Hotel Booking Cancellations Prediction Report

BA with r (buan 6356)

**GROUP NUMBER: 10**

* **LIKITH VINAYAKA GIRIDHAR (LXG210034)**
* **PRAFUL PATIL (PVP220001)**
* **SAI ROHIT BOGGARAPPU (SXB220120)**
* **RAHUL KINTALI (RXK210144)**

**Business context:**

1. Booking Dates: The dataset likely contains information about the dates on which bookings were made at the city hotel and the resort hotel. This could include details such as the date of booking, the time of booking, and possibly even the method of booking (e.g., online, phone, in-person). Analyzing booking dates can provide insights into booking patterns, trends, and seasonality, which can help hotels in their revenue management strategies, marketing campaigns, and staffing decisions.

2. Length of Stay: The dataset may include information about the duration of the stays for the bookings made at the hotels. This can include details such as the number of nights booked or the check-in and check-out dates. Analyzing the length of stay can provide insights into guest preferences, booking behavior, and demand patterns. It can also help hotels optimize their pricing strategies, room inventory management, and resource allocation.

3. Party Size: The dataset may contain information about the size of the parties for which the bookings were made. This can include details such as the number of guests per booking, the number of rooms booked, and the types of rooms booked (e.g., single, double, suite). Analyzing party size can provide insights into guest demographics, preferences, and booking behavior. It can also help hotels optimize their room allocation, amenities, and services based on guest group sizes.

4. Parking Availability: The dataset may include information about parking availability at the hotels, such as whether parking was available, the type of parking (e.g., free, paid), and the location of parking (e.g., on-site, off-site). Analyzing parking availability can provide insights into guest preferences, parking demand, and revenue generation from parking facilities. It can also help hotels optimize their parking policies, pricing, and marketing strategies related to parking services.

**Content Exploration of our dataset:**

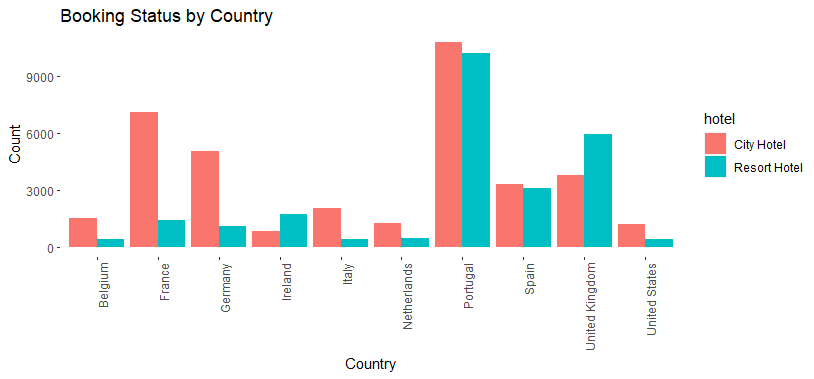
The dataset at hand contains a target variable labelled "is\_canceled" which represents binary classes 0 and 1, where 0 denotes "Not Cancelled" and 1 denotes "Cancelled". The remaining variables in the dataset provide additional details about the bookings, including lead time, arrival date, length of stay, number of guests, type of meal, country of origin, market segment, distribution channel, room type, deposit type, and more. These variables may have numerical, categorical, or binary values.

The dataset is presented in a tabular format, with rows representing individual bookings and columns representing attributes of the bookings. This layout allows for easy organization and analysis of the data. The dataset has the potential to be valuable for analysing patterns in hotel bookings, predicting future bookings, and identifying factors that contribute to cancellations. By examining the various attributes of the bookings, insights can be gained into guest behaviour, booking trends, and potential strategies for optimizing hotel operations and revenue management.

**Following are the questions we decided to ask of the dataset:**

1. **Where do the guests come from?**

**Finding.** As part of our analysis, we investigated the number of tourists arriving from different countries to determine if any country had 1500 or more tourists. Our findings revealed that Portugal had the highest number of tourists among all the countries examined due to this It is also possible that the hotel might be in Portugal.



1. **How much do guests pay for a room per night?**

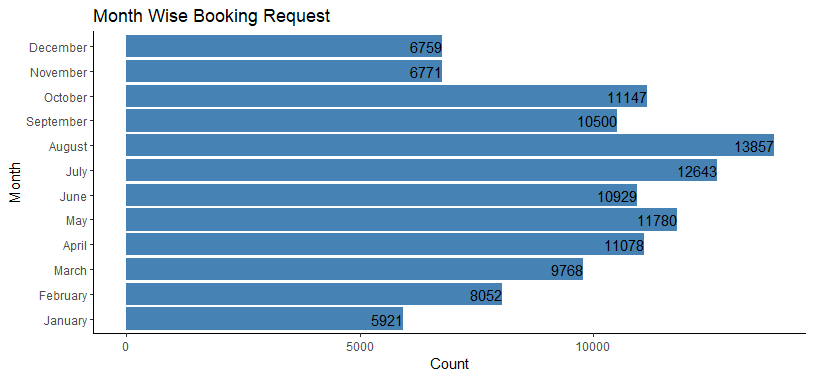
**Finding.** When compared to Resort hotel, people at city hotel tend to pay higher mainly because City hotels are often located in prime areas of urban centers and are commonly chosen by business travelers who require proximity to offices, conference centers, and other business-related facilities. The highest number of guests are from transient customer type bookings and lowest being for group customer type bookings. This trend can be seen for both the hotel types. Prices charged by both the hotel types for different customer types remains the same. The spread of outliers is extremely large for Transient and Transient-Party

Chart, histogram

Description automatically generated

1. **Which is the Busiest Month?**

**Findings.** For the City hotel, we can see that the bookings consistently remain around 7500 or more from April to October, and the bookings drop during the winter months. This trend is very similar in the Resort hotel, but the bookings hover around half of the bookings seen in the City hotel during the same period.

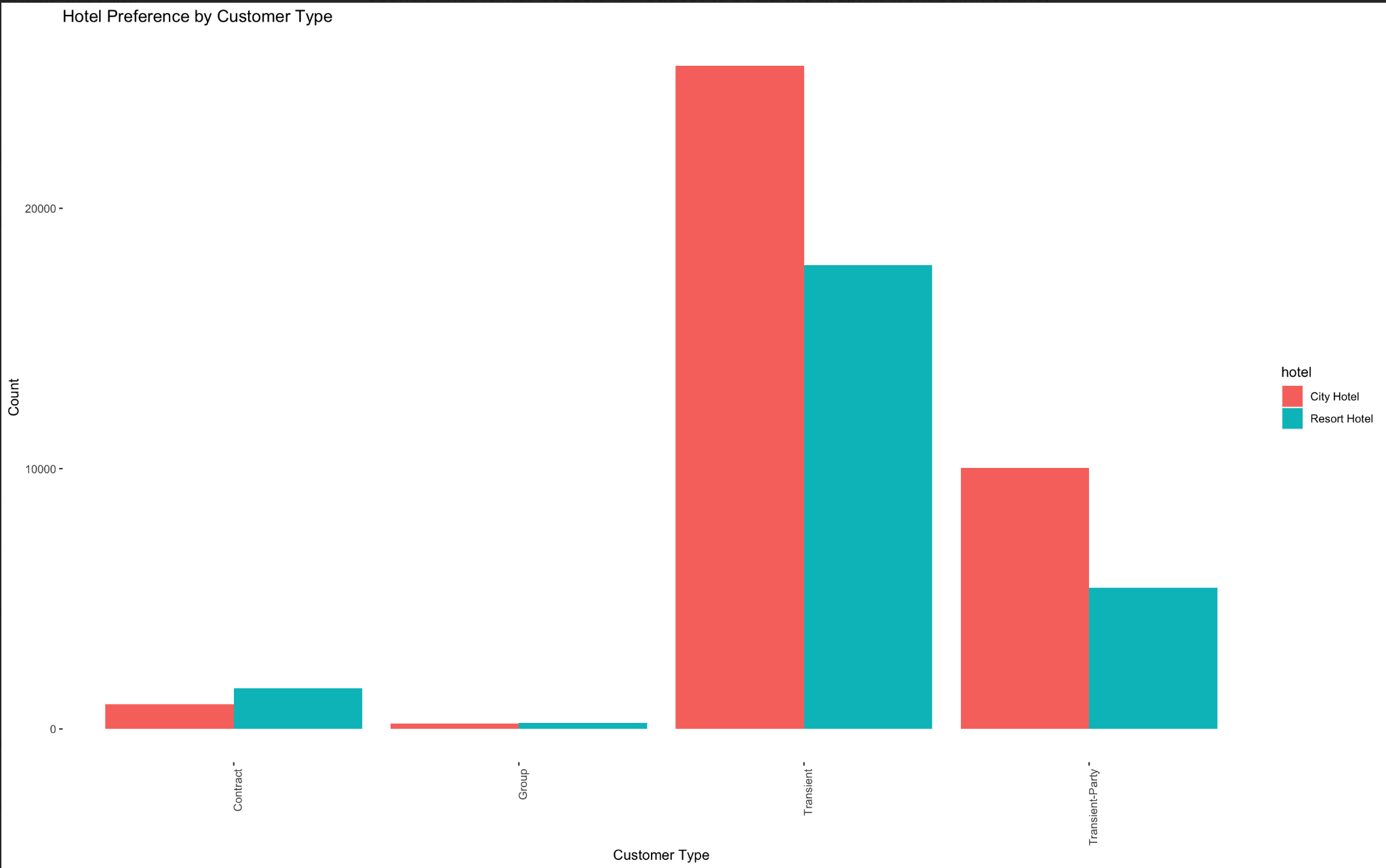


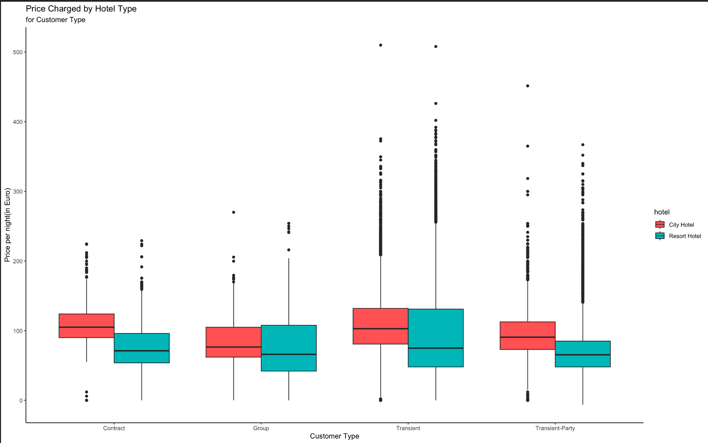
Chart, bar chart

Description automatically generated

1. **How does the price per night vary over the year?**

**Finding.** We can notice that this is a peak and valley plot. The price per night is maximum in August and Minimum in January and November for resort hotel. For Resort hotel the price hikes from May to August and drops till November. This is also because of the points mentioned in finding 3.





1. **How long do people stay at the hotels?**

**Finding.** Based on the analysis of the dataset, it is evident that most guests tend to stay for approximately 3 nights on average at the hotel. Furthermore, the highest number of guests stay for a duration ranging from 0 to 5 days, indicating that the most common length of stay falls within this timeframe.

Chart, histogram

Description automatically generated

1. **How many bookings were canceled?**

**Finding.** As per the analysis of the dataset, it was observed that a significant portion of the total bookings, accounting for 37%, were canceled. Out of the total cancellations, 28% were from the resort hotel bookings, while a higher percentage of 42% were from the city hotel bookings. This indicates that the city hotel experienced a higher proportion of cancellations compared to the resort hotel. This information provides valuable insights into the cancellation rates for each hotel type, which can be utilized for further analysis, revenue forecasting.

1. **Which month have the highest number of cancelations?**

**Finding.** It is apparent that the cancellations for the months of April, May, June, September, October, and December are relatively similar, with April having the highest number of cancellations, albeit by a slightly larger margin.

Chart, bar chart

Description automatically generated

**Challenges faced:**

1. **Imbalanced dataset:** our dataset exhibits a slight class imbalance, which could potentially impact the accuracy of our binary class statistics scores. This observation will be further analyzed during our analysis process.

> prop.table(table(df$is\_canceled))

|  |  |
| --- | --- |
| 0 | 1 |
| 0.6292521 | 0.3707479 |

1. **Missing Values:** The "Country" column contains most missing values. As a result, we have made the decision to exclude this column from our analysis to ensure data integrity and accuracy.
2. **Handling Categorical Variables:** Most of the categorical variables have over 5 categories, so for us to perform one hot encoding on this data to perform PCA the dimensionality of the data set would increase to a very high level and thus demanding extremely high computational power and computational time. Hence, we were not able to perform one hot encoding and in turn PCA for the categorical data.
3. **Feature Selection:** High Correlation with Target Variable: The "reservation\_status" column, which is a categorical variable, exhibits an almost 100% correlation with our target variable. Consequently, we have had to remove this column from our analysis to prevent any potential bias in our results. It's worth noting that these anomalies are not reflected in the correlation heat map, as it only considers numerical variables.

**Tasks Performed:**

1. **Explored Interesting Libraries:**

* **Dplyr**: Used for creating pipelines to build subsets of data frames and data frames summary.
* **Tidyr**: Used in conjunction with dplyr to clean data.
* **Plotly**: Used to create interactive plots (annotations)
* **FactoMineR**: Exploratory data analysis methods to summarize, visualize and describe data.
* **Countrycode**: Used for converting country names and codes from one format to another.
* **Lubridate**: To manipulate date.
* **Mice**: Used for imputing data.

1. **Data cleaning and manipulation for Visualization:**

* Convert character columns to factors
* Eliminate NA and other undefined values – Omit missing Records
* Some rows contain entries with zero adults, children, and babies, so we deleted these records as they have Zero guests.
* Meal column having “Undefined” values have been replaced by Mode value.
* Records with “Undefined” values in “market\_segment” and “distribution\_channel” have been deleted.

1. **Data Visualization:**

* In addition to the above data cleaning and manipulation we have used various libraries such as dplyr, ggplot 2, tidy, plotly, FacotoMineR, countrycode, lubridate to extract the required subsets from the dataframe and plot various charts such as Bar charts, Histograms, Boxplots, Line chart, horizontal bar chart, stacked bar chart, heat maps etc to draw various insights from the data.

1. **Data Pre-Processing and Cleaning for building the model:**

* Plotting Heat Map to check for correlation in numeric data: We observe that none of the numeric variables are highly correlated to the target variable “is\_canceled”
* Based on Intuition, we remove some of the irrelevant columns such as “agent”, “company”, “reservation\_status\_date”
* We also remove the categorical variable “reservation\_status” because of 100% correlation with our target variable “is\_canceled”
* We replace missing values with the mode for categorical variables, we impute children=”0”, country=”Unknown”, meal=”SC”
* We replace missing values of numeric variables with their median, for variables “lead\_time” and “required\_car\_parking\_spaces”

1. **Feature Engineering:**

* We combine arrival\_date\_year, arrival\_date\_month, arrival\_date\_day\_of\_month to form a new variable called arrival\_date
* Create a new variable called total\_guests by adding the number of adults, children, and babies.
* Then we convert the categorical variables to factors

1. **Principal Component Analysis:**

* We perform Principal Component Analysis on numeric variables.
* We merge the PCA values with the categorical variables.
* We use the first 3 components which capture majority of the data

1. **Building the Model:**

* We split the dataset into training and testing set with the split factor of 0.7 for training.
* In Logistic Regression Model we observe the following performance metrics when we test on the test data:

**Confusion Matrix:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual | | |
| Prediction |  | 0 | 1 |
| 0 | 21287 | 6315 |
| 1 | 1216 | 6943 |

**Statistics:**

|  |
| --- |
| Accuracy : 0.7894 |
| No Information Rate : 0.6293 |
| Kappa : 0.5099 |
| Sensitivity : 0.9460 |
| Specificity : 0.5237 |
| Pos Pred Value : 0.7712 |
| Neg Pred Value : 0.8510 |
| Prevalence : 0.6293 |
| Detection Rate : 0.5953 |
| Detection Prevalence : 0.7718 |
| Balanced Accuracy : 0.7348 |
| F1 Score : 0.6483634 |
| AUC : 0.7348 |

In general, these metrics suggest that the model has good sensitivity but poor specificity, meaning that it is better at correctly identifying positive cases than negative cases.

* We also build a **Random Forest Model** of 500 trees we observe the following performance metrics when we test on the test data:

Type of random forest: classification

Number of trees: 500

No. of variables tried at each split: 4

OOB estimate of  error rate: 16.52%

**Confusion matrix:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual | | |
| Prediction |  | 0 | 1 |
| 0 | 21012 | 4308 |
| 1 | 1491 | 8950 |

**Statistics:**

|  |
| --- |
| Accuracy : 0.8378401 |
| Kappa : 0.636591 |
| Sensitivity : 0.6750641 |
| Specificity : 0.9337422 |
| Pos Pred Value : 0.8571976 |
| Neg Pred Value : 0.8298578 |
| Prevalence : 0.3707391 |
| Detection Rate : 0.2502726 |
| Detection Prevalence : 0.2919661 |
| Balanced Accuracy : 0.8044031 |
| F1 Score : 0.7553061 |
| AUC : 0.8046913 |

In general, the metrics suggest that the model has good sensitivity and specificity, meaning that it is able to correctly identify both positive and negative cases. The positive predictive value is high at 0.9356, indicating that when the model identifies a case as positive, it is correct most of the time. The negative predictive value is also relatively high at 0.8944, indicating that when the model identifies a case as negative, it is correct most of the time. The prevalence of the positive cases in the dataset is relatively low at 0.3708. The detection rate is 0.2989, indicating that the model correctly identified almost one-third of all cases. The detection prevalence is 0.3195, meaning that the model identified 31.95% of the cases as positive. Finally, the balanced accuracy of the model is 0.8867, which is a relatively high value representing the overall accuracy of the model.

* **Comparing the two Models:**

Based on the obtained metrics, the Random Forest model appears to perform better than the Logistic Regression model for the hotel booking cancellation prediction problem.

1. **Accuracy:** The Rando m Forest model has an accuracy of 0.8378401, which is higher than the accuracy of 0.7894 for the Logistic Regression model. This suggests that the Random Forest model is able to correctly predict the outcomes more accurately overall.
2. **Kappa:** The Kappa score for the Random Forest model is 0.636591, which is higher than the Kappa score of 0.5099 for the Logistic Regression model. A higher Kappa score indicates better agreement between predicted and actual outcomes.
3. **Sensitivity:** The Random Forest model has a sensitivity (true positive rate) of 0.6750641, which is lower than the sensitivity of 0.9460 for the Logistic Regression model. Sensitivity measures the ability of a model to correctly identify positive cases. In this case, the Logistic Regression model has higher sensitivity, suggesting it is better at identifying actual positive cases.
4. **Specificity:** The Random Forest model has a specificity (true negative rate) of 0.9337422, which is significantly higher than the specificity of 0.5237 for the Logistic Regression model. Specificity measures the ability of a model to correctly identify negative cases. The higher specificity of the Random Forest model suggests it is better at identifying actual negative cases.
5. **Positive Predictive Value:** The Random Forest model has a higher positive predictive value (precision) of 0.8571976 compared to the positive predictive value of 0.7712 for the Logistic Regression model. This indicates that the Random Forest model is better at correctly predicting positive cases.
6. **Negative Predictive Value:** The Negative Predictive Value for the Random Forest model is 0.8298578, which is higher than the value of 0.8510 for the Logistic Regression model. Negative Predictive Value measures the ability of a model to correctly predict negative cases.
7. **F1 Score:** The F1 score for the Random Forest model is 0.7553061, which is higher than the F1 score of 0.6483634 for the Logistic Regression model. The F1 score is a measure of the trade-off between precision and recall, and a higher F1 score indicates a better balance between the two.
8. **AUC:** The Random Forest model has a higher AUC (Area Under the Curve) of 0.8046913 compared to the AUC of 0.734823287033679 for the Logistic Regression model. AUC is a measure of the model's ability to correctly classify cases, with a higher AUC indicating better performance.

**Results and Conclusions**

Based on the above metrics, the Random Forest model appears to be the better choice for the hotel booking cancellation prediction problem, as it has higher accuracy, Kappa score, specificity, positive predictive value, negative predictive value, F1 score, and AUC compared to the Logistic Regression model. However, it's important to consider other factors such as model interpretability, computational resources, and specific business requirements before making a final decision.

**Future Work**

* Performing oversampling or under sampling on the target variable and re-running the models to analyse performance.
* Check the dependence of target variable on each level of each of the categorical variables. Then select the necessary categories and one hot encode. Finally perform PCA on the one hot encoded categorical sub data frame