Firstly, the null values are checked for the listings dataset, and it was seen that the listing dataset contains nearly over 32000 null values (The details can be seen in Table3). There are various methods through which null values can be handled like using imputing using sklearn. The null values for neighnourhood\_group, licence, last\_review was handled by dropping the attributes as it is not putting any weightage to the analysis. Furthermore, the null values for name and host\_name, Reviews\_per\_month were replaced by None and 0 respectively.

### **3.2.2 Summary Statistics**

Summary statistics is used to check the descriptive analytics that includes the minimum, maximum, count, mean, standard deviation and quartile values for each of numerical attributes (The summary of numerical attributes can be seen in Table 4). Similarly, the summary statistics of categorical data includes the details count, unique, top, frequency (The summary statistics for categorical attribute can be seen in Table 5).

**Table 4 - Summary of numerical attributes**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | Max |
| id | 15418 | 2.49E+16 | 1.15E+17 | 1419 | 18925773.75 | 33777334 | 46609261.25 | 5.77E+17 |
| host\_id | 15418 | 128837316 | 126259050 | 1565 | 22378000.5 | 82723935.5 | 214030250 | 447806792 |
| latitude | 15418 | 43.68227325 | 0.049478885 | 43.58574637 | 43.64557 | 43.663535 | 43.7047775 | 43.84376 |
| longitude | 15418 | -79.39758375 | 0.06656804 | -79.63271 | -79.4265225 | -79.39692641 | -79.37617 | -79.11718 |
| price | 15418 | 159.0386561 | 370.0909273 | 0 | 69 | 106 | 169 | 13000 |
| minimum\_nights | 15418 | 25.76008561 | 38.58551747 | 1 | 4 | 28 | 28 | 1125 |
| number\_of\_reviews | 15418 | 26.00116747 | 53.29536008 | 0 | 1 | 5 | 26 | 828 |
| reviews\_per\_month | 11981 | 1.059303063 | 1.486324052 | 0.01 | 0.15 | 0.48 | 1.32 | 13.84 |
| calculated\_host\_listings\_count | 15418 | 5.559605656 | 14.44982848 | 1 | 1 | 1 | 4 | 133 |
| availability\_365 | 15418 | 132.7499676 | 133.7070044 | 0 | 0 | 88 | 258 | 365 |
| number\_of\_reviews\_ltm | 15418 | 4.826501492 | 12.45906547 | 0 | 0 | 0 | 3 | 170 |

**Table 5 - Summary of Categorical attributes**

|  |  |  |  |
| --- | --- | --- | --- |
|  | name | host\_name | neighbourhood |
| count | 15418 | 15418 | 15418 |
| unique | 15099 | 4783 | 140 |
| top | Downtown Toronto Little Italy Trinity Bellwood | Sky View | Waterfront Communities- The Island |
| freq | 7 | 133 | 2721 |

### **3.2.3 Checking and treatment of Outliers**

For the univariate analysis the outliers in each of the numerical attributes would be checked and removed or replaced as per the requirement. The treatment of outlier is important as this a noise that can affect the prediction.

**Outlier Detection and Removal for Price**

Chart, box and whisker chart

Description automatically generated

**Figure 2 - Outlier Detection in Price**

Histogram and Boxplot was drawn for Price attribute to check the outlier. As there are outliers present in the attribute and it is right skewed. So, it can be replaced with the upper bound value with the IQR method.

Chart

Description automatically generated

**Figure 3- Outlier Treatment in Price**

Above is the Histogram and boxplot for Price attribute after replacing the outliers with upper bound value.

**Outlier Detection for Minimum Nights**

Chart

Description automatically generated

**Figure 4 - Outlier Detection in Minimum Nights**

Histogram and Boxplot was drawn for Minimum Nights attribute to check the outlier. As there are outliers present in the attribute and it is right skewed. So, it can be replaced with the upper bound value with the IQR method.

Chart, box and whisker chart

Description automatically generated

**Figure 5 - Outlier Treatment in Minimum Nights**

Above is the Histogram and boxplot for Minimum Nights attribute after replacing the outliers with upper bound value.

**Outlier Detection for Number of reviews**

Chart, box and whisker chart

Description automatically generated

**Figure 6 - Outlier Detection in Number of reviews**

Histogram and Boxplot was drawn for Number of reviews attribute to check the outlier. As there are outliers present in the attribute and it is right skewed. So, it can be replaced with the upper bound value with the IQR method.

Chart, histogram

Description automatically generated

**Figure 7 - Outlier Treatment in Number of reviews**

Above is the Histogram and boxplot for Number of reviews attribute after replacing the outliers with upper bound value.

**Outlier Detection for Reviews per month**

Chart

Description automatically generated

**Figure 8 - Outlier Detection in Reviews per month**

Histogram and Boxplot was drawn for Reviews per month attribute to check the outlier. As there are outliers present in the attribute and it is right skewed. So, it can be replaced with the upper bound value with the IQR method.

Chart

Description automatically generated

**Figure 9 - Outlier Treatment in Reviews per month**

Above is the Histogram and boxplot for Reviews per month attribute after replacing the outliers with upper bound value.

**Outlier Detection for Calculated host listings count**

Chart, box and whisker chart

Description automatically generated

**Figure 10 - Outlier Detection in Calculated host listings count**

Histogram and Boxplot was drawn for Calculated host listings count attribute to check the outlier. As there are outliers present in the attribute and it is right skewed. So, it can be replaced with the upper bound value with the IQR method.

Chart

Description automatically generated

**Figure 11 - Outlier Treatment in Calculated host listings count**

Above is the Histogram and boxplot for Calculated host listings count attribute after replacing the outliers with upper bound value.

**Outlier Detection for Number of reviews ltm (last twelve months)**

Chart

Description automatically generated

**Figure 12 - Outlier Detection in Number of reviews ltm**

Histogram and Boxplot was drawn for Number of reviews ltm attribute to check the outlier. As there are outliers present in the attribute and it is right skewed. So, it can be replaced with the upper bound value with the IQR method

Chart

Description automatically generated

**Figure 13 - Outlier Treatment in Number of reviews ltm**

Below is the Histogram and boxplot for Number of reviews ltm attribute after replacing the outliers with upper bound value.

### **3.2.4 Exploratory Data Analysis**

**Distribution of price based on room-type**

Chart, box and whisker chart

Description automatically generated

**Figure 14 - Price Distribution based on room type**

From the above graphs the maximum price as per the room type is for Entire home/apt which is nearly 160 dollars followed by Private room with a price of nearly 80 dollars. Lastly, the prices for hotel room and shared room are between 40 to 60.

**Top neighbourhood and their average price**

Chart

Description automatically generated

**Figure 15 - Top neighbourhood and their average price**

From the pie chart, Waterfront communities the island is mostly distributed among all the

neighbourhoods with 78.62%, followed by Niagara which covers nearly 18% and rest 2% is

covered by rest of the neighbourhoods. On the other hand, the line chart shows the average

price of each neighbourhood in which Bridle Path-Sunnybrook-York Mills has highest price over 200 dollars and rest of the neighbourhoods has price below 200 dollar

**Reviews and Average price for different hotels**

Chart

Description automatically generated

**Figure 16 - Reviews and Average Price for different hotels**

From the above line graph, total reviews, and monthly reviews for Private-Bedroom in Toronto North York #1 and DWTN Hotel Style Apt up to 4 people are same which is nearly 175 and 10 respectively. On the other hand, in last twelve months Private Bedroom Toronto North York #1 has more reviews as compared to DWTN hotel Style Apt up to 4 people which is approximately 20.

On the other hand, in the bar graph it can be seen thar Minutes fr Eaton Centre & Dundas Sq has the highest price nearly 250 dollars followed by 2 BDRM + Sofabed - Ent. District and 40th floor perfect view of CN tower and the Lake which has price approximately 200 dollars

## **3.3 Text Pre-Processing**

This process includes cleaning of the textual information like removing punctuations, removing contractions, links, hashtags etc.

### **3.3.1 Removing Contractions**

Contractions includes words like isn’t which have same meaning if written like is not. In terms of removing punctuation the isn’t would be converted to different words with different meaning. Hence it becomes important to remove contractions before cleaning the rest of text.

### **3.3.2 Removing Noise**

Once the contractions have been identified and removed, next task is to identify the special characters, spaces, hyperlinks, hashtags, and noise in the same form. For instance – python is a good. language!!#important. Such noise is to be identified and eliminated from the text. Using regex function will identify any matching string or character as noise and later it can be removed. The cleaned sentence would be python is a good language

### **3.3.3 Converting to Lower**

Finally, after the cleaning of noise is done, in the last step words in the text are converted to lower.

### **3.3.4 Extracting English and emoji reviews**

As emojis plays an important role in sentiment, so instead of removing them they were extracted separately, and English based reviews were extracted separately. After extracting both types review it was then merged in one data frame to calculate the sentiment score for each of them.

### **3.3.5 Calculating Polarity score**

After cleaning and extracting reviews, next step is to calculate the polarity score which includes positive, negative, neutral and, compound score. It is calculated with the SentimentIntensityAnalyzer library which calculates the polarity score by removing stop words and considering only good/bad words automatically.

## **3.4 Word cloud for different Emotions**

The words that were most frequently used to describe various emotions can be shown as a plot using the word cloud. This style of plot enlarges and bolds the words with the highest frequency while reducing the size of the word as the frequency of the word increases. Four plots depicting word clouds for positive, neutral, and negative and emoji-based emotions are shown below.

**Emoji Word Cloud**



**Figure 17 - Emoji Word Cloud**

The above emoji word cloud plots the emojis as per its frequencies, emoji with higher

frequency is bigger in size and so on.

**Positive Word Cloud**

Text

Description automatically generated

**Figure 18 - Positive Word Cloud**

As per the above word cloud it shows the positive words which appeared maximum times in the reviews such as great, nice, clean, recommend etc.

**Negative Word Cloud**

Text

Description automatically generated

**Figure 19 - Negative Word Cloud**

As per the above word cloud it shows the negative words which appeared maximum times in the reviews such as dirty, noise etc.

**Neutral Word Cloud**

Text

Description automatically generated

**Figure 20 - Negative Word Cloud**

As per the above word cloud it shows the neutral words which appeared maximum times in the reviews such as arrival, automated etc.

## **3.5 Feature Engineering**

Creating features for machine learning algorithms involves employing domain expertise of the data to do so. This process is known as feature engineering. By generating features from raw data that aid in the machine learning process, feature engineering can improve the prediction capacity of machine learning algorithms. The art of feature engineering

### **3.5.1 Correlation of independent variable with dependent variables**

Chart, box and whisker chart

Description automatically generated

**Figure 21 - Correlation graph**

As per the above graph latitude and minimum\_nights are negatively correlated with price and rest of the attributes are positively correlated with the price.

**3.5.2 Conversion of categorical attribute into numerical attribute**

Categorical data can be transformed into numerical data in a variety of ways, including Label Encoding and One-Hot Encoding. Given the large number of entries in this dataset, label encoding was chosen since it makes it simpler to use the labels in the future than to create additional columns.

**Label Encoding: -** With the help of this data preparation method, we attempt to transform the categorical column data type into a numerical one (from string to numeric). This is done because string characters cannot be encoded in a way that is machine-understandable because our machine learning model cannot understand them. The Label Encoding technique is used to do this. The categories that are present under the categorical features are translated in the Label Encoding method in a way that is related to hierarchical separation. This means that label encoding should be used to encode categorical characteristics if the categorical variables are linked to one another in terms of hierarchy. Label Encoding is not a good option for non-hierarchical features because it negatively impacts the model's accuracy when applied to such characteristics.

### **Data Normalization**

There are various techniques for normalizing the data such as Min-Max Scaling and Standardization Scaling. For this dataset, Min-Max Scaling was used.

**Min-Max Scaling:** - Divide the result by the range after deducting the minimum value from the highest value in each column. The minimum and maximum values for each new column are 0 and 1, respectively.

### **Feature Selection**

Choosing the key features for the model is known as feature selection. A feature is a trait that affects or helps solve an issue. For this dataset feature score is calculated for each attribute and check which feature has less contribution (which is shown in Table 6). As in the below table it is seen that Sentiment is playing least role.

**Table 6 - Feature Selection**

|  |  |  |
| --- | --- | --- |
|  | Specs | Score |
| 0 | neighbourhood | 96.39166 |
| 1 | latitude | 168.9474 |
| 2 | longitude | 11.2541 |
| 3 | room\_type | 2486.589 |
| 4 | minimum\_nights | 140.1098 |
| 5 | number\_of\_reviews | 299.5968 |
| 6 | reviews\_per\_month | 402.6164 |
| 7 | availability\_365 | 212.0655 |
| 8 | number\_of\_reviews\_ltm | 482.4101 |
| 9 | Sentiment | 5.492908 |

## **3.6 Data Modelling (Price Prediction)**

Data modelling is the process of building models based on the type of target values, then training the model to mimic the logic of making decisions based on the information at hand. Depending on the type of data, there are various model types that can be used which are as follows:

### **3.6.1 Linear Regression**

The link between a dependent variable and an independent (predictor) variable is modelled using a machine learning technique called linear regression. The model can employ linear predictor functions to model the relationship and estimate unknown parameters from the data. To determine whether the independent variables are useful in price prediction, linear regression is used. For instance, a larger house costs more money.

Some connections, like the one between longitude and price, are not linear, though. Not only in two-dimensional space, but also in multidimensional space, are linear relationships possible. The price can be predicted using a linear regression model based on previous data if there is a linear relationship between two or more factors.

### **3.6.2 Decision Tree Regressor**

A decision tree creates tree-like models for classification or regression. It incrementally develops an associated decision tree while segmenting a dataset into smaller and smaller sections. The outcome is a tree containing leaf nodes and decision nodes.

**3.6.3 KNN Regressor**

A straightforward approach called K nearest neighbours predicts the numerical target by storing all the relevant examples and using a similarity metric (e.g., distance functions). KNN is a non-parametric technique that has been utilised in statistical estimates and pattern recognition since the early 1970s.

Calculating the average of the K nearest neighbours' numerical target is an easy way to implement KNN regression. An alternative method makes use of the K nearest neighbours' inverse distance-weighted average. The same distance functions are used in KNN regression as in KNN classification.

### **3.6.4 Random Forest regressor**

A random forest is a meta estimator that employs averaging to increase predicted accuracy and reduce overfitting after fitting numerous classification decision trees to different dataset subsamples. If bootstrap=True (the default), the size of the sub-sample is determined by the max sample’s argument; otherwise, each tree is constructed using the entire dataset.

### **3.6.5 Price Prediction with Sentiment Score**

**Table 7 - RMSE value calculation**

|  |  |  |
| --- | --- | --- |
|  | Model | RMSE |
| 0 | Linear Regression | 65.53 |
| 1 | KNN | 69.24 |
| 2 | Decision Tree | 83.58 |
| 3 | Random Forest | 61.69 |

**RMSE value: -** The standard deviation of the errors that happen when a prediction is made based on a dataset is known as RMSE. This is the same as MSE (Mean Squared Error), but when assessing the model's correctness, the value's root is considered. Different models are applied to calculate the RMSE value to check which model is accurate for the price prediction (shown in Table 7). As shown in table Random Forest and Linear Regression has less RMSE value as compared to other models.

#### **3.6.5.1 Graph Representation**

* **Linear Regression: -**

Chart, line chart, histogram

Description automatically generated

**Figure 22 - Linear Regression Prediction**

As per the above figure 22 Linear regression is not predicting the accurate value as it shows much difference from the actual price. It shows that this model is not best for predicting the price.

* **KNN Regressor: -**

Chart, histogram

Description automatically generated

**Figure 23 - KNN Regressor Prediction**

As it can be seen from the above figure 23 at the initial point it is predicting the accurate values but after 50-dollar price it is showing much difference between the actual and predicted price. So, this model is also not best suited for the prediction.

* **Decision Tree Regressor: -**

Chart

Description automatically generated

**Figure 24 -Decision Tree Regressor Prediction**

As per the above figure 24 it is visible that it is predicting the accurate values as there is not much difference between the actual and predicted values. So, as per the graph it is best suited model among all the models instead of having highest RMSE value.

* **Random Forest Regressor: -**

Chart, histogram

Description automatically generated

**Figure 25 - Random Forest Regressor Prediction**

As it can be seen from the above figure 25 at the initial point it is predicting the accurate values but after 50-dollar price it is showing much difference between the actual and predicted price. So, this model is also not best suited for the prediction.

### **3.6.6 Feature Importance**

After applying models with the sentiment score then feature importance is again checked that which attribute is playing much role in the model. This was done in order to confirm if sentiment score played any role in terms of model while predicting the price. Based on the results, the important features would be selected, and models would be implemented again using that feature. Feature importance’s for top model is shown below -

**Linear Regression: -**

Chart, waterfall chart

Description automatically generated

**Figure 26 - Linear Regression Feature Importance**

**Decision Tree: -**

Chart, bar chart

Description automatically generated

**Figure 27 - Decision Tree Regression Feature Importance**

**Random Forest: -**

Chart, bar chart

Description automatically generated

**Figure 28 - Random Forest Regression Feature Importance**

As from all the above graphs it is observed that sentiment is playing least role in every model. To check that, model is again applied on the dataset by disregarding the sentiment attribute.

### **3.6.7 Price Prediction without Sentiment Score**

**Table 8 - Calculation of RMSE**

|  |  |  |
| --- | --- | --- |
|  | Model | RMSE |
| 0 | Linear Regression | 65.54 |
| 1 | KNN | 69.36 |
| 2 | Decision Tree | 83.81 |
| 3 | Random Forest | 61.62 |

After disregarding the sentiment Score Linear Regression and Random Forest has less RMSE score seen from above Table 8

**3.6.7.1 Graph representation of actual and predicted values without Sentiment score**

* **Linear Regression: -**

Chart, line chart, histogram

Description automatically generated

**Figure 29 -** **Linear Regression Prediction**

As per the above figure 29 Linear regression is not predicting the accurate value as it shows much difference from the actual price. It shows that this model is not best for predicting the price.

* **KNN Regressor: -**

Chart, histogram

Description automatically generated

**Figure 30 - KNN Regressor Prediction**

As it can be seen from the above figure 30 at the initial point it is predicting the accurate values but after 50-dollar price it is showing much difference between the actual and predicted price. So, this model is also not best suited for the prediction.

* **Decision Tree Regressor: -**

Chart

Description automatically generated

**Figure 31 - Decision Tree Regressor Prediction**

As per the above figure 31 it is visible that it is predicting the accurate values as there is not much difference between the actual and predicted values. So, as per the graph it is best suited model among all the models instead of having highest RMSE value.

* **Random Forest Regressor: -**

Chart, histogram

Description automatically generated

**Figure 32 - Random Forest Regressor Prediction**

As it can be seen from the above figure 32 at the initial point it is predicting the accurate values but after 50-dollar price it is showing much difference between the actual and predicted price. So, this model is also not best suited for the prediction.

## **3.7 Result Evaluation**

As per the data modelling and results Random Forest has the least RMSE in case of four models with Sentiment score and without Sentiment score. But when the plots were drawn it was seen that Decision tree has given the best output in terms of testing data for actual and predicted values as seen in the Figure 31

# **4.** **Conclusion**

* Based on the analysis done below figure 33 represents the distribution of the sentiments of Airbnb hotel available in Toronto region. It can be seen that 97% reviews received are positive while only a small ratio of 2% and 1% are neutral and negative respectively.

Chart, bar chart, waterfall chart

Description automatically generated

**Figure 33 - Distribution of Sentiments**

* Based on the below graph it is clear that the positive review for Airbnb have gradually rose since 2010, while on the other hand negative remain static with time.

Chart, line chart

Description automatically generated

**Figure 34 - Distribution of sentiments over years**

* For predicting Airbnb price various models were used such as Linear Regression, KNN, Decision Tree, Random Forest. Out of them Decision Tree worked well so it should be used in terms of predicting the price.

# **5 References**

[1] Luo, Y. et al. (2019) “Predicting Airbnb Listing Price Across Different Cities.”

[2] Tang, E. (2015) “Neighbourhood and Price Prediction for San Francisco Airbnb Listings.”

[3] Q Trung Nguyen/Airbnb-Pricing-Prediction.

[4] S M Kim and E Hovy. (2004) *Determining the sentiment of opinions*

[5] A Agarwal, F Biadsy, and K Mckeown. (2009) *Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams*. Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 24–32, March.

[6] L Barbosa and J Feng. (2010) *Robust sentiment detection on twitter from biased and noisy data*. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44.

[7] Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T. (2011). *Target-dependent twitter sentiment classification*. 49th Annual Meeting of Association of Computational Linguistics, pages 151–160.

[8] Bo Pang, L Lee, (2004) A sentimental education: *sentiment analysis using subjectivity summarization based on minimum cuts,* ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Article No.271.

[9] Li, H.; Bian, Y.; Xu, X.; Sun, G. SRRTC: *Social Recommendation Based on Relationship Transmission with Convergence Algorithm*. Preprints 2017, 2017110033 (doi: 10.20944/preprints201711. 0033.v1).

[10] Kong X, Jiang H, Yang Z, Xu Z, Xia F, Tolba A (2016) *Exploiting Publication Contents and Collaboration Networks for Collaborator Recommendation*. PLoS ONE 11(2): e0148492.

[11] G. Vinodhini et al, (2012) Sentiment *Analysis and Opinion Mining*: A Survey, International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol 2, Issue 6.

[12] F Alam; A Paul, (2016) *A computational model for trust and reputation relationship in social network*, Published in: International Conference on Recent Trends in Information Technology (ICRTIT), ISBN: 978-1-4673- 9802-2

[13] Hota, Soudamini, Pathak, Sudhir, *KNN classifier-based approach for multi-class sentiment analysis of twitter data*, Volume - 7, 10.14419/ijet. v7i3.12656, International Journal of Engineering and Technology (UAE)

[14] H. Yu and J. Wu, (2016) “Real estate price prediction with regression and classification,” CS229 (Machine Learning) Final Project Reports.

[15] D. Wang and J. L. Nicolau, (2017) “Price determinants of sharing economy-based accommodation rental: A study of listings from 33 cities on airbnb. com,” International Journal of Hospitality Management, vol. 62, pp. 120–131.