Amplify Analytix Recruitment Task - Report

Question 1:

The data was loaded into RStudio to understand that there were 95842 unique searches across 34 point of sale websites in a total of 192 countries, although 74.3 % of searches performed by users belonging to only three countries. There was a total of 117378 hotels in 166 countries, of which 71 % were from three countries.

Firstly, before any feature summary statistics could be done, it was important to handle missing values. *listing_position* attribute was dropped since it is not present in testing data as understood from the data description file. A quick look at the Missing values in the raw data denote that all competitor related attributes have an incredible amount of missing values and these were dropped.

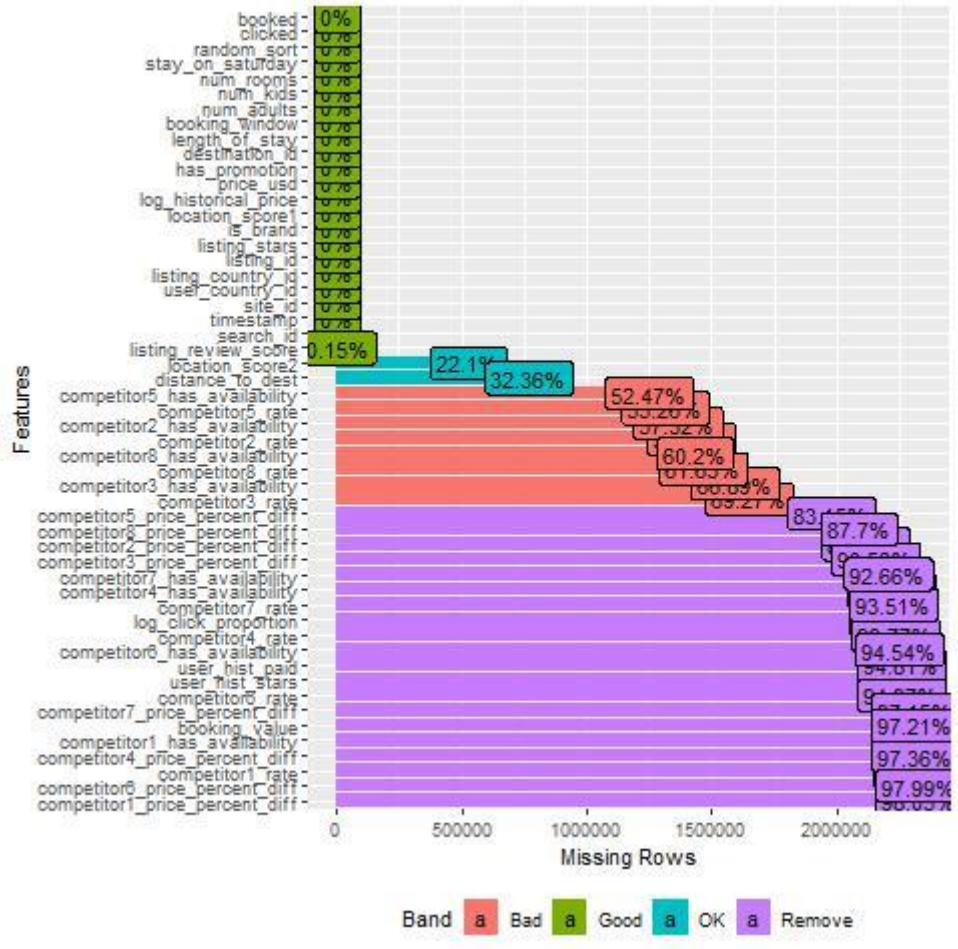

Figure 1 Missing values

Apart from these competitor attributes, all the other attributes with missing values were deleted. Detailed analysis on the choices made is available in the code. 3.4 % of the *listing_stars* attribute and 4.2 % of *listing_review_score* had a value of zero. These observations were dropped.

It was interesting to notice that data from months July to October is not available. (All plots can be found in the output folder in the code base.) From plots, it is understood that there was not much of

an effect of day of the search (Sunday- Saturday) on either clicks or bookings and which is validated by the model feature importance.

Secondly, on all the features suspected to be numeric; normality tests were performed on each to study the difference between mean and median, kurtosis and skewness, with their respective standand errors. Two features *num_rooms* and *length_of_stay* had a Poisson distribution.

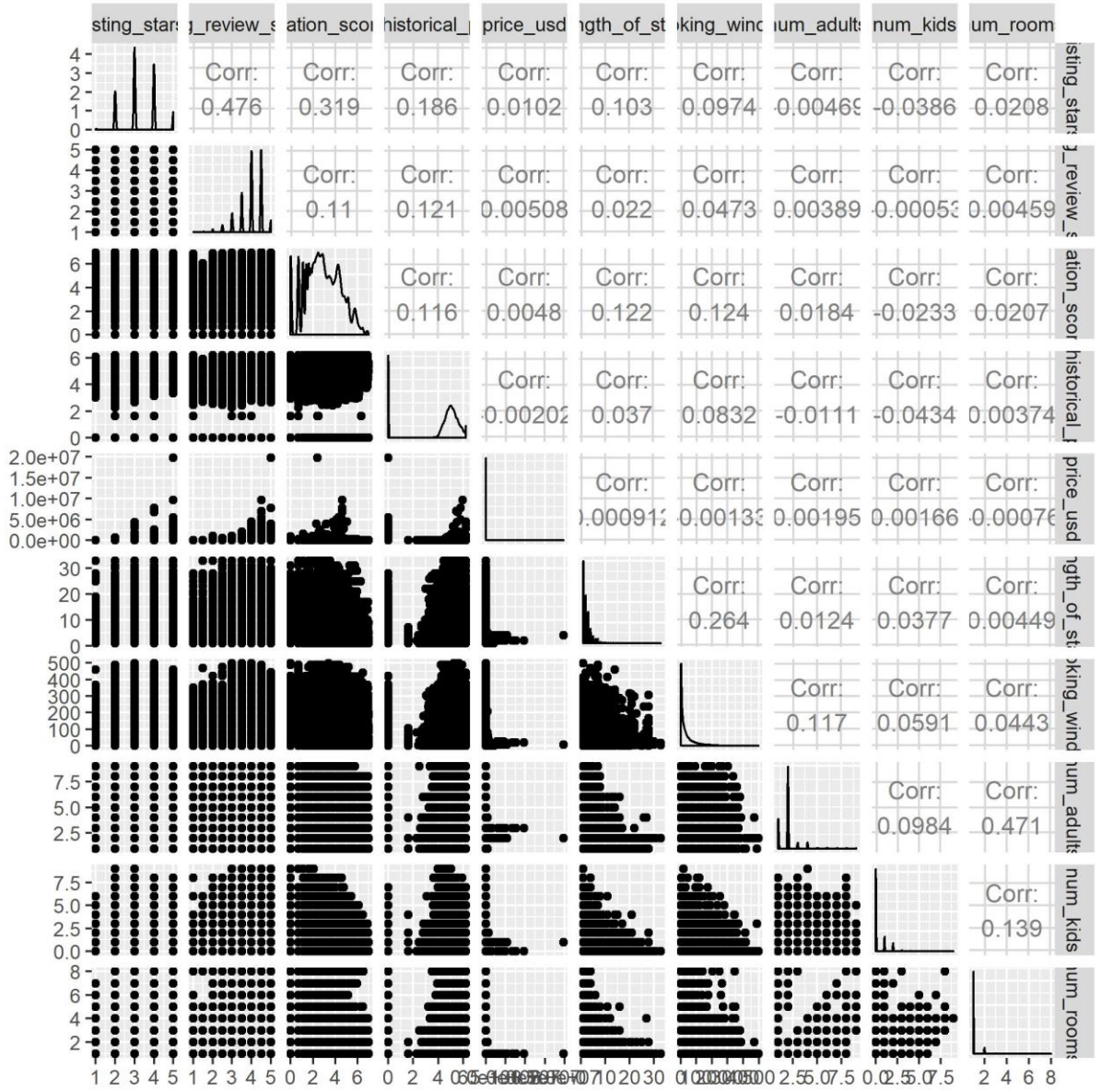

Figure 2 Summary statistics

Both *listing_review_score* and *location_score1* had noticeable correlation with *listing_stars* of 0.47 and 0.32 respectively. These two variables were averaged to create a composite feature. When an independent variable can be linearly predicted by one or more predictor variables multicollinearity happens which skews the result. Also, if independent variables are correlated with each other, the quality of the prediction comes down.

The data has high class imbalance. Minority class in *clicked* is 4.46 % and only 2.79 % in *booked.* If class imbalance is not noticed higher Accuracy maybe observed but other metrics such as Recall suffer. According to the use case and the approach, data belonging to minority class could be repeated and

resampled(over-sampling) in the pre-processing step so the machine learning algorithm understands the underlying class distribution and the variance in the dataset. Data belonging to majority class could also be under-sampled. (however, this may cause information loss) It is to be noted that with imbalanced datasets, a good cross-validation technique goes a long way.

To tackle class imbalance, one modelling approach would be to use Cost Sensitive Learning. In Cost Sensitive Learning, cost for False positives and False Negatives are defined ahead of times and total cost is calculated such that cost for False positives is multiplied with total number of False positives and then added to the cost for False Negatives multiplied with the total number of False Negatives. This total cost is calculated for multiple models and the model with least total cost is selected.

Question 2:

Click through rates and conversion rates are important metrics used by sales and marketing teams to evaluate the effectiveness of their channels and to design campaigns to improve upon either of these.

Click through rate in this scenario is calculated by first aggregating data at a listing_id level and then dividing total clicks "clicked" with total views and multiplied by 100.

Conversion rate is calculated by dividing total bookings (from the clicks "clicked") with total clicks "clicked" (not viewed) and multiplied by 100.

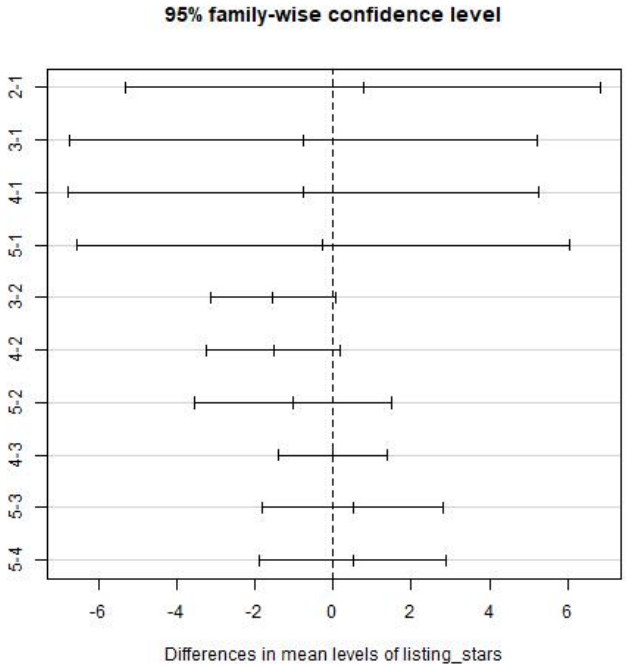

Figure 3 conversion rate significance levels - TukeyHSD

For the purpose of this analysis, *listing_stars* was considered as a categorical variable and *listing_review_score* as a continuous variable. One-way ANOVA of *listing_stars* and click through rate

showed no statistically significant results. On the other hand, analysis of *listing_stars* with conversion rate revealed that there was a 1.5 % more conversion rate for 2-star hotels over 3-star and 4-star hotels at a 90 % confidence interval (p-values of 0.06 and 0.09 – very close to 0.05 to be 95 % confident but not there yet) as seen from the above figure.

Linear relationship of *listing_review_score* with click through rate or conversion rate was not statistically significant when Kendall Tau's correlation test was applied. However, looking at the quantile-quantile plots, it is clear that there are heavy tails at the top and bottom of conversion rate although relationship is linear at the center. Since a decision tree was used to model bookings, it should be able to pick up this relationship.

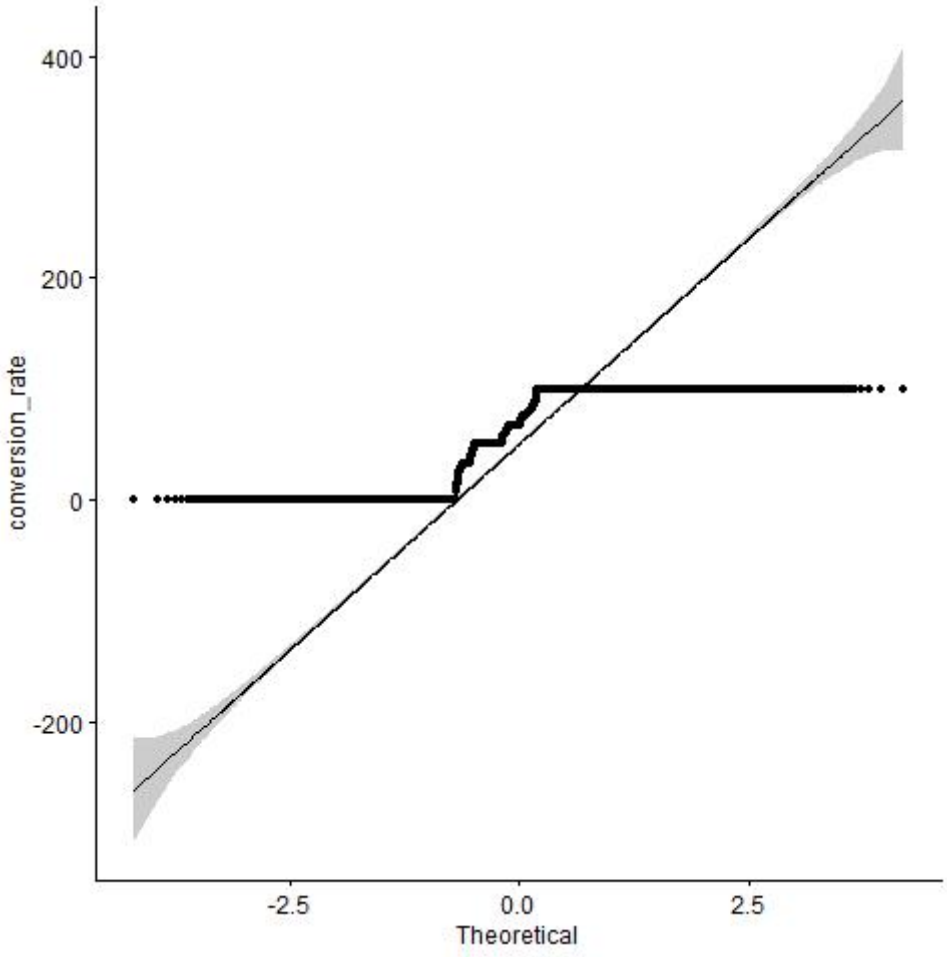

Figure 4 Quantile plot of conversion rate

Question 3:

There were three possible choice for the target variable. Since *booking_value* had an incredible amount of missing values, it was dropped initially. Although *clicked* had 4.58 % of class 1 and *booked* only 2.89 % *booked* is chosen as the target variable since for a business it is usually first priority to increase their conversion rate rather than click through rate. *clicked* was dropped from data.

As discussed in question one, all the continuous variables that established a normal distribution and all the other categorical variables where used in the analysis. Since a Gradient boosting machine was used for modelling, if a feature is not important it is dropped automatically.

Certain preprocessing was done. *price_usd* had a vast standard deviation so a new feature called *price_quantile* was created with quantile rankings. This turned out to be top 3 feature. Since *listing_review_score* and location_score1 had a correlation around 0.46 the average of this was created as a new feature which was second most important feature after *listing_stars*. Few other features where created such as top 6 user countries, top 6 destination countries, top 6 point of sale websites, weekday of booking but all of them were not significant.

A gradient boosting method was used to build a model. Decision trees are good in this scenario because it is easier to interpret why the model predicts in the way it does and it helps to discard rules that doesn't make sense at a later point. Since there was a high class-imbalance, multiple gbm models where built with various techniques to handle such as over sampling, under sampling and cost sensitive learning. Assigning weights to classes solved the problem and gave higher AUC of 0.58

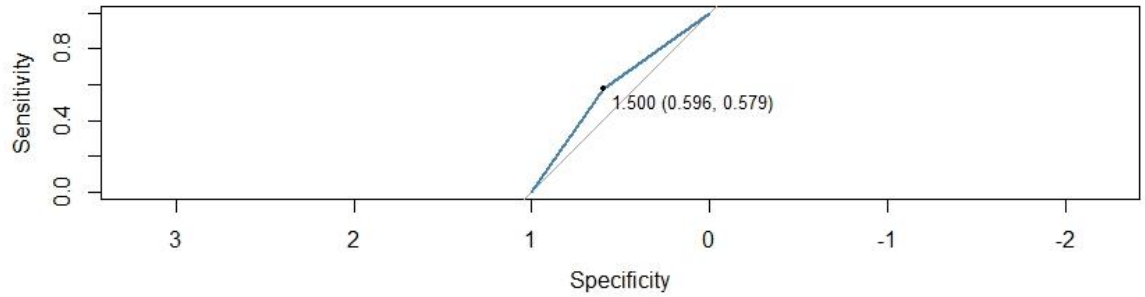

Figure 5 Best model AUC

AUC was chosen as evaluation metric. Due to class imbalance Accuracy can be very high but the model would not be practical. Using AUC balances sensitivity and specificity there by learning the under sampled data efficiently. Final AUC was 0.58 but at least it learned over 50 % of minority class. A model where if Accuracy was chosen as evaluation metric, a value of over 0. 90 would still be bad.

Since GBM was chosen as the model, the significance levels of each of the features can be understood and this gives a good intuition on what another feature engineering could be done. It gives a good start and is a good balance between explicability and model complexity.