Leveraging data mining techniques to predict rating and price using Airbnb Dublin data.

Sahil Wadhwa
MSc Data Analytics
National College of Ireland
Dublin, Ireland
X17170133@student.ncirl.ie

Sai Girish
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x17170401@student.ncirl.ie

Zainul Abedin Khatik
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x18110096@student.ncirl.ie

Atif Feroz
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x17169992 @student.ncirl.ie

***Abstract*—Airbnb a $25 Billion-dollar company is one of the most recognizable international brands in the emerging "accommodation sharing economy" which is dealing with the increasing number of property owners making dwellings available for short-term rentals. For the customer to undergo a special experience a detailed investigation is needed to systematically examine the lodging experience. Also, the overall ratings given by customer narrates the satisfaction which affects the listing price. Therefore, a study is required to bridge the gap of overall ratings and customer listing price. For this study Dublin data was gathered for both review and listing. In this paper we try to attempt an approach to predict the overall rating based on length of review using Random Forest and further an attempt to predict listing price was carried using Linear Regression, SVR and tuned SVR. XGBoost algorithm is used to increase the accuracy of regression models. For evaluation and results Root Mean Error Square (RMSE) was used. (*Abstract*)**

Keywords—Support vector regression, Airbnb price, review length, raiting, XG Boost, .

I. INTRODUCTION

Airbnb was established in 2008, since its inception it has grown at a tremendous pace, it has more than 200 million guests from more than 65,000 cities. Airbnb has disrupted the tourism and hospitality market with an estimated value of $30 billion [14]. Airbnb listings not only provide us a window into how members in the sharing economy advertise their offerings, but also provide us deep insights about the characteristics of a city and its neighborhoods. Also, reviews give insight to improve your business, products, and the overall customer experience. We gathered data from Inside Airbnb specifically for Dublin city which is publicly available. For this research project we have considered reviews dataset and listing dataset. The listing dataset has 106 features and reviews dataset has 6 features. In this research we propose to analyze Airbnb's publicly available reviews given by the customers and listing information for historical data and try to validate our hypothesis: (1) prediction of overall rating using length of the review (2) prediction of optimum price using important predictors such as neighborhood, room type, cleaning fee, reviews score rating, reviews per month, accommodates, etc.

Hypothesis (1): Prediction of ratings using length of the review.

Customers mostly rely on reviews given by other people to get an information about a service. But, due to overwhelming number of reviews, it is nearly impossible for the potential customer to go through each and every review and make a decision whether to book a place or not. The larger the number of reviews also makes it difficult for the service provider(host) to keep track of customer feedback. Furthermore, we observed that when a customer is angry about a service, the customer writes a longer review to complaint what problems they have faced in detail and contrary to that, they leave a shorter review when they are happy about the service. To simplify the decision process, we propose a novel approach based on the intuition, to predict overall rating using length of the review. We have used Linear regression, Random forest and XGboost models. Our analysis shows that rating can be predicted using length of review.

Hypothesis (2): prediction of listing price using review length and important predictors such as neighborhood, room type, cleaning fee, reviews score rating, reviews per month, accommodates, etc.

Despite having a good amount of success, Airbnb has struggled to optimize pricing by its host [15]. So, proper pricing strategy can lead to success of the company and host who is listed their house. Knowing the right parameters which are affecting the price can help the hosts to keep a reasonable price for their listing. In this way the company, hosts and the guests can benefit from the sharing economy.

In this research we will make an attempt to predict the price of the listing using important predictors such as neighborhood, room type, cleaning fee, reviews score rating, reviews per month, accommodates, etc. For this study we have selected **15 features** from the listing dataset. Several machine learning methods, such as Linear Regression, Support Vector Regression (SVR), tuned SVR and XGboost is being used. Results are evaluated on RMSE between the logarithm of the predicted value and the logarithm of the observed value. Our analysis shows that RMSE value of tuned SVR is 0.3019 and after boosting using XGboost it was 0.2893 with 74% accuracy.

The rest of this paper is organized as follows: Section 2 discusses literature review, and Section 3 presents the data and methods used in this research project; next, in Section 4 we present the main estimated model results and their implications, and in Section 5 conclusions and suggestions are presented.

Research Question:

1. Can we predict the ratings using the review length?
2. To what extent data mining techniques can help Airbnb host in determining listing price for their property?

II. LITERATURE REVIEW

This section has been classified into two categories.

1) Literature based on the analysis of review.

It has been witnessed that opinion of other people in regard to what to eat, where to stay, what to do majorly effects the decisions taken by the concerned people and internet plays a vital role in molding these suggestions **(Json Jong 2011)** [7]. There are various online sites that act as a database of reviews and ratings typed by people about various restaurants and places for which different researchers have explored models like SVM and Support word vector to capture information present in each words in association for review and ratings where Yelp reviews were used to estimate star ratings and the model came up with accuracy of 70% depicting a strong model [7] . **Further,Turney's** work in [13] uses a specific unsupervised learning technique based on the mutual information for document phrases and the words "poor" and "excellent" .However, sentiment information is not able to be captured by unsupervised vector based approaches, therefore, **Andrew.L.Maas in 2011** [8] presented a model that incorporates a mix of supervised as well as unsupervised techniques in order to learn the word vectors so that it could capture the sentiment content as well as information in the document where huge data of movie reviews were taken for analysis. Similarly, **Pang et al.** in [12] used movie data to examine several supervised machine learning methods for sentiment classification of movie reviews. More often merchants trading their products online request customers to rate & review products giving rise to huge data of product review making it difficult for the potential customer to go through all the reviews **(Minqing hu and Bing Liu, 2011)** [9] .This problem is increasing with people now shopping on web .To tackle this, authors in [9] proposed a set of techniques for mining and summarizing the reviews of the products based on the natural language processing and data mining methods, authors believe summarizing these reviews proves to be useful which was done by mining product features and identifying opinion sentences. The technique was effective enough in performing the task. Airbnb customers as we know usually consider the overall ratings in purchasing the specific product, therefore a study was required that would bridge the gap between the experience gained by the customers in lodging and overall ratings. With this aim **Yi Luo in 2018** in [10] used a total of 250,439 reviews for 6,946 listings. It was noted that 60% of Airbnb listings had overall of 5 stars. According to the author the length of review is directly proportional to the customer's sentiments and opinions, therefore reviews for most and least words were extracted. Using latent aspect rating analysis, result concluded that reviews with higher words has low ratings. Further in 2015 **Walid Majeed** in [11], made use of several probabilistic techniques and heurists for classifying the app reviews of Apple and Google into ratings based on their metadata such as length of text and star ratings involving keyword frequencies and linguistic rules. An estimate of 1.1 million reviews for 1100 apps were gathered and 146,057 reviews were taken from 40 apps from Google apps. According to the author the length of the text of review can be used as a feature where lengthy reviews can be more informative, and tense of the review used as review type.

Going through the previous work, in our literature, we try to attempt a novel approach of using the feature word length of reviews to predict the overall ratings of all Airbnb listing by which review length can be used to determine the overall satisfaction of visitors.

2) Literature based on price prediction.

Since the onset of Airbnb in the year 2008 has absolutely transformed the way of travelling for the people. This site has advertised more than 1.5 million listings by 2015 around 34000 cities throughout the world. These listings not only aware about how participants in the new sharing economy market their offerings but also provide an exclusive insight in the attributes of a city along with its neighborhoods. In 2015**, Emily Tang and K. Sangani** in [1] have attempted to develop SVM classifier which was able to predict the neighborhood's listings and their prices for the region of San Francisco. The analysis was a success in price prediction at 81.2% accuracy for all feature whereas failed to attain accuracy for neighborhood listing prediction where only bag of words text features was used.[1] But however, the model couldn't be used to predict non-linear and non-stationery processes for which (**Jiaoyang Wu ,2017)** in [3] used a regression model which could accurately predict the trends for non-linear and non-stationary processes for 20 explanatory features of 21,613 housing sales in King Country, US. Author used Support Vector Regression (SVR) for the purpose of prediction of house prices where PCA was used for feature extraction and RFE, Ridge, Lasso, and Random Forest Selector algorithms were used for feature selection. The results show after a proper feature selection price prediction accuracy increased from 65% to 86% using SVR model

Real estate market price prediction was a problem as it is exposed to lot of price fluctuations which is generally due to existing correlations along with many variables for what some variables could be controlled and some may not be. To tackle this problem (**Alejandro Baldominos in 2018**) in [2] aimed to develop a machine learning application that would identify various opportunities in the real estate market. A Similar research was carried by **Park and Kwon Bae** in [6] who analyzed housing data for 5359 houses in Fairfax County, Virginia. Authors used logistic regression and t-test for feature selection and did a comparative study of decision tree (C4.5), RIPPER, Naïve Bayes and Ada Boost for prediction where RIPPER outperformed the rest.

Further **P. Ye et al.,** [4] have explained the pricing strategy model that is deployed at Airbnb. Author introduced a set of metrics that is used to estimate the effectiveness of a pricing strategy and have proposed a customized regression model which will help them learn a pricing strategy which would

eventually minimize the bad suggestions. Whereas for proper understanding of features like location, bedrooms, house type to predict new listing price (**Choudhary 2018**) in [5] developed a model which was intentionally developed to prove helpful in internal pricing using Random Forest Regressor on balanced and imbalanced dataset.

III. Exploratory Data Analysis

After data collection and load in R, the next step is to examine the data in detail.

A. Data Distribution

CASE-I:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	85.0	94.0	78.4	98.0	100.0

Table.1

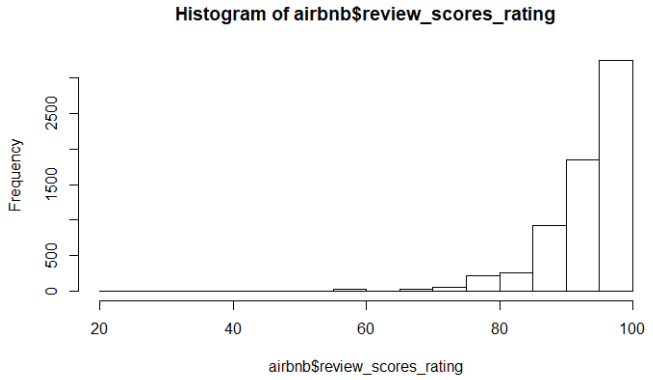


Fig.1 Histogram of review score rating

In this research model the dependent variable is review score rating, which measures the overall review rating given by customer. The data is not normally distributed as the mean is less than median, it means, our dependent variable is left-skewed showed by Table.1 and Fig.1.

CASE-II:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	54.0	85.0	110.6	135.0	999.0

Table.2

In this case the dependent variable is price, which measures the daily price paid by customer for booking a listing customer.

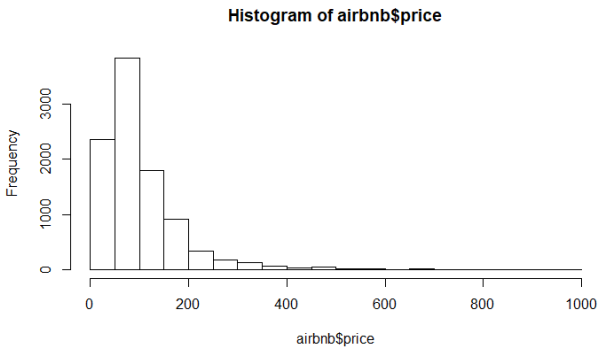


Fig.2 Histogram of price

The data is not normally distributed as the mean is greater than median, meaning, our dependent variable is right-skewed showed by Table.2 and Fig.2.

B. Normalizing the data

Regression assumes data to be normally distributed however, it is not considered to be a strict assumption. But in the real world generally the data will not be distributed normally and as our data is distributed in different scale we use normalization technique which changes the values of numeric columns in the dataset to use a common scale, without loosing information. We have used min-max normalization technique in which we fit the data, in a pre-defined range because it is more efficient and gives accurate results. To normalize the data in the boundary of [A,B], the min-max normalization is defined as:

xi_normalized = (xi - min(x)) / (max(x)-min(x)) * (B-A) + A

C. Correlation Matrix and Feature Selection

If the independent variables have identical information, keeping both variables in regression model will not show any great effect, this is called as multicollinearity between the variables. Multicollinearity leads to unstable estimates as it tends to increase the variances of regression coefficients and also it increases the process time when a huge data is considered while regression. The solution to this problem is to keep only one of the two independent variables that are highly correlated in the regression model as they both have same impact on the dependent variable.

Case-I: Histogram depicts distribution of values for each feature. Since the correlation is less than 0.7 hence, our predictor variable is suited for this research.

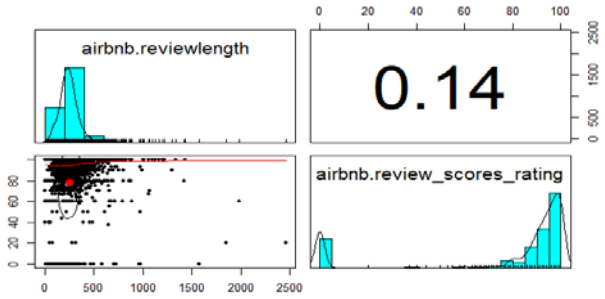


Fig. 3

Case-I: Histogram depicts distribution of values for each feature. Since the correlation is less than 0.7 hence our predictor variable is suited for this research.

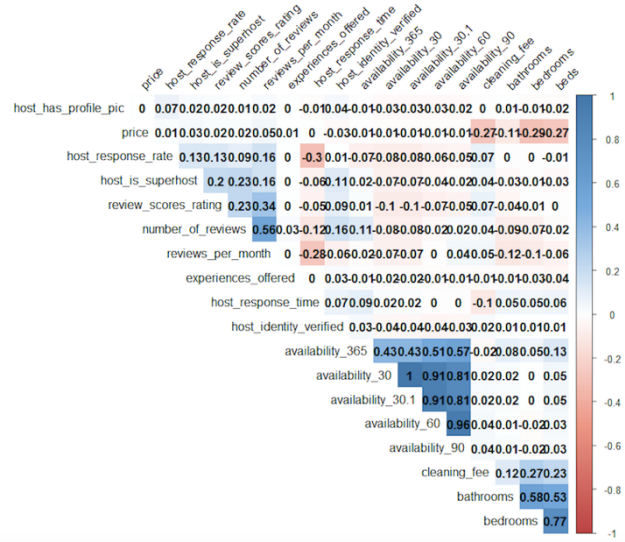


Fig. 4 Correlation matrix

Case-II: Checking correlation is an essential step before doing any analysis and for that we use Pearson's correlation. Red color indicates negative correlation and blue color indicates positive correlation. The largest correlation is between availability_60 and availability_90 is 0.96 and availability_30 and availability_60 is 0.81. Also, we saw a strong uphill linear relationship between bedrooms and beds which was 0.77. The procedure is repeated for all variables. We remove the feature that have correlation value greater than 0.7.

D. Box Plot

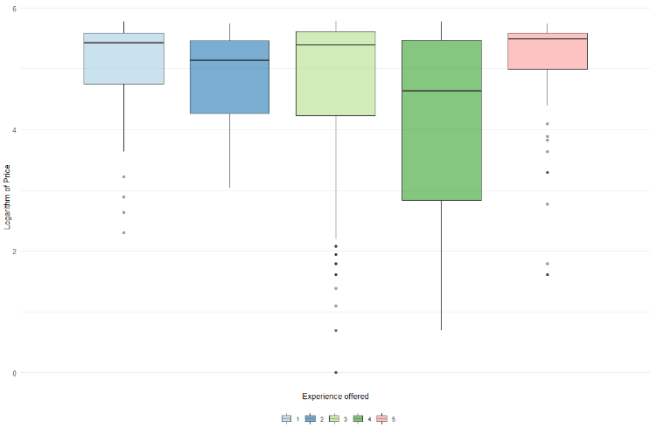


Fig.5 Different types of experienced offered vs price

Case-II: Fig. 5 illustrates boxplots of different types of experienced offered vs price. The different types of experienced are business, family, mixed, romantic and social which are represented in the boxplot as 1, 2, 3, 4 and 5 respectively. It can be observed in boxplot that prices among business, family and none do not vary much. However, prices increase tremendously when it comes to romantic.

E. Box Plot

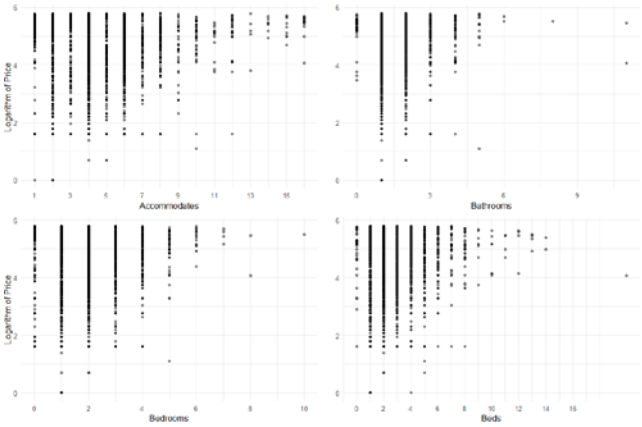


Fig. 6 Logarithm of price w.r.t. accommodates, bathrooms, bedrooms and beds.

Fig shows scatterplot of log of price versus accommodates, bathrooms, bedrooms and beds. It can be observed that it has positive relationship as price will increase if it can accommodate more people, and has more bedrooms, bathrooms and beds.

IV. METHODLOGY

There are nearly 4 machine learning models and two feature extraction models with one accuracy boosting models are applied on the dataset to solve the Business query.
The details about the models applied are explained in the below table.

Applied Machine learning models	
1	Linear Regression
2	Random Forest
3	Support Vector Machine (SVM)
4	Support Vector Regression (SVR)
Feature Extraction models	
1	Correlation
Improving Accuracy models	
1	XGBoost

A. Linear regression

The linear regression is a method concerned with providing the relation between one dependent which is numeric in nature (the value to be predicted) and one or many independent variables (predictors) [1]. In this model the process tries to assume that the dependent and the independent variables follows a straight line. The simple equation for regression analysis is **y=a+bx**, where y is dependent x is independent variables. In this project for one question the price is dependent variables like y and other variables like availability for 30 days, host response rate and time, room type, accommodation and other 10 variables.
To solve the next question the dependent variables is rating, and the independent variable is Review length.

B. Support vector Regression (SVR)

The machine learning algorithms are generally divided into two types they are supervised and unsupervised learning.

Now in this project the supervised machine learning SVM algorithm is used and this can also be used for classification and as well as regression purposes and here the regression model is used in which the dependent variable is price which is continuous in nature and the independent variables are nearly 20 which includes categorical and numerical variables so that to solve the regression the SVM is used to solve the issue.

The major target of SVM or SVR analysis in this case is to define a hyperplane between the classes of the data so that the model can predict the continuous dependent variable easily.

In the external word the data will not be distributed linearly and if the data is not distributed then the models will not fill on the data. To solve this problem the kernels where introduced in SVR or SVM algorithms. These kernels convert the data to a higher dimensional feature space values by applying some mathematical formulas and functions. The details about the functions and kernels in SVR analysis is given below.

Type of Kernel	Function
Radial basis	K (xi, xj) = exp (- [‖xi – xj ‖2] / 2σ2)

In the above function Xi and Xj are the vectors in feature space and K is their cross product.

Steps for implementing SVR model

As mentioned above the steps for implementing the model for linear regression same are followed here.

1)Preparing the data for analysis

2)For auto tuning the model of SVM, in this research the CARET package is used, and the RADIAL kernels used in the analysis with gamma and epsilon values like 0.047 and 0.1.

3)model is trained and tested using train, test data and plotting of results gives the idea about results obtained.

C. Tuned SVR model

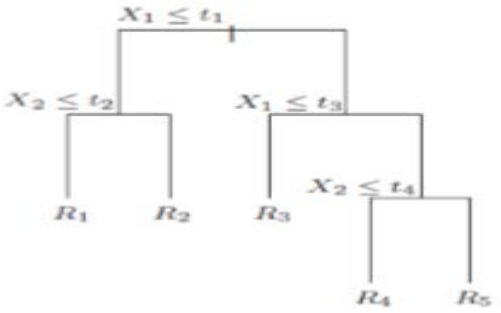
This model is the same as SVR, but the difference is that the number of iterations done by model by considering different cost and epsilon values and predicting the best value of dependent variable.so that the accuracy of the model will be high when compared with normal SVR model. The steps for undergoing this model is same as normal model but they differ in giving the range of cost and epsilon values. Where they should be given by the user externally.

In this research model the cost value ranges from 0 to 1 with an interval of 0.1 whereas for epsilon the vales are like 0.5 to 8 with an interval of 0.5

D. Random Forest

The model of Random forest for classification and Regression (CART) was proposed by L. Breman in the year 1984. This model was later modified by Breman by introducing Bagging technique for RF in 1996.[2]

We have made use of Random Forest model to predict review score rating based on review length calculated by counting the number of characters in the comment and aggregating the review length by taking the average of that length by listing id. Random Forest is a classifier which consists of a collection of tree structured classifiers {h (x,Θk), k=1,...} where { Θk } are independent identically distributed random vectors for which each tree casts a unit vote for the most popular class at input x .[1] RF is basically an ensemble learning method used for classification based on 2 techniques called CART and Bagging. CART is a technique which is tree structured classification model which maps the item observation to its conclusion. The above illustration depicts the working of CART).



CART iteratively splits each node into 2 sub-nodes by checking for best split variable along with best split value till minimum node size is attained. It makes decision based on the split of variable. CART's advantage is that it fits into data perfectly but however for prediction CART accuracy isn't good to counter this Bagging technique was introduced. The Bagging would reduce the variance of CART and keeping the bias low. [2]

RF can also be used for regression, depending on random vector Θ for tree predictor h(x,Θ) which takes numerical values .The output is of numerical form and training set is drawn independently of random vector Y, X where mean square generalization error for predicitor h(x) and in our case reviewlength is :

$$EX,Y(Y-h(X))2$$

The predictor is formed by taking average of K of the trees {h(x,Θk)} which is also similar in classification phase. [1]

E. XG Boosting

Various machine learning algorithms such as Support Vector Machines, Logistic/Linear regression, tree based models, neural networks, etc have been applied in order to forecast and analyse the markets but researchers have found out that Extreme Gradient Boosting is one among those machine learning algorithms that has got more success rate as compared to other algorithms in terms of attaining accuracy [3]. This algorithm was first developed by Tianqi Chen that was basically used for building predictive tree based models. Therefore, XG Boost is an approach wherein new models are developed that help in predicting the errors or residuals of previous models and then are summed up together so that final prediction is made. It is called so, as it implements a gradient descent algorithm for minimization of loss as new models are accumulated. It not only supports classification predictive modeling problems but regression models as well.(3)

Objective for XG Boost model is given by:

$$Obj = L + \Omega$$

Where, L= Loss function that regulates the predictive power,

Ω=Regularization component that controls over fitting and simplicity.

Loss Function (L) needs to be optimized and could be Log Loss for binary classification, Root Mean Squared Error for Regression, and M Log Loss for Multi Class Classification.

Whereas, the Regularization Component (Ω) depends on prediction score and number of leaves that are attributed to the Tree Ensemble Model. (4) It is basically a regression tree that follows the similar decision rules as that of a decision tree. The difference between the two is that in the former, leaf nodes with scores depict a decision, whereas, the inside nodes depict the values for the attribute test and resultant of the overall prediction is the sum of the scores predicted by K trees

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F$$

XG Boost follows the similar pattern as that of gradient tree boosting with small alteration; it incorporates a minor improvement on the regularized objective.

$$\text{Obj}(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k}^{K} \Omega(f_k)$$

As we noticed, former is a loss function that measures whether the whole model is fit for the training set data, whereas, later punishes the complex nature of the model. But when the complexity increases in the model, corresponding score will be deducted. Thus the purpose of XG Boost algorithm is to avoid over fitting of the model. (5)

RESULTS *(Heading 5)*

Case study 1:

Rating	Average Length
2.5	296.7273
3	277.6684
4	246.3212
4.5	250.4148
5	248.8797

Fig

The figure shows overall rating along with the average length after boosting the random forest using XGboost model. Our analysis shows that when the length of review is longer there is high probability that the rating will be low and when the length of review is short, there is high probability that the rating will be high

Case study 2:

Value	Lin regress	SVM model	Tuned SVM model	XGboost mo
MODEL_RMSE	0.3043658	0.3019419	0.2952528	0.289382
DEL_ACCURACY	0.6944661	0.7388932	0.7381138	0.743569

XGboost model has the lowest RSME at 0.2893, Linear regression has RMSE value 0.3026 with 69% accuracy. SVM has RMSE 0.309 with 73% accuracy and after tuning the SVM using Tuned SVM we get RMSE value 0.2952 with 73.88% accuracy. Even though there not much difference between the RMSE values of SVM and Tuned SVM, but it as a significant effect when we increase the size of the data. Lastly, XGBoost has RMSE value 0.2893 with 74% accuracy.

CONCLUSION

This study uses a sample of 319,511customer comments. Case study 1 shows that when the length of review is longer there is high probability that the rating will be low and when the length of review is short, there is high probability that the rating will be high. Also, in case study 2 our analysis shows that we were able to achieve 74% of accuracy in predicting the price of the hosts property by considering different features such as neighborhood, room type, cleaning fee, reviews score rating, reviews per month, accommodates, etc.

LIMITAION & FUTURE WORK

Failed attempt on text analytics:

We made an attempt to do opinionated text data analysis. The aim was to decompose this overall rating into ratings on different aspects such as rooms, location, service, etc. so that we can get detailed understanding of the reviewer's opinion about the place.

Approach: Initially we did sentimental analysis on the Airbnb reviews given by the customer. Using textcat we selected only English language. Using wordcloud we plot the cloud of words on review given by the customer.

Fig 7. Word Cloud

We used Qdap package to get polarity (positive and negative words).

Fig. 8 Positive vs Negative words

Then, we used tidytexts bing library to determine sentiment score.

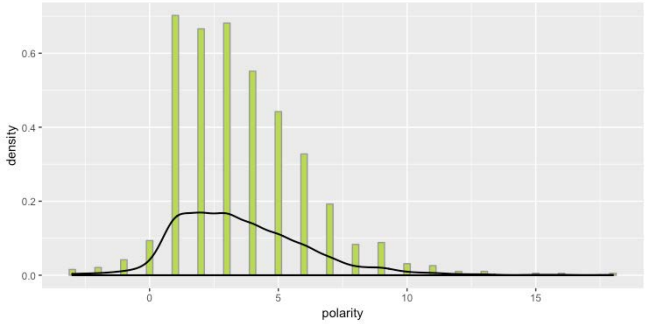


Fig. 9 polarity distribution for both Qdap's polarity score and TidyText's BING

Next, we made a polarity-based corpus and applied TFIDF on it.

Using NRC words are divide in emotions like anger, anticipation, disgust, joy, sadness, surprise, trust, negative and positive.

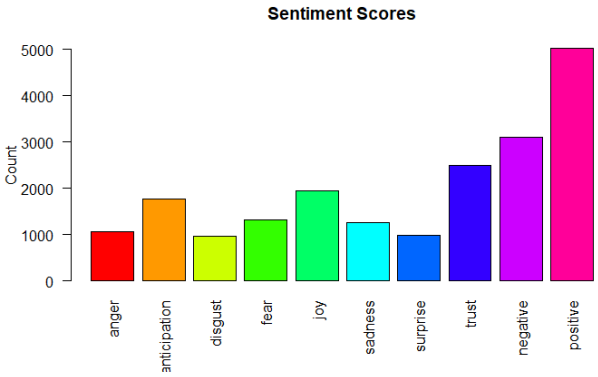


Fig. 10 NRC (showing emotions)

On the same corpus we applied Latent Dirichlet Allocation using Gibbs method with 4 topics to determine group of comments that belong to the same cluster.

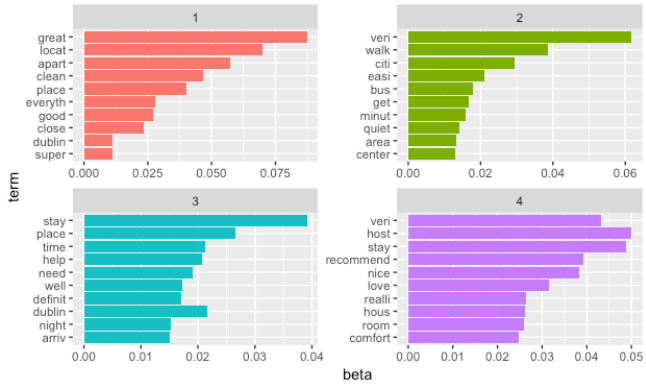


Fig. 11 LDA- Topic modelling.

Due to our lack of knowledge and short of time we could not build the model.

Future work:
For future work of this research, Latent Aspect Rating Analysis can be used which allows a detailed analysis of reviews on overall ratings. Based on this approach we can decompose ratings on different aspects like rooms, location, and service which can allow us to get detailed analysis on reviewer's ratings on Airbnb listings.

REFERENCES

1. Tang, Emily Wai-Ho. "Neighborhood and Price Prediction for San Francisco Airbnb Listings."(2015)
2. A. Baldominos, I. Blanco, A. Moreno, R. Iturrarte, Ó. Bernárdez and C. Afonso, "Identifying Real Estate Opportunities Using Machine Learning", *Applied Sciences*, vol. 8, no. 11, p. 2321, 2018. Available: 10.3390/app8112321.
3. J. Wu, "Housing Price prediction Using Support Vector Regression." Available: 10.31979/etd.vpub-6bgs [Accessed 20 April 2019].
4. P. Ye et al., "Customized Regression Model for Airbnb Dynamic Pricing", *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, 2018. Available: 10.1145/3219819.3219830 [Accessed 20 April 2019].
5. Choudhary, Paridhi & Jain, Aniket & Baijal, Rahul." Unravelling Airbnb Predicting Price for New Listing". 2018
6. V. Chiarazzo, L. Caggiani, M. Marinelli and M. Ottomanelli, "A Neural Network based Model for Real Estate Price Estimation Considering Environmental Quality of Property Location", *Transportation Research Procedia*, vol. 3, pp. 810-817, 2014. Available: 10.1016/j.trpro.2014.10.067 [Accessed 20 April 2019].
7. Jong, Jason. "Predicting Rating with Sentiment Analysis." (2011): 1-5.
8. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*(HLT '11), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 142-150.
9. M. Hu and B. Liu, "Mining and summarizing customer reviews", *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, 2004. Available: 10.1145/1014052.1014073 [Accessed 20 April 2019].
10. Luo, Yi. (2018)."What Airbnb Reviews can Tell us? An Advanced Latent Aspect Rating Analysis Approach". Graduate Theses and Dissertations. 16403. https://lib.dr.iastate.edu/etd/16403
11. W. Maalej and H. Nabil, "Bug report, feature request, or simply praise? On automatically classifying app reviews", *2015 IEEE 23rd International Requirements Engineering Conference (RE)*, 2015. Available:

10.1109/re.2015.7320414 [Accessed 24 April 2019].

12. Pang, B., Lee, L., and Vaithyanathan, S., 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In Proc. of EMNLP 2002
13. Turney, P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. ACL'02.
14. E. Newcomer, "Airbnb Seeks New Funding at $30 Billion Valuation", *www.bloomberg.com*, 2019. [Online]. Available: https://www.bloomberg.com/news/articles/2016-06-28/airbnb-seeks-new-funding-at-30-billion-valuation. [Accessed: 23- Apr- 2019].
15. Hill, "How much is your spare room worth?", *IEEE Spectrum*, vol. 52, no. 9, pp. 32-58, 2015. Available: 10.1109/mspec.2015.7226609 [Accessed 23 April 2019].