Hotel Booking Demand Dataset

i. Supply Chain Management Scenario

The dataset provides information regarding the arrival dates, cancellation status and several other hotel and booking characteristics which would enable us to forecast demand in terms of booking cancellations. This kind of forecasting aids the hospitality industry in managing the required number of supplies with expiry dates (e.g. shower gel, soaps and handwashes). Instead of ordering all the supplies in bulk at once, the hotel management could optimize their supplies according to the prediction of booking cancellations. This would ensure that the products are used before they expire and the waste is reduced. The dataset also discusses about the type of meal preferred by the guests. Similarly, using visualization, the type of meals preferred can be observed and the raw materials could be purchased accordingly. Therefore, using demand forecasting on this data set would not only help in saving costs but also in optimizing the revenue.

ii. Data Description

The dataset contains booking information for a city hotel and a resort hotel and comprises of 32 columns that include information such as when the booking was made, length of stay, the number of adults, children, and/or babies, preferred meals and the number of available parking spaces, among many other things. Data types used in this dataset are boolean, integer, strings and different types of categories. The detailed description about the columns are attached in the Appendix 2.

iii. Data Pre-processing

Data exploration and pre-processing is usually the first step done before performing data analysis. The first step in data processing involves dropping the null values since they hamper the dataset. In this dataset, features such as agent; company; children; and country had null values. To deal with such features for example '*Country*'. I created a function to reduce the amount of dummy columns by keeping the top 10 Country values and set the rest to 'other' .Finally, only 4 features were dropped due to two important reasons: agent and company were dropped since they had many null values and '*reservation_status'* and '*reservation_status_dat'e* columns were dropped as they are directly related to our target variable.

The next step involved adding new features. Since the hotel is just concerned with total number of guests arriving, rather than segmenting guests according to age and other attributes them, a new column was made indicating only the total guests arriving for any particular booking. Along with this a new feature '*adrpp'* was introduced which was calculated from arrival daily rate ('adr') divided by total members that arrived in the hotel. Another new column *'weekend_or_weekday'* was created from given dataset which differentiate between if the customer only stayed just for weekends or weekday or both.

With the cleaned dataset we visualized different columns in the pre-processed dataset to understand the dataset better.

Visual graphs are in Appendix 1.

The last step was dealing with categorical and numerical features i.e. finding out how the features are correlated to each other. For this, Correlation Matrix Spearman Method for categorical data and Correlation Matrix Pearson Method for Numerical data is used.

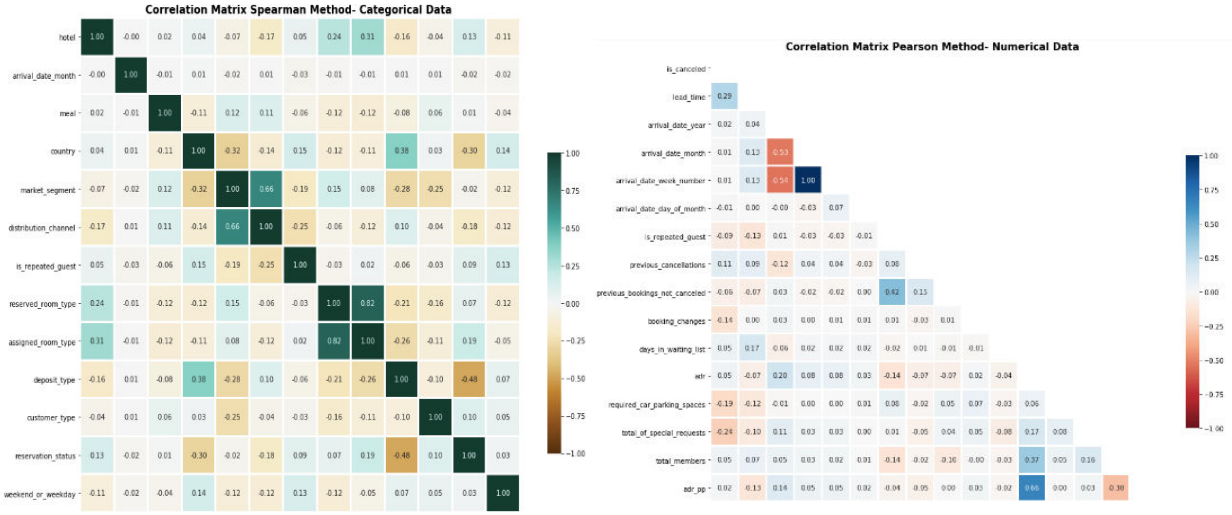

Figure 2 Correlation matrices

The above matrix shows the relationship between 2 variables. To understand better dark blue represents positive strong correlation and dark red represents negative strong correlation between the variables

iv. Data Analytics

After the data has been transformed into the correct format and also getting the better understanding of the dataset ,classification algorithm was performed using Random Forest.

i. Random Forest

The Random Forest technique uses is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object (Wiener', 2002). Classification trees have a built-in method for feature selection, which chooses the feature with the most impact at each split. Thus, adding more features should generally increase predictive accuracy. The training set includes all 70 features for the model to randomly select the subset and build the ensembled trees. The random forest model built for this project includes following parameters: n estimators=200 and max depth=50. The figure below shows the evaluation metrics for this model.

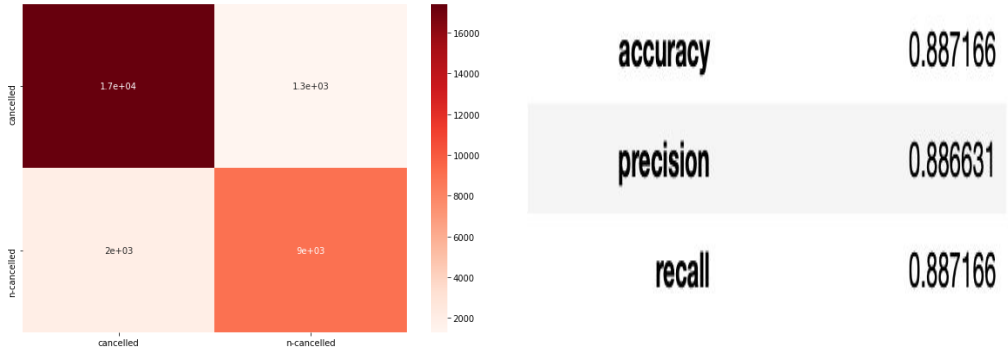

Figure 3 Confusion metrics (left), Random forest model evaluation (right)

These metrics show that the model is able to accurately classify 88% of booking present in the test dataset. To further improve our classification performance, the number of features at which our models give the best accuracy was explored. The below figure also shows the number of features with respect to the model accuracy. From the plot, it can be seen that the model performs the best with around 28 features.

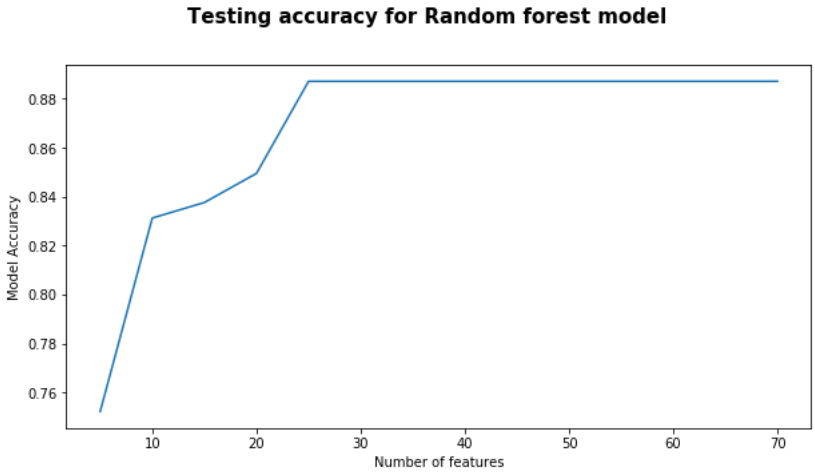

Figure 4 Testing accuracy for Random Forest model

ii. Gradient Boosting Classifier

Similar to Random Forests, Gradient Boosting is an ensemble learner. Gradient boosting is a technique for performing supervised machine learning tasks, like classification and regression. The idea behind using the gradient boosting after random forest is that in gradient boosting trees are built sequentially i.e. each tree is grown using information from previously grown trees unlike in bagging where we create multiple copies of original training data and fit separate decision tree on each. *('Xu', Gradient boosted feature selectio, 2019) (kharshit G. B.-a., 2018).) (kharshit, Gradient Boosting vs Random Forest, 2018)*

The parameters chosen were learning rate, max depth, min samples leaf, max features and the n estimators. The max depth and n estimators were also the same parameters chosen in a random forest. Here, the min samples leaf and max features. Figure 4 shows the evaluation metrics for this model with 28 best features.

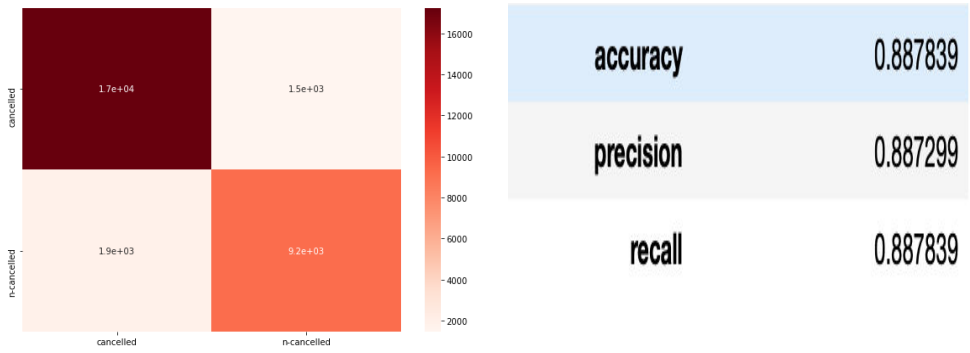

Figure 5 Confusion metrics (left), Gradient Boosting model evaluation

The new accuracy score of 0.8878 is almost identical to the one obtained with the Random Forest.

v. Evaluation and Comparison

The objective is to predict if guests actually end up coming (i.e. not cancel their booking). Therefore, for this we need a classifier which has the low false positive rates, i.e. bookings that were cancelled but predicted as not cancelled.

When we look at the confusion matrix for the random forest classifier, Figure 3 we see that for the 'cancelled' class, out-of 18,687 booking in the test data-set falling in this class, the model correctly classifies 17,381 bookings but miss-classifies 1306 bookings as not-cancelled.This. shows that this model does a better job in identifying the bookings that are 'not cancelled'.

But if we look at the confusion matrix for the Gradient Boosting , Figure 5 we see that for the 'cancelled' class, out-of 18,687 booking in the test data-set falling in this class, the model correctly classifies 17,220 bookings but, miss-classifies 1467 bookings as not-cancelled.This shows that this model does not work well in identifying the bookings that are not cancelled.

If we are just concerned about the accuracy then gradient boosting works a bit better than Random Forest .But since our goal is to minimize the cost for the utilities bought by the hotels for the guest, Random forest is better.

vi. Osterwalder's Business Model Canvas

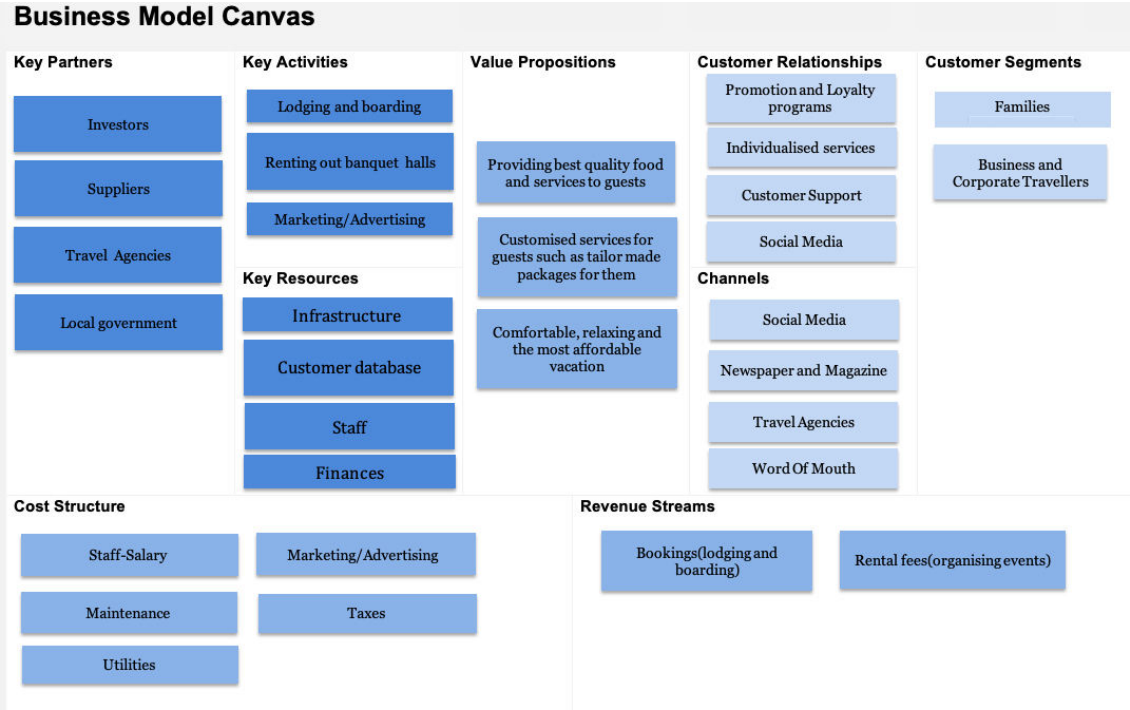

Figure 6 Business model canvas 1

Strengths: Some of the strengths of this business model for hotels are affordable, culinary experience, various offerings, the vast range of services such as hotel pickup and drop,

organizing game nights, movie nights and others. Mainly, it serves to provide great customer satisfaction.

Weaknesses: Some of the weaknesses for this hotel business model include obtaining the required funds, the need for hotels to continuously adapt to changing demands and trends, and the complexity of managing the work force, high infrastructure maintenance and high taxes.