Hotel Review Analysis

EDA and NLP Sentiment Analysis

Lillian Leong

Table of Contents

1. Background

The client is an established hospitality group who runs several hotels across Europe. They noticed that the occupancy rates have suffered recently and their book out rate has been sluggish. The senior management received feedback from their hotel managers that their business has suffered due to backlash from their hotel guests, citing that the hotel fell short of expectations resulting in drop in referrals.

The hotelier had faced frequent staff turnover and the workforce has always been supplemented by part time workers especially during peak seasons in summer and yearend. Therefore, they were unable to nail the issues brought up by guests as different staff would provide different feedback based on their interactions with the hotel guests.

The Customer Experience Director has tried doing some research on their own by visiting booking websites but they realised that there were conflicting reviews left on the travel websites such as www.booking.com. They were unable to summarise any statistics efficiently as it was too tedious to read through all the reviews online page by page.

With that, the client would like us to provide some insights as follows:

- How the hotels across Europe were doing as a whole.
- Which traveller's nationalities are the reviewers and is there any differences in rating score across different nationalities.
- Apply text analysis on the review text to assess the sentiments.
- Is there any common topics/ characteristics in the negative reviews.
- Whether the guest is likely to recommend the hotel

They hope to pinpoint the areas of improvement through our consultancy services and allocate funds to resolve the issues but they were unsure what to invest in:

- More marketing and promotion activities to create brand awareness
- Renovation uplift as there were several hotels which were quite old and rundown but they were also unsure which areas to spend on.
- Staff training so they are intuitive to customer's demands.

2. Stakeholders & Business Question

The stakeholders in this project would be mainly the hotel shareholders, the key management who is operating the business. The immediate stakeholder is the Customer Service Experience Director who has been in touch with the hotel staff and guests, as well as having read some of the reviews from www.booking.com website.

With our firm's service and data science tools, they can focus on the operations – to better manage the value chains which will result in higher satisfaction across. The secondary stakeholders would be the hotel guests since the firm will be able to gain more insights from the reviews and able to understand the pressing issues amongst these customers. The end customer would also benefit from this project indirectly as the project can help to identify the popular demand and avoid unpleasant situations or unsatisfied, resulting in higher customer satisfaction.

The key business question in this case is to empower the stakeholders with business insights as follows:

- How the hotels across Europe were doing as a whole in terms of rating score?
- Which countries are the guests from and is there any differences in rating scores across different nationalities?
- Any other characteristics with guests giving bad reviews?
- Apply text analysis on the review text to assess the sentiments of the review text.
- Is there any common topics/ characteristics in the bad reviews

3. Data Question

From the sample reviews extracted from the website, it was observed that that the review scores between 7 to 9 might were considered good, hence the midpoint of 8 is used as the threshold to determine if the review is considered good or bad.

The project will also include tokenisation and ngram analysis with TFIDIF to find out which are the more prominent topics so as to better understand the primary concerns in the guest reviews.

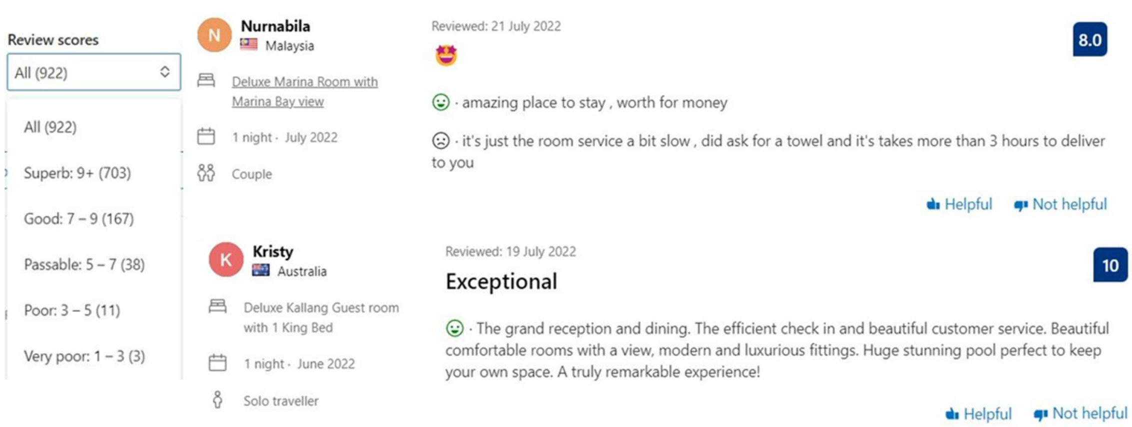
Last but not the least, an classifier model will be built to assess the guest's sentiment of the review, whether they are likely to recommend the hotel or not.

4. Overview of Dataset

The data is published on Kaggle, the original author has performed webscraping from www.booking.com website.

The source of dataset: https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe

A sample of the review has been extracted from the website.



This dataset contains 515,000 customer reviews and scoring of 1493 luxury hotels across Europe. The file has 17 columns with description as below:

Field	Description
Hotel_Address	Address of hotel.
Review_Date	Date when reviewer posted the corresponding review.
Average_Score	Average Score of the hotel, calculated based on the latest comment in the last year.
Hotel_Name	Name of Hotel
Reviewer_Nationality	Nationality of Reviewer
Negative_Review	Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Negative'
ReviewTotalNegative WordCounts	Total number of words in the negative review.
Positive_Review:	Positive Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Positive'
ReviewTotalPositive WordCounts	Total number of words in the positive review.
Reviewer_Score	Score the reviewer has given to the hotel, based on his/her experience
TotalNumberofReviews ReviewerHasGiven:	Number of Reviews the reviewers has given in the past.
TotalNumberof_Reviews	Total number of valid reviews the hotel has.
Tags	Tags reviewer gave the hotel.
Dayssincereview	Duration between the review date and scrape date.
AdditionalNumberof_Scoring	There are also some guests who just made a scoring on the service rather than a review. This number indicates how many valid scores without review in there.
Lat	Latitude of the hotel
Lng	longtitude of the hotel

5. Data Science Process

a. Preliminary Data Preparation

There are two sections to the project:

- Exploratory Data Analysis (EDA) on dataset to provide insights from the business aspects
- Text analysis, particular sentiment analysis on the review text.

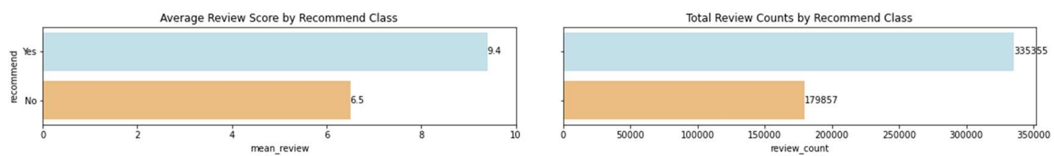
Feature engineering was performed as EDA was performed.

i. Duplicates and Null Values in Dataset:

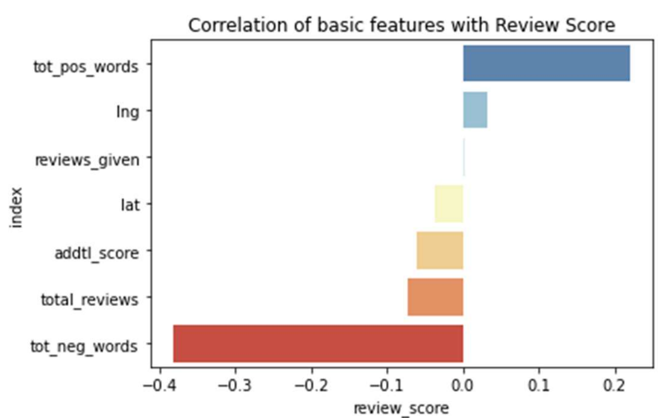
Firstly, 526 duplicates were removed from the original dataset of 515,738 records. Thereafter, there were 3,268 null values in the latitude and longtitude information which were patched by matching the hotel name against the web resources; the values using the website http://latlong.net/

ii. Recommend Indicator

For the purpose of this project, the reviews above 8/10 would be classified as recommend and anything below as would not recommend. The resultant distribution is:

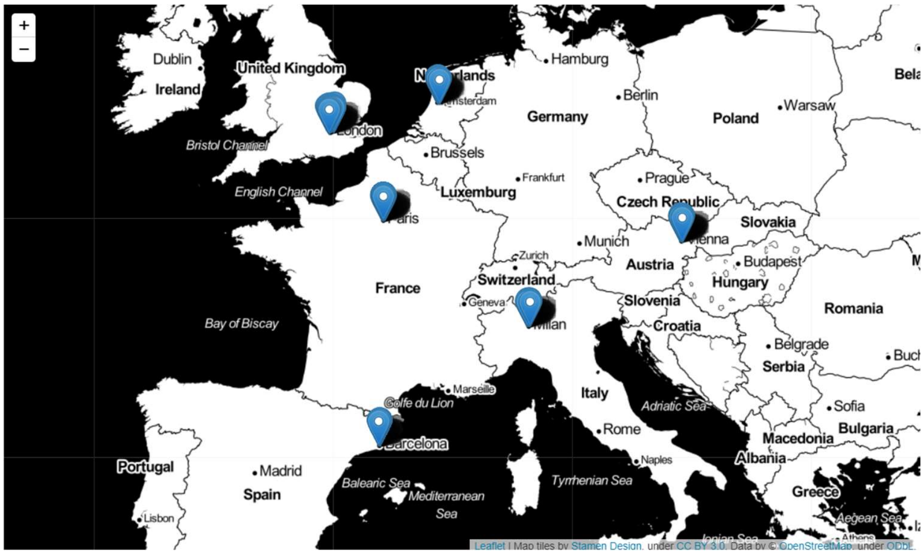


There is about 65% of instances which relates to recommend and another 35% which is not recommend. From preliminary assessment of the correlation map, it is observed that the total negative word count has a higher correlation to the review score as compared to the total positive word count which means that negativity sentiment seems to correlate more with the review scores.

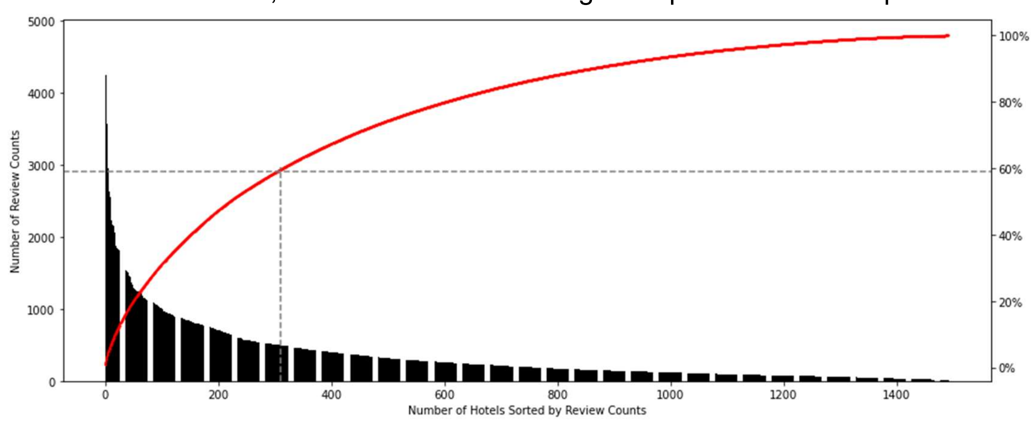


iii. Overview of Hotels

To visualise the locations these hotels, a folium map was used and it was observed that most the hotels were located in the 6 major cities, namely, Amsterdam, London, Paris, Vienna, Milan and Barcelona.

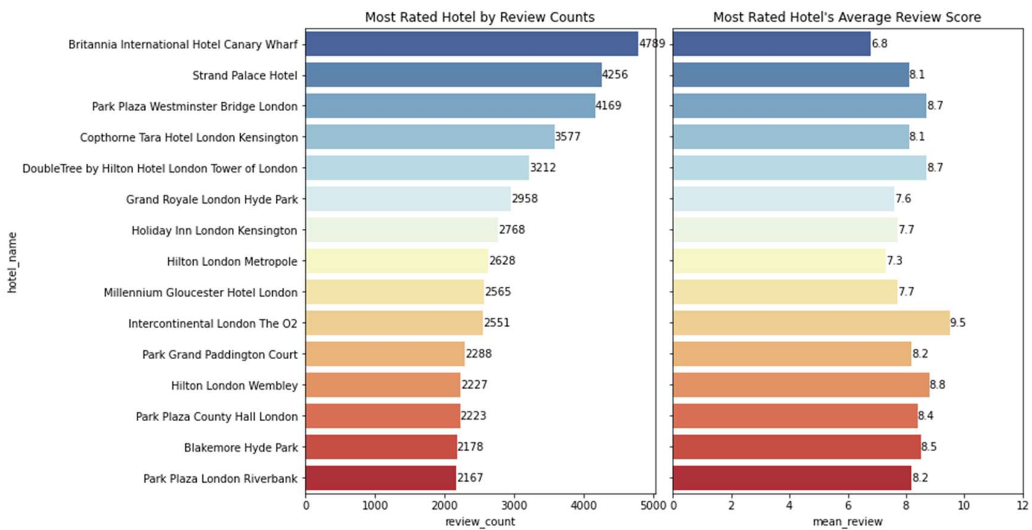


With the Pareto Chart of all the 1,493 hotels, the top 300 hotels in terms of review counts only account up to 60% of total 512k reviews and the 300th hotel only had about 300 reviews in the dataset. It would appear that there were many hotels with smaller review counts, hence no outlier handling were performed at this point.



iv. Most Rated Hotel and Average Reviewer Score

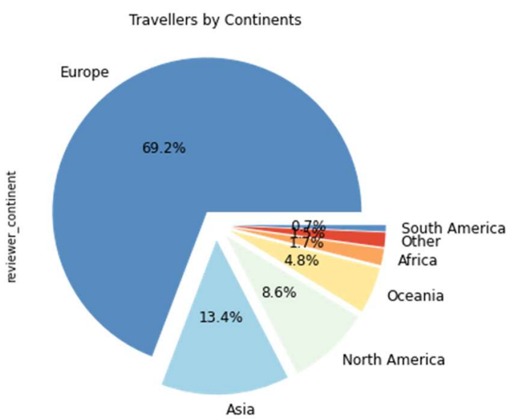
It was observed that the most rated hotel, Britannia International Hotel Canary Wharf with almost 5k review counts had the lowest average review score compared to other most rated hotels. In the other hand, the best rated hotel, Intercontinental London The O2 had only half the review counts at 2.5K. Therefore, there is a difference between most rated vs best rated hotel.

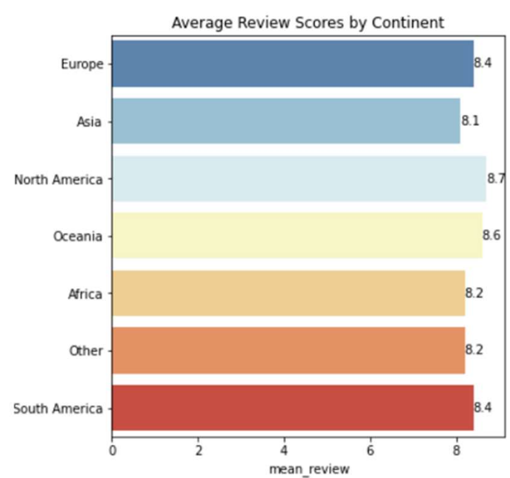


v. Hotel Guests' Geographical Profiles

From the reviewer's nationalities, the continent column was featured engineered by looking up their nationalities against a continent dictionary[1],

From the continent analysis, it was observed that 70% of the reviews were from guest nationalities in Europe, followed by 13% from Asia. However, the average review scores were not the highest from Europe guest but from North America and Oceania. It was also noted that Asia travellers had rated lowest.



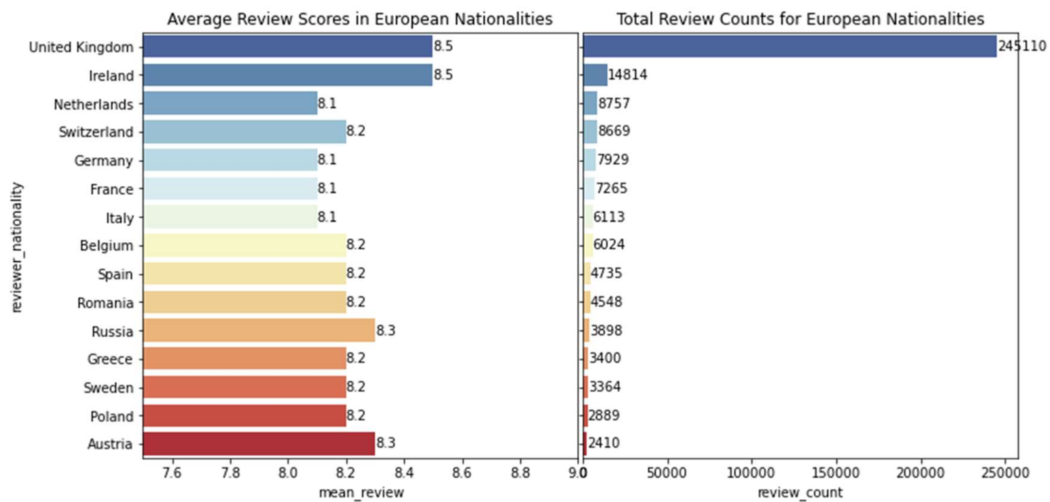


vi. Breaking down Hotel Guests - European Nationalities

As Europe took up 70% of the guest reviews, the dataset was broken down further to see where were the guests coming from. Most of the guests came from United Kingdom and Ireland who had rated higher than all their counterparts. The lowest

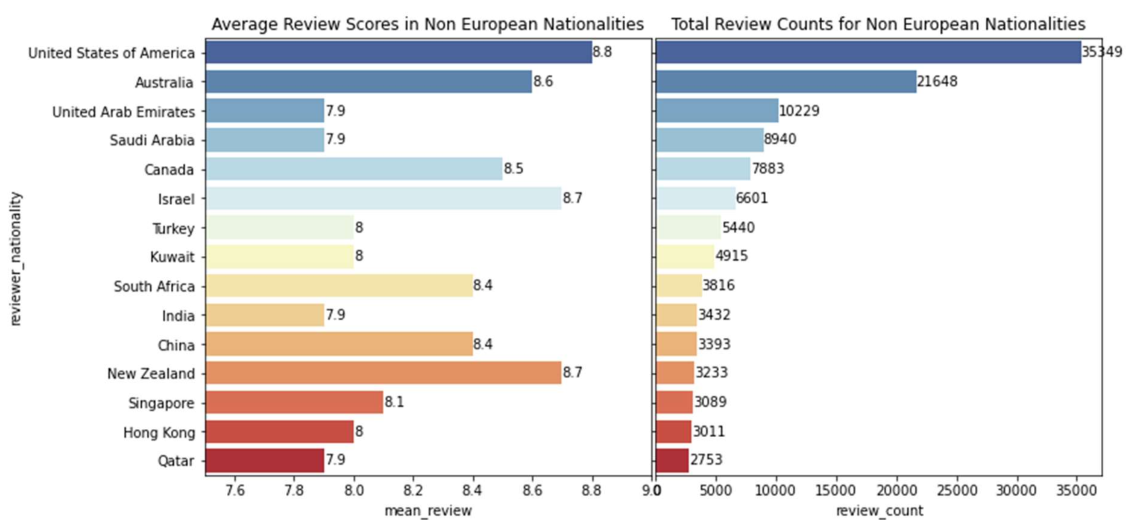
[1] https://gist.github.com/Desperado/3293395#file-countryinfo-py

rating reviews came from Netherland, Germany France and Italy. It is also interested to know that Russia had rated relatively higher as well.



vii. Breaking down Hotel Guests – Non European Nationalities

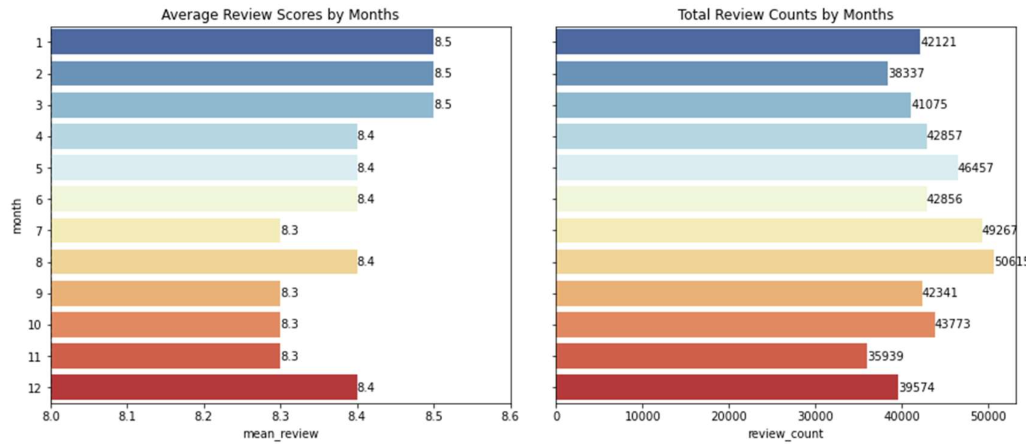
Outside Europe, American guest form the bulk, followed by Australia and UAE guest. The higher review score appears to be coming from English speaking countries like America, Australia, New Zealand while the lower review scores came from India and Middle East countries (UAE, Qatar, Saudi Arabia). It is also interested to know that Singapore has rated lower than China with both having similar review counts.



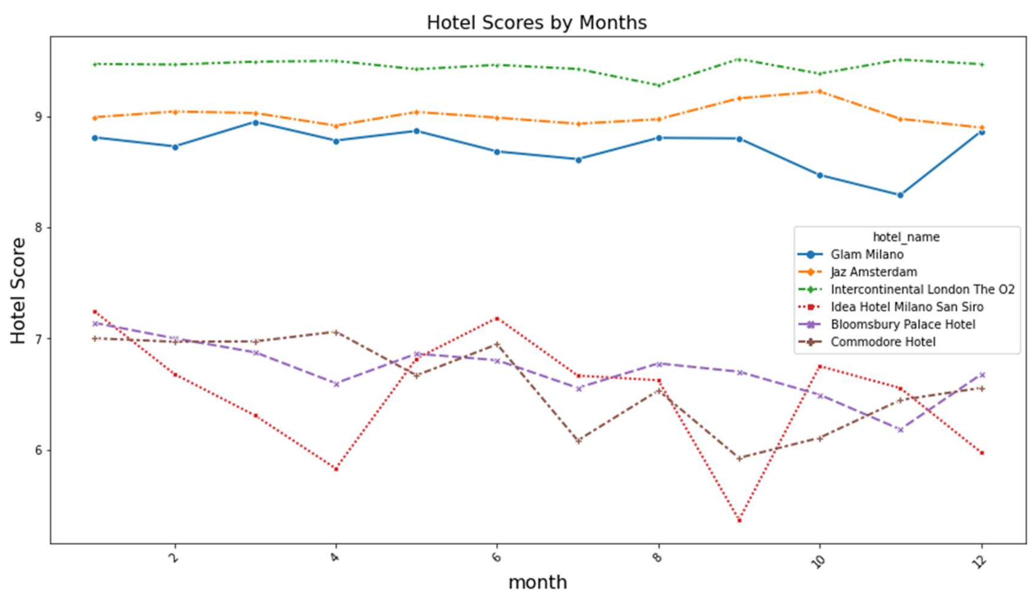
viii. Review Months

It was observed that the rating score tends to peak highest at the start of the year and starts to drop as the months passed and bottomed out in the months starting

September-October. The reverse is observed for review counts. This might suggest that the rating score were affected during busy travelling periods.

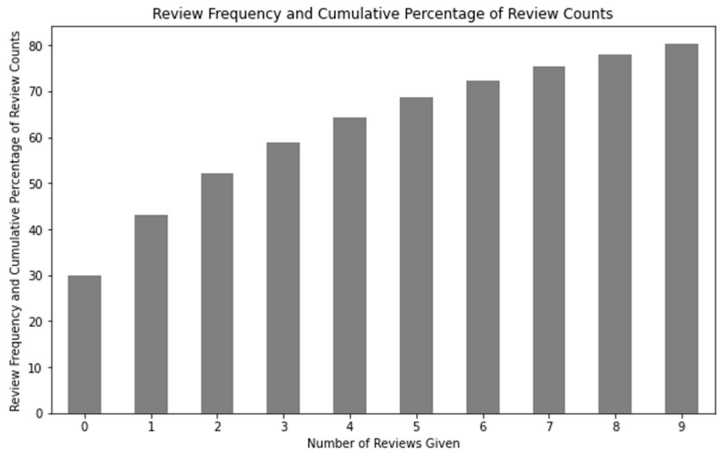


To add on, it was observed that hotels with higher review scores tends to be more consistent throughout the months while the hotels with lower review scores tends to display more fluctuations. That could suggest that the guest experience is more consistent resulting in a more consistent review scores for better hotels.



ix. Review Frequency

About 30% of the guests was giving the reviews for the first time and up to 60% of the guests had given reviews at least thrice previously which might suggest the amount of new and repeat customers that was funnelling through the booking websites.



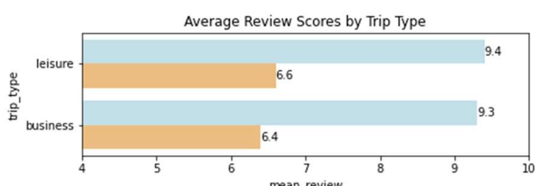
The records with review frequency > 51 were removed as outlier handling. The records for review frequency were 5,408 records out of the 500k records, so 99% of the dataset can be kept.

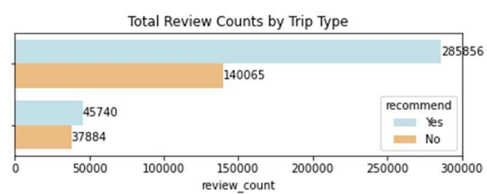
b. Feature Engineering

From the dataset, I learnt that there was additional information from the column tags, where I can extract trip type, traveller type, room type, stay duration.

i. Trip Type

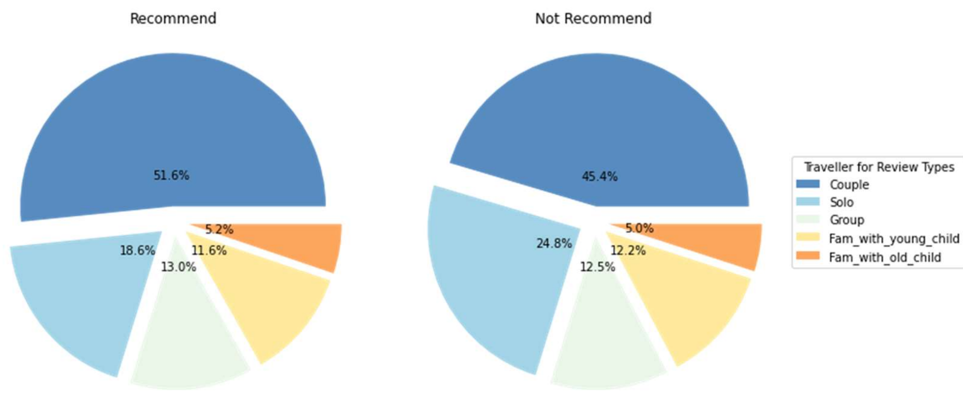
There are 2 categories within the trip type – Leisure and Business. There appears to be not much difference in the mean score between the 2 trip types. However, it is noted that there were 6 times as many leisure travellers than business travellers who had rated good reviews (review score above 8) while there is 4 time as many leisure travellers than business travellers who had rated bad reviews.





ii. Traveller Type

It appears that there are slightly more solo travellers than other segments that had given bad reviews and couple travellers tend to give good reviews.



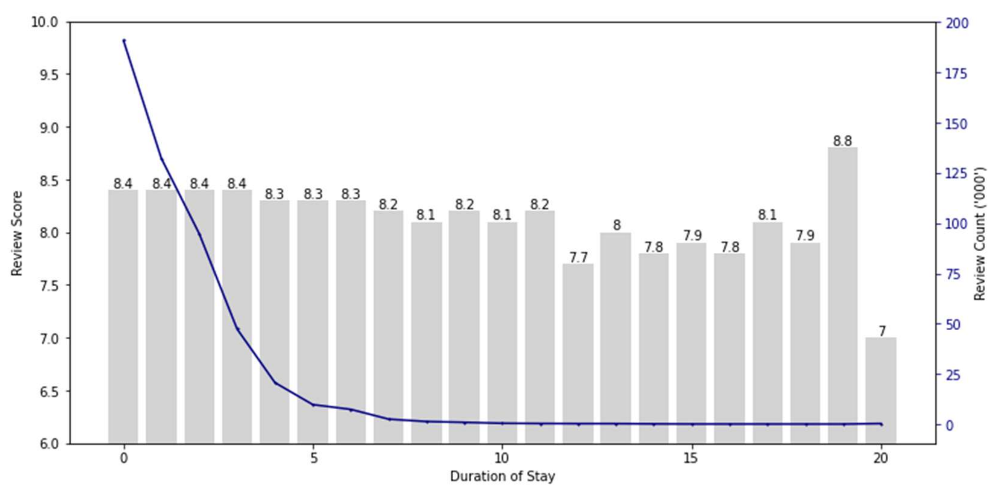
iii. Room Type

There is no analysis done on this information because different hotel has different room category and it would lack a common basis of comparison.

iv. Stay Duration

The rating score tends to be consistent for shorter trips but starts to drop every 3-4 days as they stay extend and dips drastically once the duration stretches beyond 10 days.

There was an outlier towards day 20 which was caused by outliers. Out of the 19 instances for duration of 20 days, almost half of them rated their stay more than 9/10.



c. Text Classification Analysis

For the purpose of text analysis, 10% of the dataset were sampled from the dataset.

i. Pre Processing Tasks

The dataset had the positive and negative reviews kept separately hence data cleaning had to perform separately on both the negative and positive comments separately so that no important information are missed out before combining it.
It was observed that, when the guest leaves the “negative review” or “positive review” empty, the system will auto populate “No Negative” and “No Positive” in the dataset. These values were replaced with null in these instances.

Thereafter, the “negative review” and “positive review” were merged to form the “combine review text” before the below text cleaning function was applied to the dataset:

- Transform the text into lower case
- Tokenize the text into words and remove the punctuation
- Remove useless words that contains numbers.
- Remove useless words like “the”, “a” which are in the stopwords
- Tag Part of Speech (POS) whether the word is a noun or verb using the WordNet lexical database.
- Lemmazation of the text by transforming the word into their root form.

Below is an example of the text cleaning process.
The “negative review” before text cleaning:

```
' Very expensive parking When I booked a 3 adult room I were s
queezed to a 2 people room where the extra bed barely fit the b
ooking details were very misleading Had to complain and argue w
ith the staff to finally get upgraded All and all painful expe
rience '
```

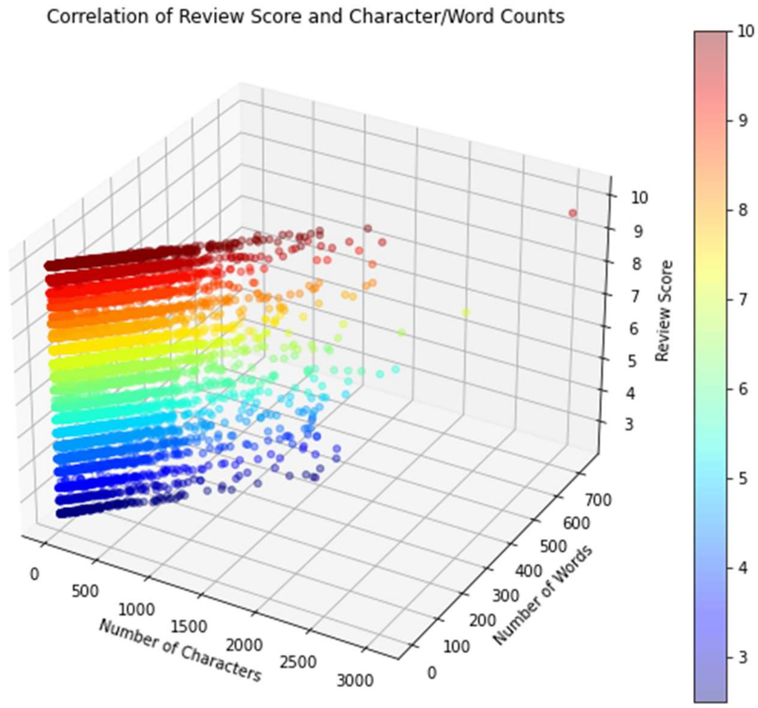
The negative review after text cleaning:

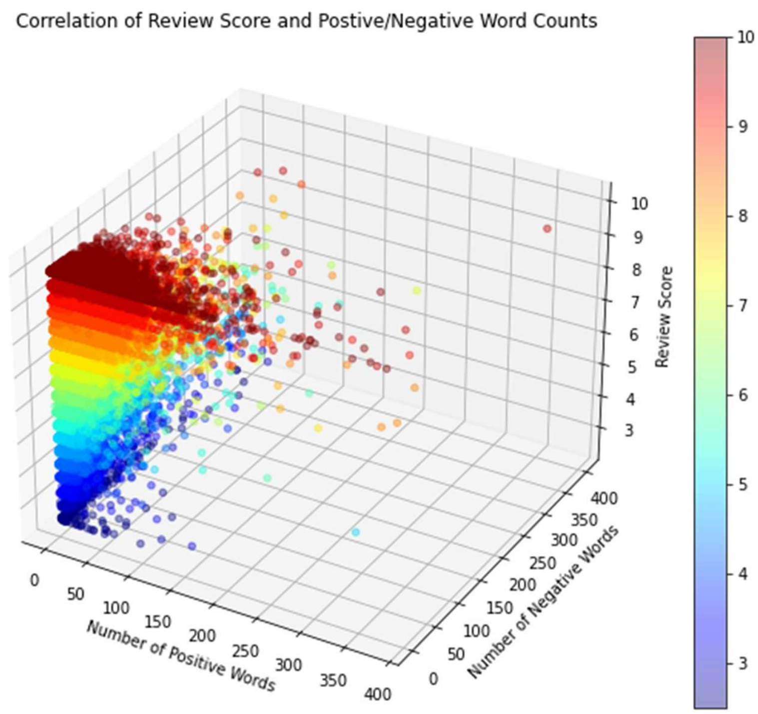
```
'expensive parking book adult room squeeze people room extra b
ed barely fit booking detail mislead complain argue staff fina
lly get upgraded painful experience'
```

ii. Count columns

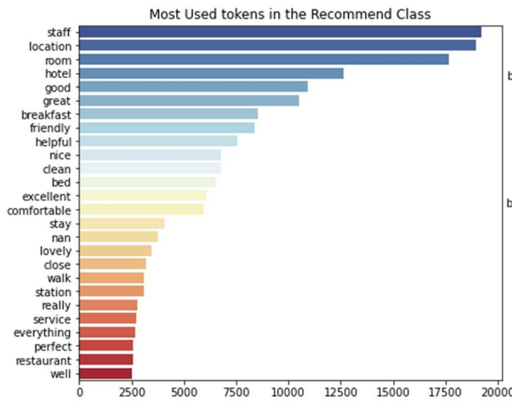
Count columns for the number of characters and number of words in the review text were added for further analysis. The review text were at least 150 words in length. It was observed that review text which are higher scoring (with warmer colour tones) appears to be longer, at least 300 to 400 words while the bad reviews are usually shorter (100 to 200 words).

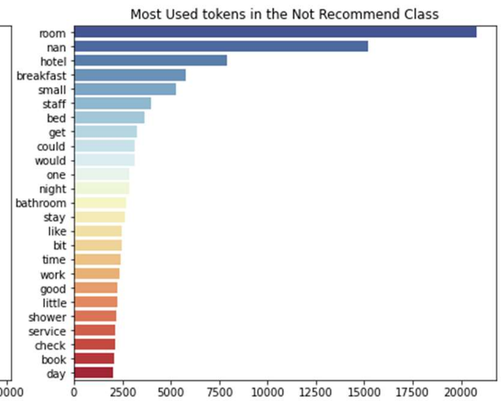
It was also observed that there are more outliers for higher scoring reviews with high count of negative words as well





The commonly used tokens were extracted from the reviews and their frequency are illustrated below. It was observed that there were many null values (nan) in the bad reviews. That is due to the data cleaning for “no negative” when the guest did not leave any reviews in the negative review text which suggest that many guests do not have much to mention in the negative review field.

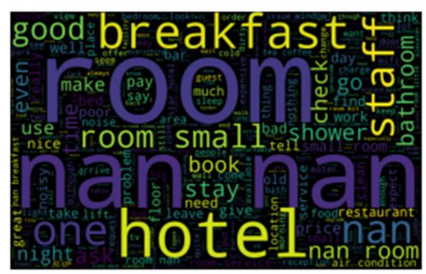




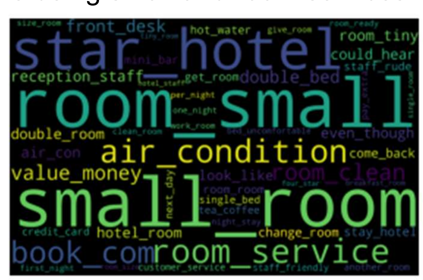
However, some of the tokens does not give context hence bigrams and trigrams were extracted to better understand the common issues underlying the good and bad reviews.

The following word clouds were created to illustrate the above.
The good reviews are in white background while bad reviews are in black background.

The tokens in good and bad reviews:



The bigrams in good and bad reviews:
It was learnt that the good reviews surrounds friendly staff and the great location while bad reviews focuses on the room size being small and room services.



The trigrams in good and bad reviews:
The good reviews echo the same message in the bigrams that the hotel is within walk to the guests' destination while bad reviews started to mention "two single

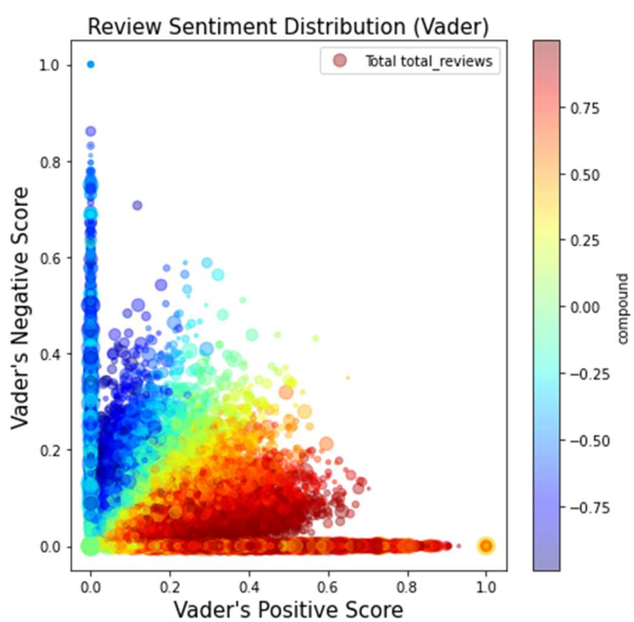
bed” and “air conditioner”, suggested that the room setup might have fallen under guests’ expectations.

iii. Sentiment Analysis with Vader

Using the original review text, Vader sentiment algorithm was applied to the review text, to assess the overall sentiment behind the review. Vader uses a lexicon of words to find which ones are positives or negatives. It also takes into account the context of the sentences to determine the sentiment scores. For each text, Vader returns 4 values:

- a neutrality score
- a positivity score
- a negativity score
- an overall score that summarizes the previous scores

It was generally quite accurate in detecting the sentiments behind the review scores; there was a clear distribution between the Vader’s negative and positive score and it translates to the compound score.



Below were a few examples of how Vader analyse the positive review text:

text	pos	cmp
A perfect location comfortable great value	0.931	0.9260
Clean comfortable friendly helpful staff	0.923	0.9001
Friendly Smiling Efficient Helpful staff	0.922	0.8957
Efficient friendly smiling fair price	0.919	0.8834
More friendly I hope Smile	0.904	0.8568

Below were a few examples of how Vader analyse the negative review text:

text	pos	cmp
poor facility poor staff poor management	0.756	-0.974
Poor staff Poor gym Poor room	0.756	-0.851
dirty and tired breakfast not good	0.732	-0.802
Very bad atmosphear noisy weird smells	0.713	-0.838
No things Bad room Bad bathroom Bad	0.707	-0.926

iv. TFIDFVectorizer (Term Frequency - Inverse Document Frequency)

TFIDF was used to build a vocabulary bank for the frequently appearing words in each review text, where the relative importance of the words in the text. The rationale is a word that appears in almost every text would not likely bring useful information for analysis. On the contrary, rare words may have a lot more of meanings.

The TFIDF yield about 166 additional features, consisting unigram, bigrams, trigrams. The algorithm was also set such that it does not consider any elements that appears in less than 2% of the document, to reduce the size of the output.

An example of the TFIDF output in one of the rows as extracted:

	access	air	air condition	airport	location	location excellent	location good	location great
51515	0.0	0.0	0.0	0.0	0.191186	0.0	0.474625	0.0

6. Modelling

i. Features Used

The feature columns from TFIDF and sentiment score elements from Vader to form the feature columns for modelling. The target columns would be “recommend” which was derived from the review score. If the review score is less than 8, it would be considered “not recommend”.

ii. Models Used

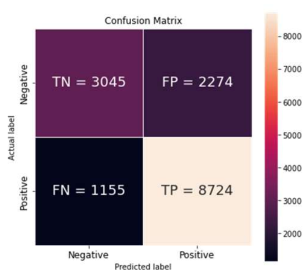
For this classification problem, the following classifiers were used:

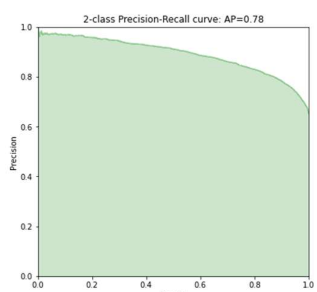
- Logistic Regression
- Naïve Bayes Classifier
- Random Forest Classifier
- XGBoost Classifier
- LightGBM Classifier

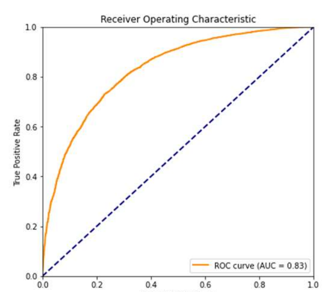
iii. Evaluation of Models

Logistic Regression was used as the base model and yield the following results:

Accuracy	Precision	Recall	F1	ROC_AUC	Eval_time
77.4	79.3	88.3	83.6	82.8	1.6

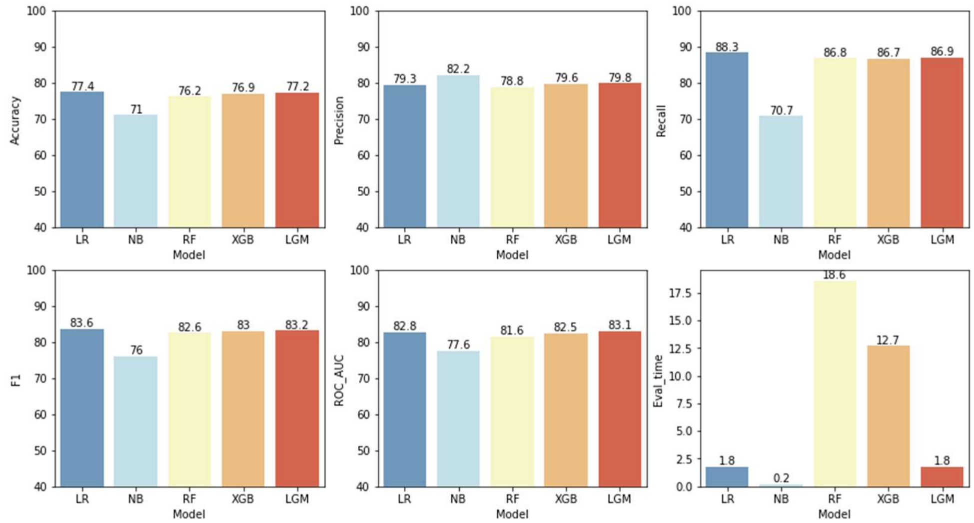






Naïve Bayes and other ensemble models were used for further testing.
The ensemble models XGBoost and LightGBM model’s performance were quite comparable to Logistic Regression as well, especially F1 score and the time taken.

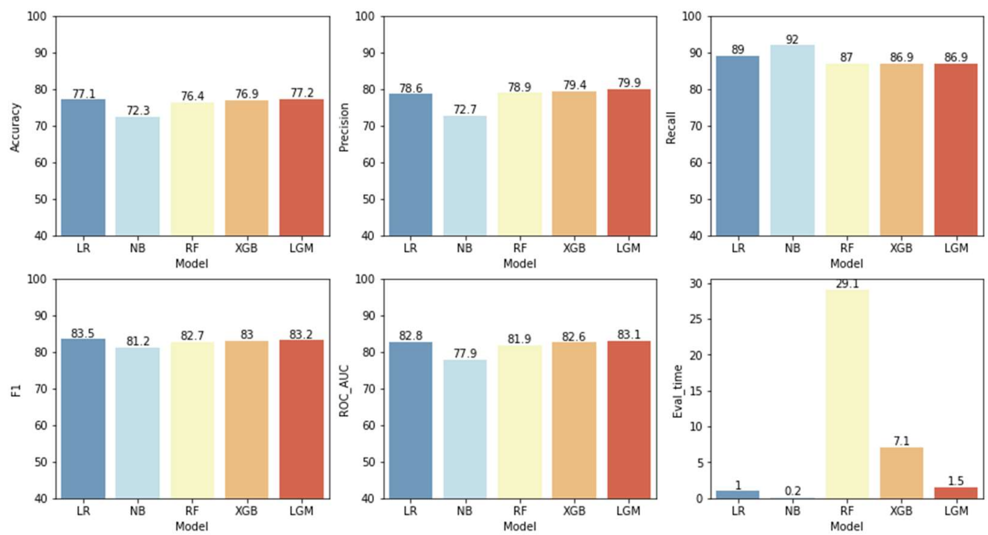
The graph below showed the performance metrics for the base models before hyperparameter tuning:



Thereafter, RandomSearchCV was used during the parameter tuning; the RandomSearchCV was mainly to increase the number of estimators in the ensemble models. However, it did not yield much improvement.

Logistic Regression topped the F1 score though the scores were all very near each other. The final metrics are tabulated in the below illustration

This graph below now shows the revised metrics after hyperparameter tuning. Unfortunately, the RandomSearchCV did not yield much improvement.



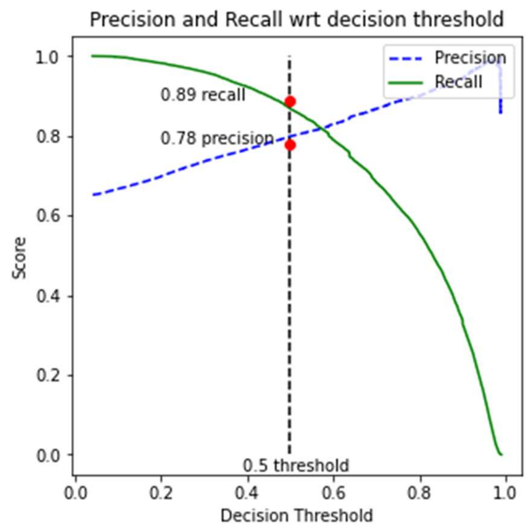
However, to finalise the model's decision threshold to proceed, the precision recall tradeoff issue needed to be understood.

iv. Precision and Recall tradeoff

With precision and recall having an inverse relationship, the precision decreases while the recall increases. The decision threshold needs to be customise according to the precision/recall tradeoff preference. If the goal is to detect most observations of the positive class (true), then high recall (but low precision) is set.

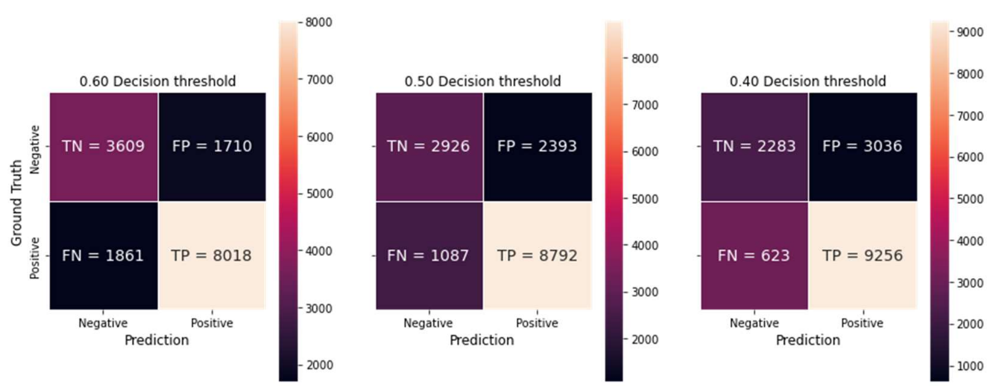
On the other hand, if the focus is on being confident about our predictions but does not matter about not finding all the positive observation, a high precision threshold (but at the expense of a low recall).

By shifting the decision threshold from 0.50, the precision/recall point of the model can be shifted and the confusion matrices would look different.



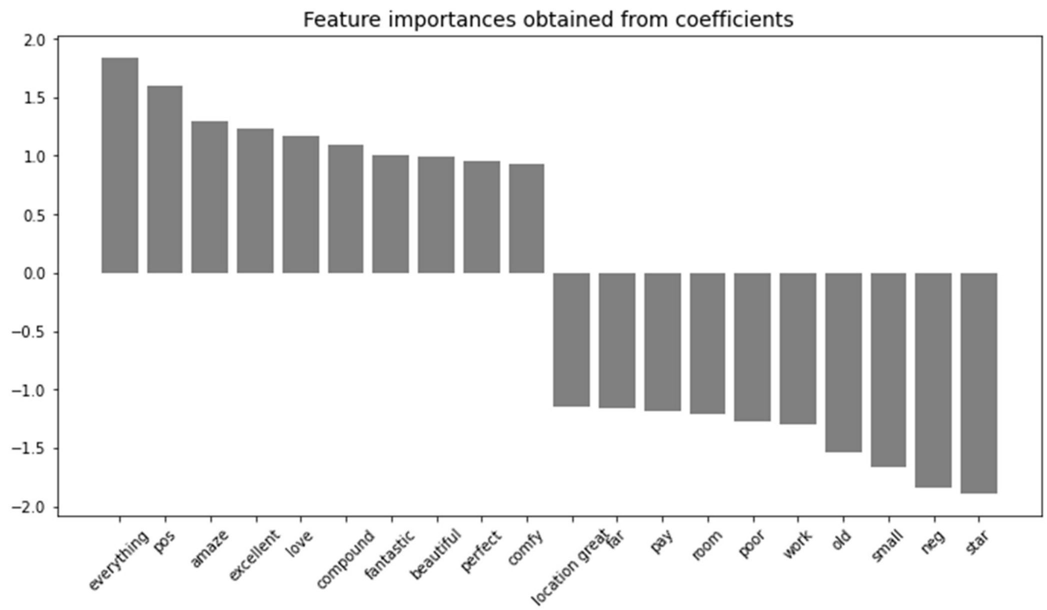
v. Confusion Matrices & Different Thresholds

To illustrate this, the confusion matrices at different thresholds were computed as below. For every 0.10 downshift in the decision threshold, there was a drop of 500 false positive but at the expense of an increase of false negative.



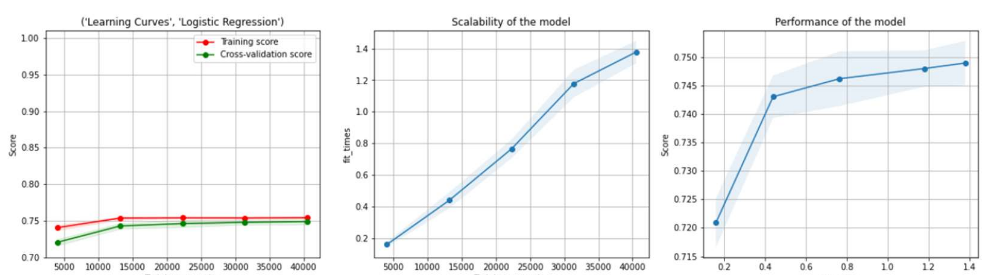
vi. Feature Importance of the Model

From feature importance of Logistic Regression, it was observed that the number of words in positive/negative reviews as well as the vader scores played an important role. Further, the TFIDF word vocab of “room”, “location”, “small”, “staff” and “breakfast” also came up which suggest that these are the aspects that the guests places importance.



vii. Cross Validation & Learning Curves of the Model

A cross validation fold size of 5 was performed and there were no indications of overfitting or underfitting of the tuned Logistic Regression model.
It was evident from the learning curves below that, the training sample size appears to be appropriate as the curves have converged though the model is computational expensive to be scalable in huge datasets.



viii. Summary of Modelling Process

- Models used: Logistic Regression and Naïve Bayes before applying Ensemble Models (Random Forest, XGBoost, LightGBM).

- Logistic Regression topped the F1 score and clocked a relatively short evaluation time.
- The feature importance in Logistic Regression surrounds Vader's sentiment scores (pos, neu, neg, compound) as well as certain keywords like location, room, staff.
- For every 0.10 upshift in the decision threshold, we see a drop of 500 false positive but at the expense of an increase of false negatives.

ix. Conclusion of Model and Future Works

While the random probability of predicting a bad review is about 35%, the average precision of the model is about 78% twice the predictive power of that. However, there is still room for improvement for the model.

The area of improvements as follows:
- To perform further multivariate analysis to find the connection between the text classification features and the existing geographic/guest features to uncover more business insights.
- Relook at the text pre-processing to clean the data further.
- Relook at hyper parameter tuning to see if precision/recall/F1 can be improved further.
- Apply the model to a larger sample size to see if the score differs.
- A neural network might need to be considered to cater to the size of the eventual dataset.

7. Implementation

Before implementation, the precision recall trade off needs to be considered before finalising the decision threshold to set for the model.

The model should be applied to a larger training sample to a large database to assess if the results are consistent so that the model is not underfitted before deploying into production, exposed to larger unseen data. One downside of TFIDF is that, the vocabulary bank of TFIDF is currently limited to the 50k sample reviews and might run the risk of having to predict a bad review inaccurately as a good review.

8. Data answer

Through this use case, an extensive EDA on the geographic and demographic profiling of the guests as well as applying text sentiment using Vader on the review text has been performed. A predictive classification model that was able to predict up to 78% precision of the predictions was also built to assess the sentiment behind the review text.

The precision% could be improved further with the further works abovementioned.

9. Business answer

Below are the key summaries to the findings of the EDA and text analysis:

a) Business Insights

The below is the summary of the characteristic of the low scoring review text:

- 70% of the travellers were from Europe but the best reviews came from outside Europe (N America & Oceania)
- Within the travellers in Europe (mostly from UK) , yet those from Netherlands, Italy, Germany, France gave lower score than their counterparts.
- For non Europe travellers, the lower rating travellers came from Middle East (Saudi Arabia, Qatar, UAE) and India.
- There is more 4x as many leisure vs business travellers who had bad reviews.
- There is also a higher proportion of solo who had gave bad reviews compared to the segments.
- The review scores drops a little every 3 days and beyond day10 of the hotel stay.
- The lowest scoring review months (where review counts were highest) were Sep & October, possibly after the summer break vacations.
- The hotel with higher review scores were more consistent through the months as compared to the hotels with lower review scores.

b) Text Classification Analysis

The below is the summary of text & sentiment analysis:

- Most reviews averaged 250 words though top reviews can go beyond 450 words.
- Positive reviews tend to be longer than negative reviews.
- From the unigram analysis, it was observed that unigram removes context from the original review especially when they are used along with words like “not, no” which reverse the original intent of the review.
- From bigram, trigram analysis, these areas were mentioned in the reviews:
    - Room size being small
    - Air conditioning
    - Room amenities like mini bar, Wi-Fi, bed, room service, bed
    - Hotel location, walking distance
    - Helpful, friendly staff.

10. Response to stakeholders

The recommendations to the business problems:

- While the size of the hotel room cannot be easily altered, the solution proposed would be for the hotel can be more transparent in the facts on their website by providing more precise room dimensions between furniture, 3D room layout so that guests can better visualise the room prior to booking.

- Similarly, there were concerns over the hotel locations being far from their destinations and the guests had to walk. Perhaps the hotel can look into shuttle bus service that chauffeur hotel guests to the common destinations or nearest subway stations.
- To address quality of staff service, the client would need to invest in staff training and perhaps look into their staff retention to avoid high staff turnover to maintain the service level standards in the hotel. The hotel might also need to look into flexible staffing strategies during busy months to maintain the service level of the hotel staff.
- The guest also placed much importance on "breakfast" as evidenced in the TFIDF analysis and the model feature importance, hence the hotel has to ensure that they continue to improve the negative aspect in their breakfast offerings.
- Last but not the least, if the client has budget to uplift or renovate, it would be ideal that they look into the air-conditioning system, bed condition.

11. End-to-end solution

For end-to-end solution, the following features would have to be considered and required for the trained model:

i. Vader Sentiment Analysis performed on the review text
ii. Application of the text cleaner before TFIDF is measured against the trained model
iii. With the above, the values required in the feature columns are extracted to be fed into the classifier to make prediction of the overall review score.
iv. If the review score prediction is above 8, it would be considered as a good review.

12. References

The dataset can be found at the following website:
https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe

I used a variety of libraries as follows:

- Basic libraries: numpy, pandas, time, collection,
- Data visualisation libraries: matplotlib, seaborn
- Text processing libraries: nltk, spacy, textblob, wordcloud
- Model training libraries: sklearn, xgboost, lightgbm, imblearn
- Misc libraries: re, folium, random, string, itertools, tqdm

Last but not the least, the following articles and notebooks has provided great guidance to this project:

1. Detecting bad customer reviews with NLP (Jonathan Oheix, 2018)
   https://towardsdatascience.com/detecting-bad-customer-reviews-with-nlp-d8b36134dc7e
   https://www.kaggle.com/code/jonathanoheix/sentiment-analysis-with-hotel-reviews

2. Sentiment Analysis using NLP on Hotel Review Dataset (Rishan Sanjay, 2021)
   https://medium.com/analytics-vidhya/sentiment-analysis-using-nlp-on-hotel-review-dataset-fa049e23de29

3. Text analysis basics in Python (Sophia Yang, 2020)
   https://towardsdatascience.com/text-analysis-basics-in-python-443282942ec5

4. How to solve 90% of NLP problems: a step-by-step guide (Emmanuel Ameisen, 2018)
   https://medium.com/insight-data/how-to-solve-90-of-nlp-problems-a-step-by-step-guide-fda605278e4e

5. SENTIMENTAL ANALYSIS USING VADER (Aditya Beri, 2020)
   https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

6. Máster – Big Data & Data Science UCM - Anál. Sent. (JAIME SÁNCHEZ, 2021)
   https://www.kaggle.com/code/jaimesz11/m-ster-big-data-data-science-ucm-an-l-sent