Hotel rating prediction from crowd-sourced hotel reviews

ADM Report

Nowadays, ubiquitous access to the internet has strengthened the online communication among people through various crowdsourcing platforms such as e-commerce,
travel booking sites, etc. Most of the crowd-sourcing platforms include the customer’s reviews for making better Decisions. This experience is continuously shared in the form of review comments, ratings, photos etc. In the tourism industry, hotel ratings are the key indicators for the growth of business and help the guests choose right accommodation. Few research works have been conducted on the relation between the conventional hotel ratings and online guest reviews. In this project, we have used multi-criteria rating provided by the hotels and online guest reviews that influence the prediction of overall ratings. In this paper, we present a novel approach of how online guest reviews and multi-criteria ratings together influence the future of hotel rating using LightGBM algorithm. Our experiments were conducted on Airbnb data, the proposed LightGBM methods were used to make suitable predictions.Using lightGBM methodology, there is a significant difference in rating prediction with the addition of sentiment score of guest reviews.

rating prediction with addition of sentiment score of guest reviews.

Research question:
To what extent multicriteria ratings and online guest reviews affects overall ratings of hosting business?

The project aims to predict two cases firstly, how services affect star rating of hosting business. Secondly is there any change in prediction of rating, with combination of text review score by users and services provided by hosting business.

Introduction

Word of mouth is a key aspect of any business it refers to good or a bad opinion of a particular product or a service . The growth of online market for business has replaced word of mouth with reviews and ratings. The reviews and ratings provided by users influences crowd in turn increases sales. For an example a customer has an option to choose between few hotels, it's more likely that the customer would choose a hotel based on rating and review. A good reputation always attracts customers. Also Quality of the service can be explained through reviews. (https://ieeexplore.ieee.org/document/6586254/).

Independent organizations are involved in booking hotels for guests. With growth in technology there is a significance increase in online booking platforms, these platforms influence the customers for making better decisions by continuously sharing their information regarding their travel experience in the form of review comments, ratings, photos etc by the host and

Ratings on hotels is a sign if a hotel is successful and popular. For hosting business, star rating and reviews on online platform are vital factors. The reviews show the quality of services and it will help to drive more guests. Stars are provided by the online platforms to classify the hosting based on quality of services provided.

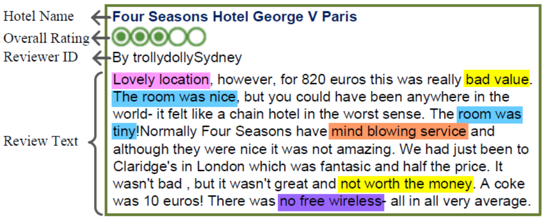
For our project we have considered ratings, reviews and services provided by hotels in Airbnb platform. Ratings of a hotel is based on the cumulative rating provided by the guests. Guests can submit overall ratings and rate the hotels based on categories i.e. cleanliness, accuracy, value, communication, arrival and location. The hosts can rate these hotels based on the facilities provided and eventually their satisfaction on a scale. With addition to this a written review can be provided by the guest in the Airbnb platform based on the stay.

Its of a higher interest for hotel owners to know, how their hotels perform on Airbnb platform. In this project we like to predict the ratings of hotels in Airbnb platform based on features, such as price, cancellation policy, Minimum nights, accuracy rating, textual reviews, etc. This project can tell what attributes should customers focus on for a better star rating. To be more precise, we collected listings and text comments of Airbnb hotels in berlin city in order to predict the overall rating as the experience shared by the users. The primary factors that affect the rating are cleanliness score, price, amenities, text reviews etc. Some of the earlier works considered multi criteria ratings like services , cleanliness, location etc and some researches predicted the hotel ratings based on sentimental analysis of text reviews. In this project we are predicting overall hotel ratings considering multi criteria ratings and text review comments from the user. Tree-based methods such as Decision tree and LightGBM has been used for predicting the overall ratings. By the way the predicting variable i.e. overall rating which as continuous scores ranging from 1 to 100 so preferred regression techniques for our analysis.

Literature Review

The advent of information and technology has strengthened the online interactions among the people through various online platforms such has travel booking website such has airbnb, tripadvisor and Expedia. All these online platforms which influences the customers for making better decisions by continuously sharing their information regarding their travel experience in the form of review comments, ratings, photos etc. From last two decades, in tourism industry there is a remarkable increase in crowd-sourced data which drawn the attention of researchers and business in building the personalised recommendation systems.

In order to perform any prediction model based on reviews, it is crucial to identify and extract the opinion of users on various features of a hotel from the review sentences. [1] proposed that developing Latent aspect rating analysis(LARA) model can aid to analyse the opinions given in the comments of online hotel reviews. This method allows to find out various levels of aspects of opinion on each feature of an object, thus forming an overall perception of that object. Authors proposed Latent regression method to resolve the Latent aspect rating analysis and conducted empirical experiments on their data. In their article they claimed that their proposed model i.e. latent regression can successfully fit for developing LARA which enables learning customized model for each individual user.



Four Seasons Hotel George V Paris

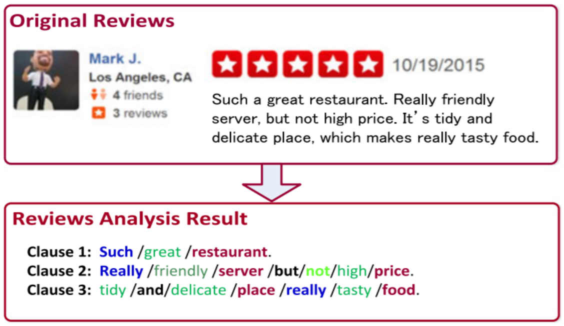
By trollydollySydney

Lovely location, however, for 820 euros this was really bad value. The room was nice, but you could have been anywhere in the world- it felt like a chain hotel in the worst sense. The room was tiny!Normally Four Seasons have mind blowing service and although they were nice it was not amazing. We had just been to Claridge's in London which was fantasic and half the price. It wasn't bad , but it wasn't great and not worth the money. A coke was 10 euros! There was no free wireless- all in all very average.

Fig. 1 Identifying aspects of a review [1]

[2] in their paper described a system to distinguish phrases from English sentences by extracting typed dependencies parses from phrase structure parsers of sentences. Authors used NP relations to form grammatical relations to observe the inherit relations among the collection of tests and compared their model's performance with other parsers such as Minipar and Link parser.

[4] proposed a sentiment-based rating prediction method (RPS) to improve prediction accuracy. Authors took sentimental measurement approach to identify user's sentiment on objects and considered sentimental attributes and interpersonal sentimental influence of user as factors. As a third factor, authors considered product reputation and clubbed all of them by similarity. Authors tested their method on yelp data set and concluded that sentimental factors can have huge impact on user characterization that helps in recommendation system.



Original Reviews

Mark J.
Los Angeles, CA
4 friends
3 reviews

10/19/2015

Such a great restaurant. Really friendly server, but not high price. It's tidy and delicate place, which makes really tasty food.

Reviews Analysis Result

Clause 1: Such /great /restaurant.
Clause 2: Really /friendly /server /but/not/high/price.
Clause 3: tidy /and/delicate /place /really /tasty /food.

Fig 2 Example of review analysis for identifying user's sentiment on Yelp. [4]

[5] proposed RAS (Ratings Are Sentiments) model which is an extended Hidden Factors as Topics Model (HFT) based on the Sentiments Unification Model (ASUM). This model learns more accurate latent factors of users and objects by combining user’s sentiments in review comments and scores. Authors conducted performance evaluation and compared the results of RAS to the HFT model. The results came in the favour of RAS model.

The extracted information from reviews will be converted into numerical scale and will be considered as factors in predicting the overall rating. Some of the research addressed /cite:anand personalised recommendation by prioritized the items in the top-n recommendation hotels list based on the interest of users. [3] proposed a recommendation model which prefers the top-n products by analysing the user's opinions regarding the different items based on the popularity in merchandise. In order to achieve the similar interest among the user about the products prefered collaborative filtering technique, opinion mining for finding a review of product and the missing values are imputed by matrix factorisation method.

[6] implemented a hybrid Multi criteria CF Recommender system for increasing predictive accuracy. Expectation Maximization algorithm was used to cluster user ratings and PCA was applied for handling collinearity issues, they used ANFIS for prediction and concluded that this model achieved very good accuracy. As ANFIS were deployed offline , they had some issues with large amount of updated data,to address this problem they proposed a incremental learning approach as a future work.

[7] using sentimental analysis methods predicted the overall ratings of the hotel from user review comments. They used three sentiment analysis algorithms Opinion finder, Sentiment Annotator and Sentiwordnet for extracting text comments, the comments showed good correlation with the actual ratings. Predicted algorithm results showed good reliability in comparison to the actual ratings.

[8] developed a relationship between kansei words (consumers psychological feeling and image) and hotel service characteristics using Kansei engineering and text mining methods which was used in the improvement of hotel service guidelines.

[13] provided an insight on conventional hotel rating and online reviews from social media platforms , they presented case studies on how focus is shifting from conventional rating systems to online ratings and user feedback. They predicted in future conventional hotel ratings system will integrate with online media platforms and customer feedback will have more importance in rating systems.

[14] Analyzed hotel ratings and reviews for finding the trends and patterns in the tourism business, they carried out offline analysis using multiple linear regression to find out the relationship between trends and prices related to hotel ratings. Using word cloud, they analysed the positive and negative reviews from the customers. Using this information along with price

and ratings online analysis platform was developed which helps the user to select the top hotels based on price , reviews and ratings.

Light GBM

[9] For high dimensional data the scalability of features were yielded unsatisfactory results by using tree based models such has decision tree, random forest and Gradient boost methods.Even for finding relevant data instances the information gain for all possible splits it is tedious job and time consuming.In order to overcome the author [10] proposed two tree based models such as LightGBM and XGBoost. Both of them are the variants of gradient boost methods that were mainly used to reduce the variance and bias in the datasets. It yielded significant results [11] In his research evaluated boosting algorithms which aimed to convert week learners into the strong learners. In which LightGBM facilitates the high prediction accuracy.

[12] proposed a model for predicting the next destination based on check-in status updated by users in the social media. In this analysis collected spatial,temporal and historical data for determining the next point of location. To minimize the error rate he evaluated using tree based boosting methods such as xGboost, Bagging trees and LightGBM for recommending new places The results obtained for LightGBM model that yielded least error rate.

When encountered with mixed data , categorical data LGBM is performing better than Xgboost as LGBM is designed to handle categorical data and it address the overfitting of models better when compared to Random forest and Xgboost.

Methodology

In this research we have used CRISP-DM data mining approach for this analysis which is generally used in execution of business objectives. This process constitutes of 6 stages such as Business understanding, Data understanding, data preparation, modelling, evaluation and deployment which primarily followed in the analysis. Since travellers trust hosting platforms to book their accommodation, the main goal of this project is to find the relationship between star rating of hotels with services provided by the hotels and written reviews by guests. In this project we also aim to determine the critical attributes responsible in determining changes in overall rating using tree based method LightGBM to get a desired results.

3) Dataset Description

Data Source:

Airbnb is established online marketplace which provides a service to the people for renting/booking the residential properties.It has more than 2 million listings in 191 countries. In this research we have based our study on the Airbnb dataset for Berlin city in germany which is publicly available on internet [Link of the website].The dataset has around 20k listings with 95 features that includes the hotel services and user rating scores. In addition to that which includes text reviews of over 2,66,555 for listings in Berlin. As observed, the dataset consist of mixed proportion of categorical and numerical data types. Somehow we encountered features having more missing values which were eliminated. Meanwhile we filtered relevant features that may be responsible for predictive analysis.

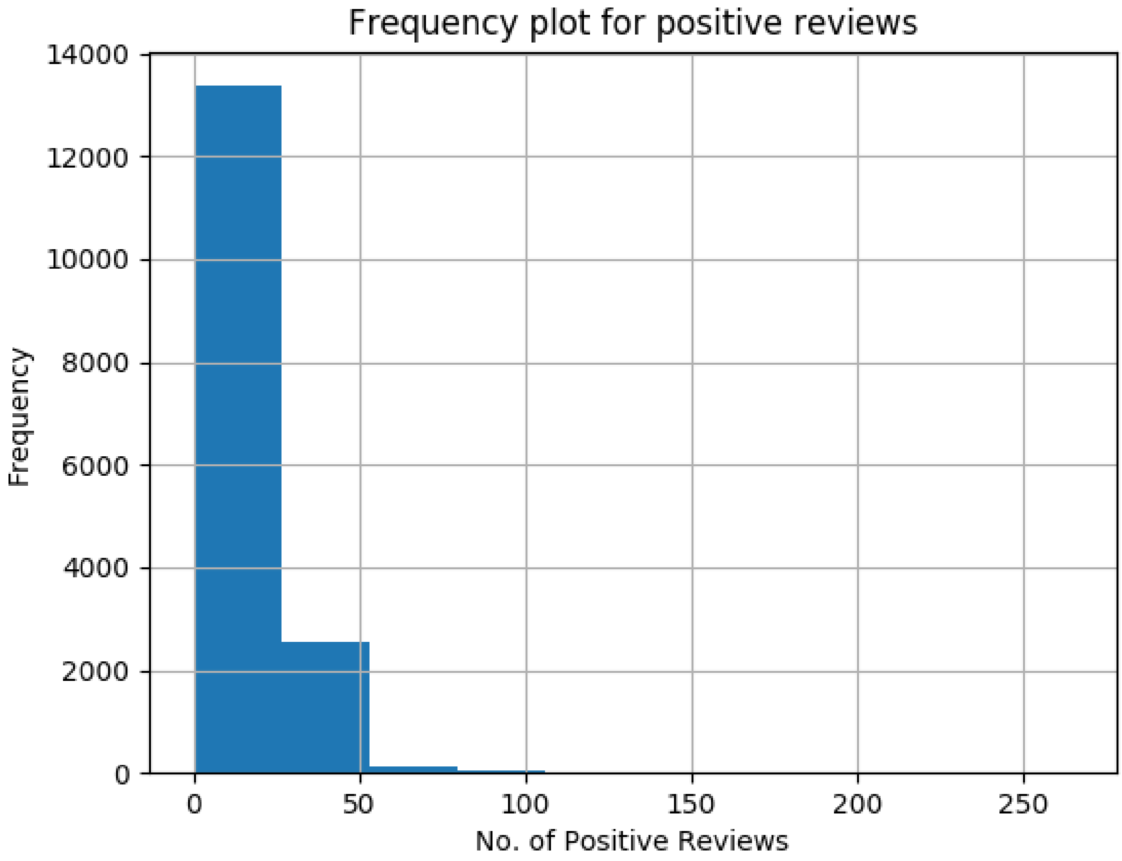
Data Preparation:

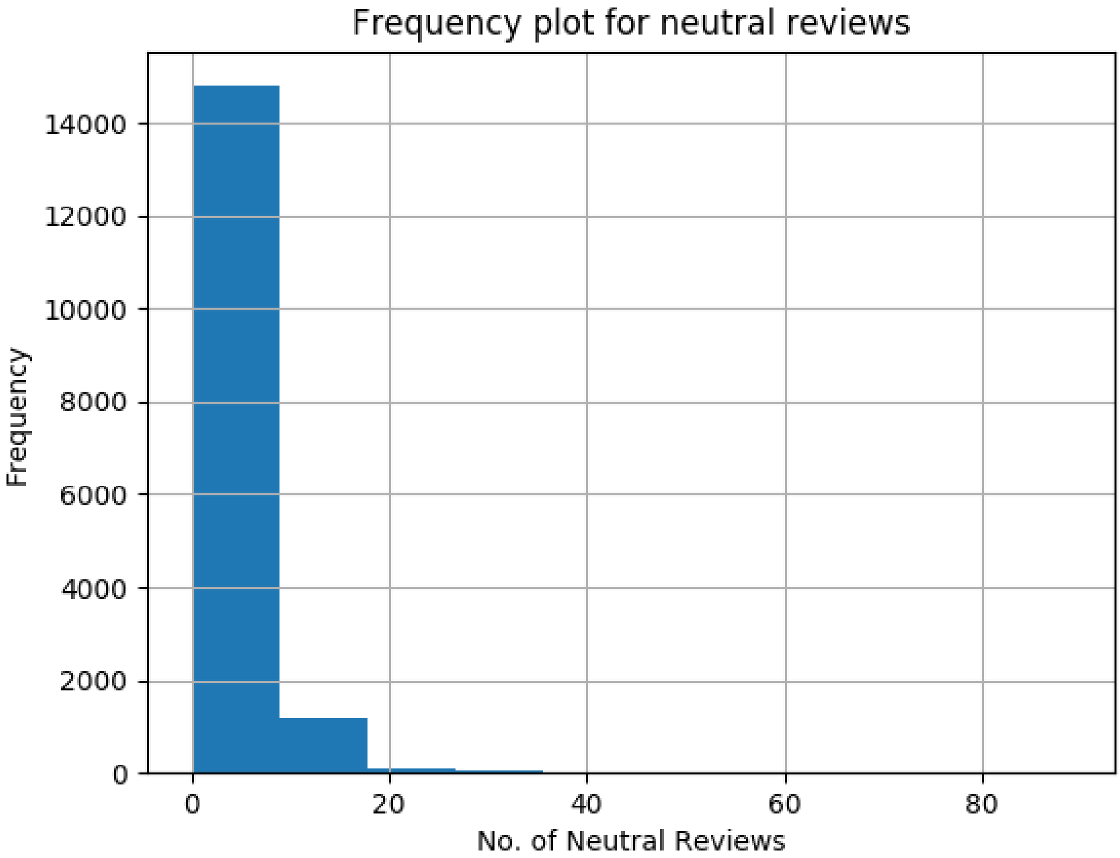
The extracted dataset had 96 attributes and only few attributes were filtered and considered for analysis including the reviews provided by the guest as shown in the below table.Since the selected data consist of missing values, to get rid of it we have imputed with mean and median values for quantitative fields and for some part of listings the text reviews were missing so we eliminated the missing reviews. After cleaning the data we encoded the categorical variables by representing with unique numbers. Variables such as amenities, host verifications and few other variables were numbered with number of services offered. [author].Furthermore we encountered with informative fields i.e. summary, description, hostabout, interaction such field were encode by their text length as taken into consideration.

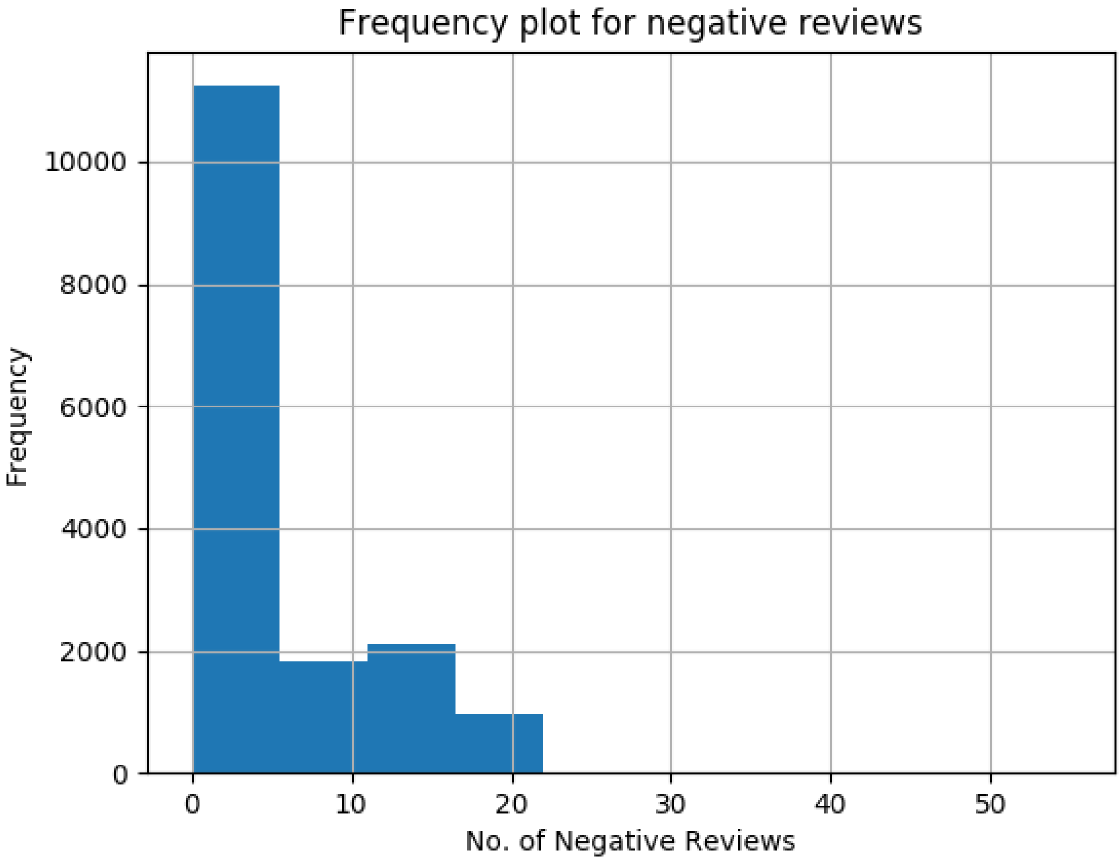
Opinion mining: The collective opinions were acquired from the Text reviews provided by users. Initially the comments were cleaned for special characters and spaces and each review is applied through sentiment function using TextBlob [ ] which calculates the sentiment score. The sentiment function which returns the properties like polarity and subjectivity for each sentence. Based on the polarity scores each reviews are grouped as positive, negative and neutral reviews and finally aggregated for each listings. For rows with missing positive, negative & neutral reviews imputation was applied using mean values for each column. This sentiment orientation was achieved using textblob library using python programming.

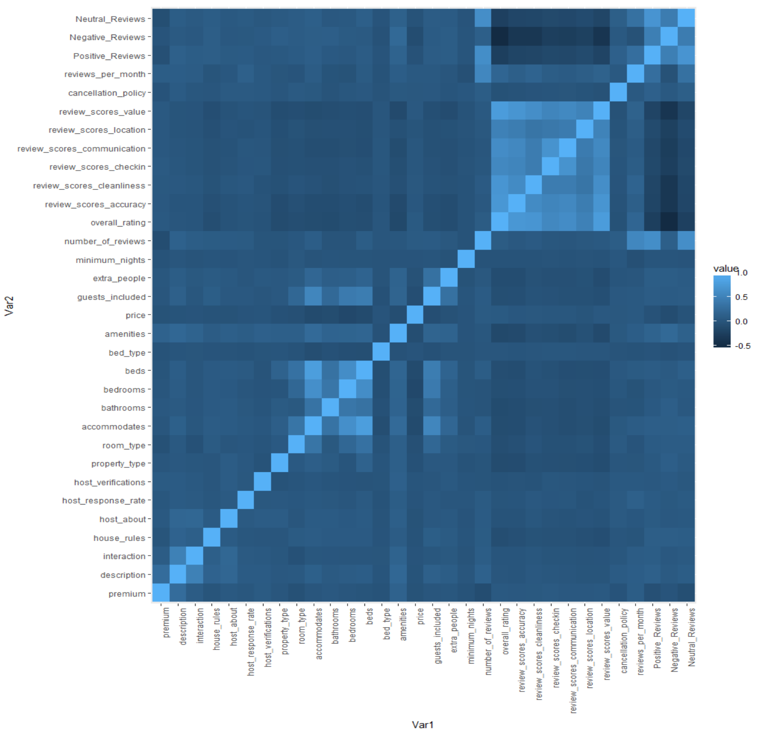
Preliminary Analysis on Statistical aspects

After preprocessing the data, we conducted statistical tests for exploring information from the dataset to accomplish prediction task. In technical jargon, the overall rating is highly dependent across all the variables so considered as targeted variable in this analysis. Initially we considered all the variables in the dataset to find the correlation between the dependent and independent variables. As observed in the correlation matrix the we identified highly correlated variables are identified as shown in the figure[1]. To check the normality of the variables bar graphs are used. The targeted continuous variables are plotted to check their frequency is normally distributed as shown in the figure [2].


Frequency plot for positive reviews
Frequency
14000
12000
10000
8000
6000
4000
2000
0
0
50
100
150
200
250
No. of Positive Reviews


Frequency plot for neutral reviews
Frequency
0
2000
4000
6000
8000
10000
12000
14000
0
20
40
60
80
No. of Neutral Reviews


Frequency plot for negative reviews
Frequency
10000
8000
6000
4000
2000
0
0
10
20
30
40
50
No. of Negative Reviews



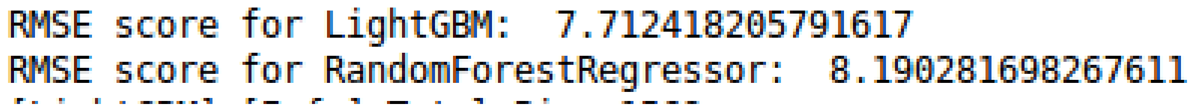
Data Modelling

After the preliminary analysis we observed multicollinearity among the independent variables such as review score cleanliness, review score location, review scores value, and review score communication, positive reviews and negative reviews. Collinearity has a negative impact on predictive power on statistical methods but Tree based methods can handle multicollinearity better than linear models. Further, in our dataset, many features have been derived the length of informative texts and score of negative reviews, positive and neutral reviews that is why we gave precedence to Decision tree based algorithms such has random forest regressor and LightGBM.

LightGBM is a gradient boosting framework that uses tree based learning algorithm.The underlying idea in gradient boosting is to add a regressor at a time, so that the next regressor is trained to improve the already trained ensemble. Random Forest Regressor differs from GBM as in Random Forest each decision tree is trained independently from the rest. LightGBM library is relatively new (released in 2016) and has largely replaced Random Forest Regression in machine learning, as in model the tree grows leaf wise rather than level wise horizontally.For selecting the best fit model for our dataset we compared Light GBM and Random forest.

To evaluate our models,we have divided the dataset into train and test set and performed 5-fold cross validation was used across the training set i.e. training set was divided into 5 validation sets and for each validation set remaining 4 sets are the training sets. The advantage of this method is that reduces the significance of how the split is performed for training & test set. we applied cross fold validation on random regressor and LightGBM algorithms to get a best fit results. But LightGBM outperforms Random Forest Regressor for our dataset with less error rate as show in the figure[4] .

Working with our algorithms helped us in understanding the crucial reasons behind LightGBM superior predictive power. Our dataset has a mix of categorical and numerical variables and light GBM directly supports working with encoded categorical variables which is not the case for most other popular models including Random Forest Regressor. These models require transformation of categorical variables into multiple binary variables using one hot encoding method which leads in high dimensionality and in return negatively impacts predictive performance. This made us choose LightGBM model for this analysis.



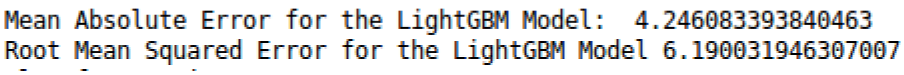
```
RMSE score for LightGBM:  7.712418205791617
RMSE score for RandomForestRegressor:  8.190281698267611
```

Metric Used: Metric used for the evaluation would be the Root Mean Squared Error (RMSE) Mean absolute error (MAE). Root mean square error widely used to verify experimental results of forecasting and regression analysis. It shows how data points are accumulated along the line of best fit (regression line). The measures of the distance of these data points from the regression line is known as residuals (prediction errors) and RMSE is the standard deviation of these residuals. The difference between the measured values and true values is known as Absolute error. It denotes the amount of error in the measurements.

Rating Prediction

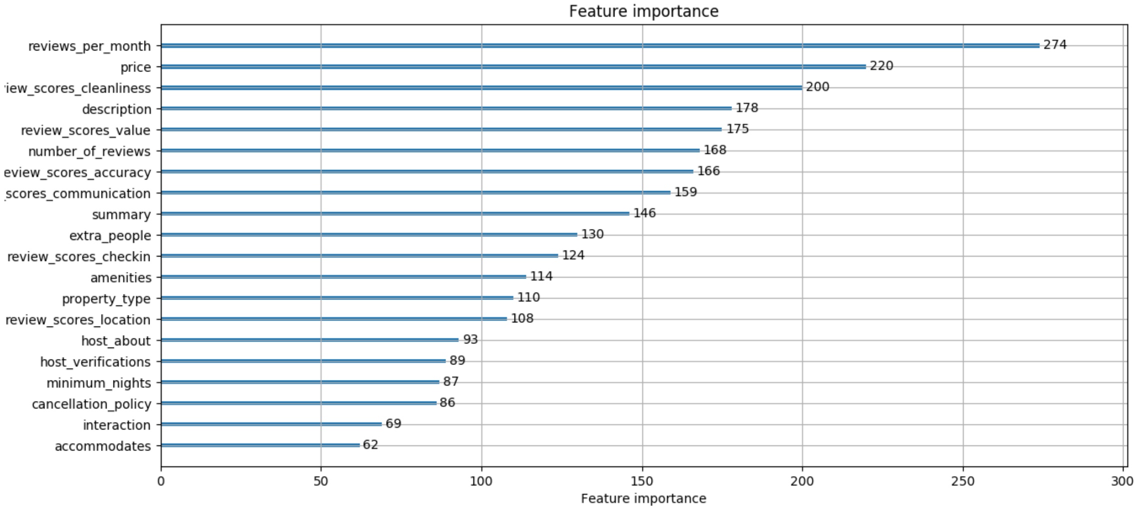
In the last step of our analysis we conducted two experiments by including and excluding the sentiment scores of text reviews in the dataset.
To find out the importance of text reviews with multi- criteria ratings that affects overall ratings prediction:

Case1: We performed analysis on the entire dataset to predict overall rating by excluding sentiment scores of text reviews provided by guests. The dataset is splitted into train and test data as mentioned earlier. We applied LightGBM algorithm on train dataset and validated using testdata. We evaluated our results using root mean square and mean square error rate.

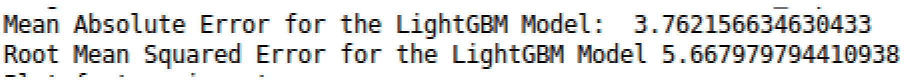


```
Mean Absolute Error for the LightGBM Model:  4.246083393840463
Root Mean Squared Error for the LightGBM Model 6.190031946307007
```

The feature importance plot showcases the relative importance of features for the lightgbm model. The reviews_per_month is the most important variable for this dataset followed by price, review_scores_cleanliness, description & review_scores_value as shown in the figure[]

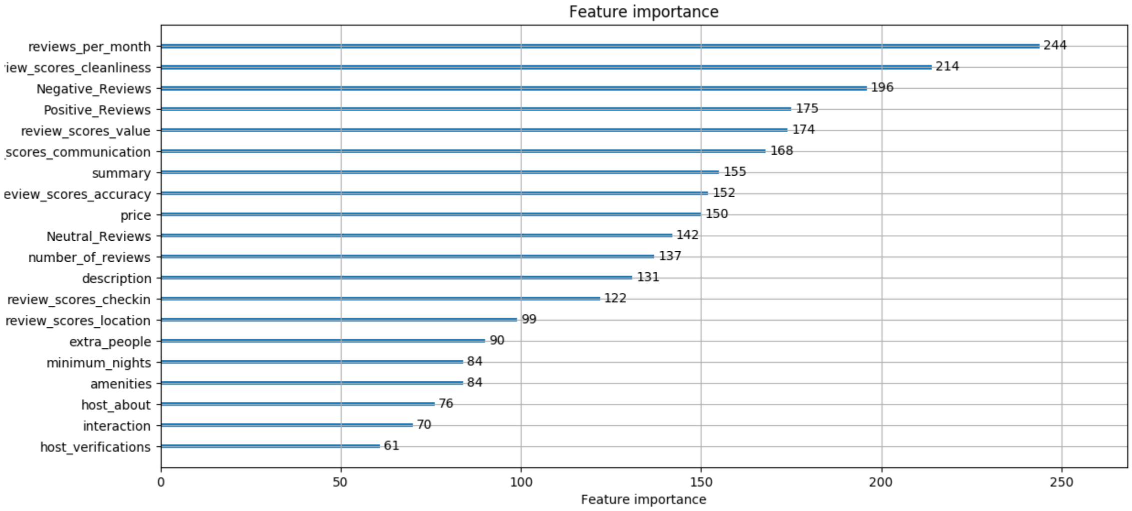


Case2: In this step the dataset was augmented by combining sentiment scores of text reviews with the dataset. The dataset is splitted into train and test as mentioned in earlier. The regression is performed using lightgbm algorithm and obtained least error rate as shown in the below fig [].



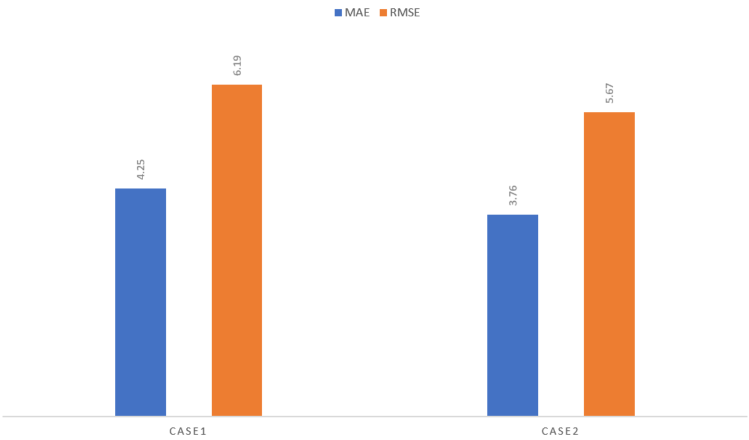
```
Mean Absolute Error for the LightGBM Model:  3.762156634630433
Root Mean Squared Error for the LightGBM Model 5.667979794410938
```

In this step we noticed that 'reviews_per_month' is still the most important variable for this dataset and which is followed by positive_reviews and negative reviews, etc which are the key factors that affecting the overall rating prediction as shown in the below figure[].



As we see from the results case 2 is performing better than case 1 and the variable importance plot showcases that Negative, Positive Reviews are among top 5 most important variables used for creating leaves by the model. Including text review comments from the user has improved the prediction results of overall hotel ratings. Hence, we can conclude that prediction of overall ratings is much more accurate if text reviews are also included as features.

	MAE	**RMSE**
Case1	**4.25**	**6.19**
Case2	**6**	**8.03**



Conclusion:

The prominence of crowd-sourced platforms i.e. Airbnb leads customers to continuously generate and share a huge volume of feedback data on online resources. By using this valuable information we outlined the trends in predicting the overall rating based on online guest reviews and multi-criteria ratings. Initially we extracted sentiment scores from the online guest reviews and represented as positive, negative and neural review scores for each listings in berlin. By using LightGBM algorithm we performed analysis to figure out the impact of online guest reviews for prediction. As observed that Light GBM method performed better with inclusion of online guest reviews with a low RMSE value as compared to multi-criteria rating upon overall rating prediction. So finally the key drivers - cleanliness score, positive reviews, negative reviews and price are most important factors responsible for change in overall rating. we evaluated LightGBM algorithm with mean squared error which having least error rate. From this analysis we can conclude that online guest reviews plays an integral part for rating the hotels. In future we extend our research on finding how the hotel service offering such as amenities, guest included and host rules that affects the positive and negative reviews provided by the guests.

References

[1] Wang, Hongning, Yue Lu, and Chengxiang Zhai. "Latent aspect rating analysis on review text data: a rating regression approach." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACm, 2010.

[2] De Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning. &quot;Generating typed dependency parses from phrase structure parses.&quot; Proceedings of LREC. Vol. 6. No. 2006. 2006.

[3] Tewari, Anand Shanker, and Asim Gopal Barman. "Sequencing of items in personalized recommendations using multiple recommendation techniques." Expert Systems with Applications 97 (2018): 70-82.

[4] Lei, Xiaojiang, Xueming Qian, and Guoshuai Zhao. "Rating prediction based on social sentiment from textual reviews." *IEEE Transactions on Multimedia* 18.9 (2016): 1910-1921.

[5] Yu, Dongjin, Yunlei Mu, and Yike Jin. "Rating prediction using review texts with underlying sentiments." *Information Processing Letters* 117 (2017): 10-18.

[6] Nilashi, M., bin Ibrahim, O., Ithnin, N. and Sarmin, N.H., 2015. A multi-criteria collaborative filtering recommender system for the tourism domain using Expectation Maximization (EM) and PCA–ANFIS. *Electronic Commerce Research and Applications*, *14*(6), pp.542-562.

[7] López Barbosa, R.R., Sánchez-Alonso, S. and Sicilia-Urban, M.A., 2015. Evaluating hotels rating prediction based on sentiment analysis services. *Aslib Journal of Information Management*, *67*(4), pp.392-407

[8] Y. H. Hsiao, M. C. Chen, and M. K. Lin. ”Kansei Engineering with Online Review Mining for Hotel Service Development.” In 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI) (pp. 29-34). IEEE, 2017.

[9] LightGBM: A Highly Efficient Gradient Boosting Decision Tree

[10] Deep Embedding Forest: Forest-based Serving with Deep Embedding Features

[11] Applications of Python to Evaluate the Performance of Decision Tree-Based Boosting Algorithms

[12] Novelty Detection for Location Prediction Problems Using Boosting Trees

[13] Hensens, W., 2015. The future of hotel rating. Journal of Tourism Futures, 1(1), pp.69-73.

[14] Leal, F., Dias, J.M., Malheiro, B. and Burguillo, J.C., 2016, July. Analysis and visualisation of crowd-sourced tourism data. In *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering* (pp. 98-101). ACM.

Conclusions :The prominence of hosting platforms i.e. Airbnb, allows its users to share feedback via online sources. By using this information we could outline the trends in predicting the overall rating based on online guest reviews and multi-criteria ratings. Initially, we extracted sentiment scores from the online guest reviews and represented them as positive, negative and neutral review scores for each listing in Berlin city. By using LightGBM algorithm we performed analysis to figure out the impact of online guest reviews for prediction. As we have observed, Light GBM method has performed better with the inclusion of online guest reviews with a low RMSE value compared to multi-criteria rating upon overall rating prediction. So finally the key drivers are - cleanliness score, positive reviews, negative reviews, and price. Price being an important factor responsible for the change in overall rating based on multi-criteria but cleanness had a better effect when the comments were added . From this analysis, we can conclude that online guest reviews play an integral part in rating the hotels. In future, we extend our research on precisely giving a score to hotel services such as amenities, guest included and host rules that affect the positive and negative reviews provided by the guests.`