Glossary

Attrited Customer The negative class in the classification and represents the customers who are no longer registered for a credit card.

Avg_Open_To_Buy One of the features in classifying credit card customers and represents the average open-to-buy which is the remaining amount of credit that a customer can utilise.

Binning The process of partitioning non-categorical data into equally sized bins

Boosted Trees An ensemble of individual decision trees which are weak learners but are trained to learn on the errors of its predecessor.

Bootstrapping The process of resampling which randomly samples, from the original sample, with replacement.

Credit_Limit One of the features in classifying credit card customers and represents the credit limit which is the maximum amount of credit that is extended to a customer.

Decision Tree A model that is similar to a flow-chart with a set of arguments which ultimately leads to leaf nodes that represent the labeled class

Existing Customer The positive class in the classification and represents the customers who are still registered and using their credit card.

Fall-out The number of incorrectly labeled negative classes divided by the number of actual negative classes. Larger the fallout, higher the misclassification of negative classes.

Hyperparameters Parameters which dictate the learning process and final model parameters.

Laplace Smoothing A smoothing technique to ensure posterior probabilities are never 0 despite whether the category was present in the training set or not.

Precision The number of correctly labeled positive classes divided by the sum of the number of correctly and incorrectly labeled positive classes. The lower the precision, the higher the misclassification of negative classes.

Principal Component Analysis (PCA) A technique used for dimensionality reduction on the feature set. Linear combinations of the original features form a set of components which are ranked in order of the variance each component explains.

Recall The number of correctly labeled positive classes divided by the number of actual positive classes. Lower the recall, higher the misclassification of positive classes.

Receiver Operating Characteristics (ROC) A curve which plots the relationship and tradeoff between the TPR and FPR. A straight diagonal line connecting (0, 0) and (1, 1) on the ROC plot is the equivalent of random guessing while plots which are singularly kinked at (0,1) represent an ideal discriminatory model.

Sensitivity Equivalent to recall.

Specificity The number of correctly labeled negative classes divided by the number of actual negative classes. Lower the specificity, higher the misclassification of negative classes.

Supervised Learning A method of training algorithms with labeled data.

Synthetic Minority Oversampling Technique (SMOTE) A process used to oversample the minority class to solve any imbalances within the data. When oversampling, new samples are synthetically generated from existing samples of the minority class.

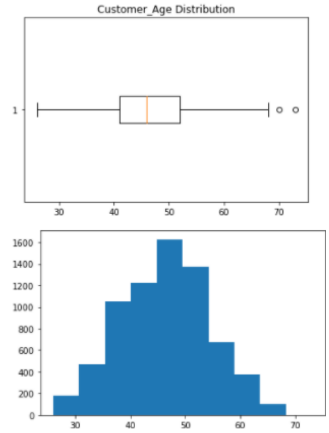
Total_Revolving_Bal One of the features in classifying credit card customers and represents the total revolving balance which is the of credit card spending that goes unpaid.

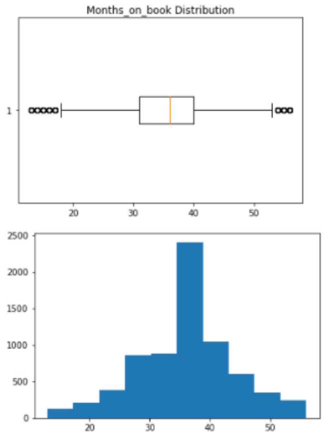
Variance Inflation Factor (VIF) A measure of multicollinearity which can be computed for each feature. The larger the VIF, the more collinear the feature is with all other features.

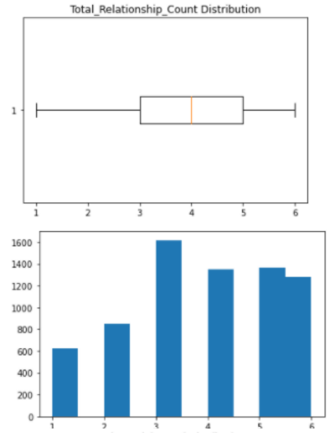
Intermediate Results

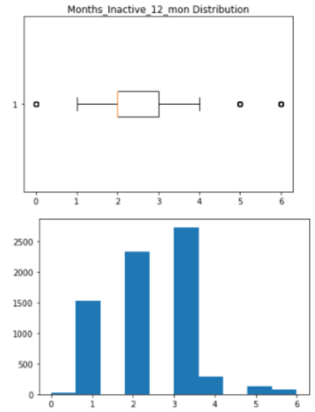
Distribution of Features

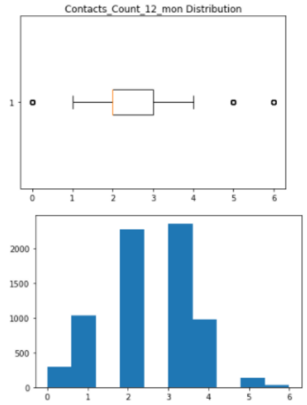
To assess for outliers, boxplots and histograms are plotted for non-categorical features and pie charts are plotted for categorical features. For features with points beyond the maximum and minimum of the boxplot, distribution was either mostly skewed or multimodal and thus no outliers were removed.

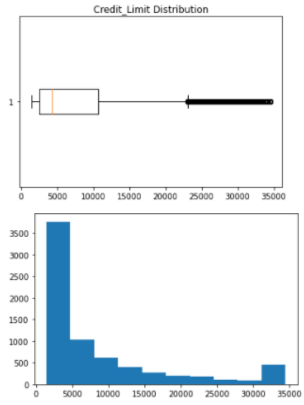


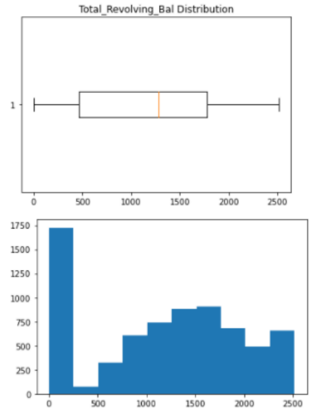


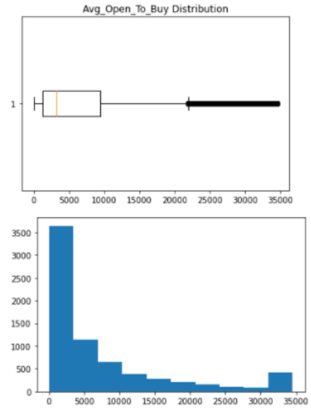


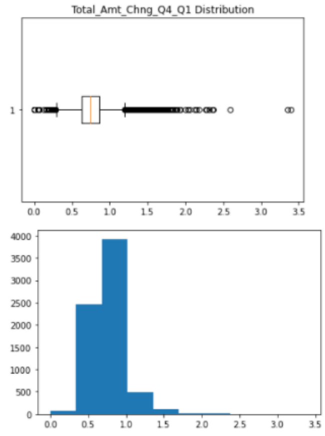


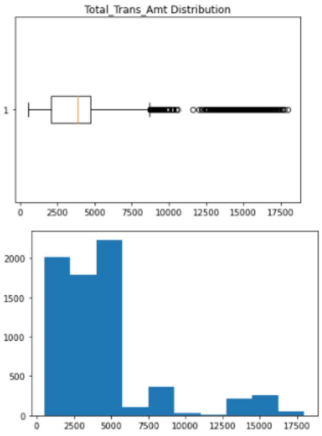


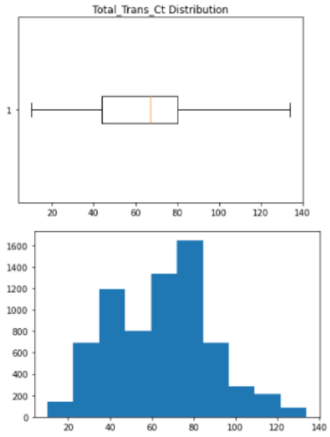


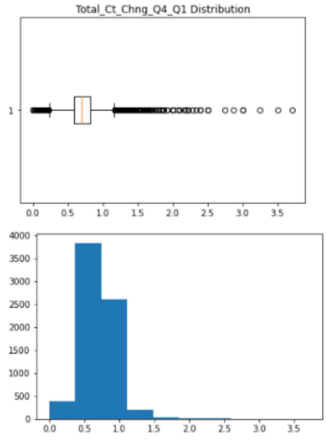


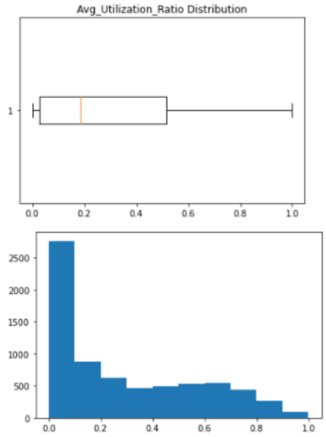


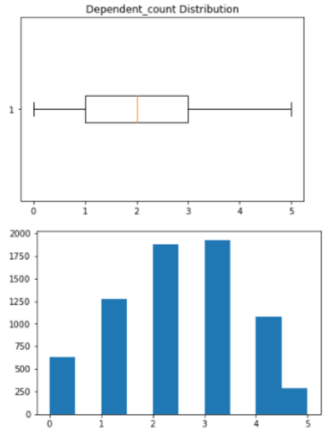


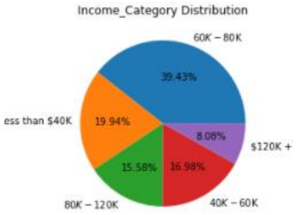


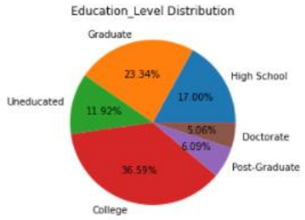


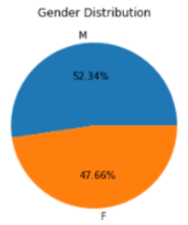


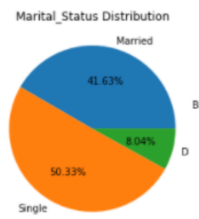


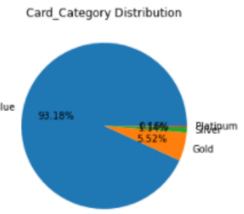




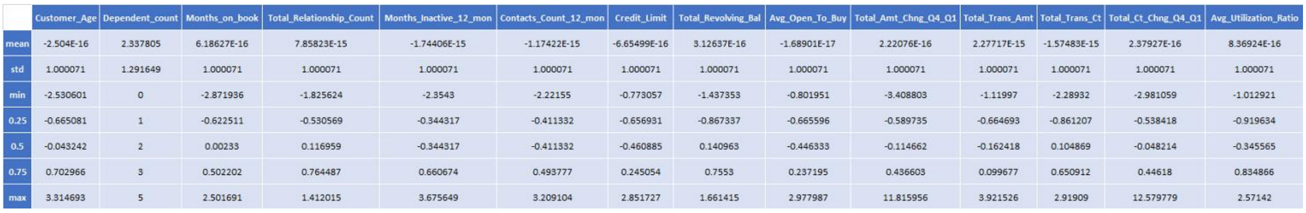








Statistical Overview of Features



	Customer_Age	Dependent_count	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Amt_Chng_Q4_Q1	Total_Trans_Amt	Total_Trans_Ct	Total_Ct_Chng_Q4_Q1	Avg_Utilization_Ratio
mean	-2.504E-16	2.337805	6.18627E-16	7.85823E-15	-1.74406E-15	-1.17422E-15	-6.65499E-16	3.12637E-16	-1.68901E-17	2.22076E-16	2.27717E-15	-1.57483E-15	2.37927E-16	8.36924E-16
std	1.000071	1.291649	1.000071	1.000071	1.000071	1.000071	1.000071	1.000071	1.000071	1.000071	1.000071	1.000071	1.000071	1.000071
min	-2.530601	0	-2.871936	-1.825624	-2.3543	-2.22155	-0.773057	-1.437353	-0.801951	-3.408803	-1.11997	-2.28932	-2.981059	-1.012921
0.25	-0.665081	1	-0.622511	-0.530569	-0.344317	-0.411332	-0.656931	-0.867337	-0.665596	-0.589735	-0.664693	-0.861207	-0.538418	-0.919634
0.5	-0.043242	2	0.00233	0.116959	-0.344317	-0.411332	-0.460885	0.140963	-0.446333	-0.114662	-0.162418	0.104869	-0.048214	-0.345565
0.75	0.702966	3	0.502202	0.764487	0.660674	0.493777	0.245054	0.7553	0.237195	0.436603	0.099677	0.650912	0.44618	0.834866
max	3.314693	5	2.501691	1.412015	3.675649	3.209104	2.851727	1.661415	2.977987	11.815956	3.921526	2.91909	12.579779	2.57142

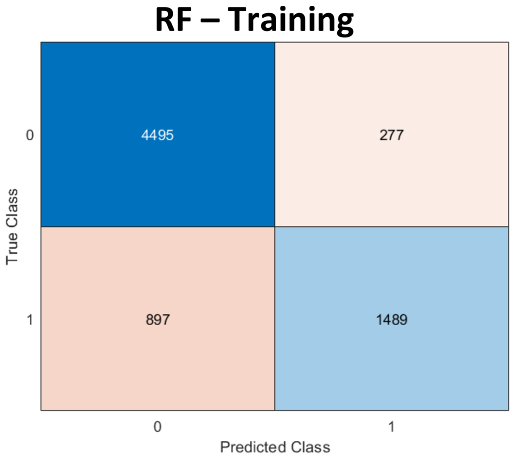
VIF After Removing Avg_Open_To_Buy Feature

Feature	VIF
Customer_Age	2.70
Gender	3.00
Dependent_count	3.78
Education_Level	2.84
Marital_Status	2.06
Income_Category	3.48
Card_Category	1.46
Months_on_book	2.68
Total_Relationship_Count	1.14
Months_Inactive_12_mon	1.03
Contacts_Count_12_mon	1.06
Credit_Limit	2.57
Total_Revolving_Bal	2.56
Total_Amt_Chng_Q4_Q1	1.25
Total_Trans_Amt	3.33
Total_Trans_Ct	3.26
Total_Ct_Chng_Q4_Q1	1.33
Avg_Utilization_Ratio	3.08

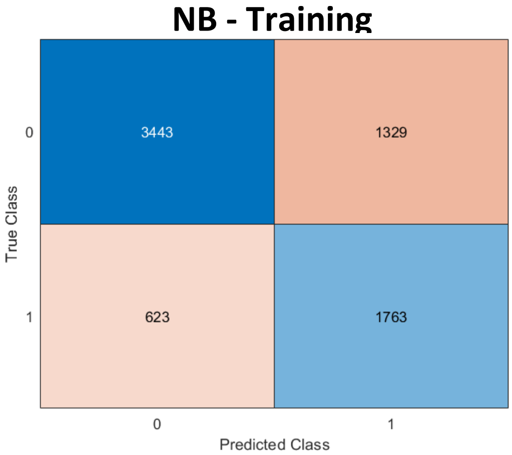
$$\begin{aligned}&Average\ Open\ To\ Buy\\ &\qquad = Credit\ Limit - Total\ Revolving\ Balance\end{aligned}$$

- It was noted that the Total_Revolving_Bal feature is also a function of the Credit_Limit and Avg_Open_To_Buy, but the Total_Revolving_Bal feature is idiosyncratic (i.e. it depends on the customer) while the Avg_Open_To_Buy is the systemic difference between the Credit_Limit and Total_Revolving_Bal.
- This is further supported through the correlation matrix where the correlation between the Credit_Limit and Avg_Open_To_Buy is 1 while the correlation between the Total_Revolving_Bal and the other two features are close to 0. Thus, Avg_Open_To_Buy was omitted instead.
- After omission, the VIF's of all features are finite.

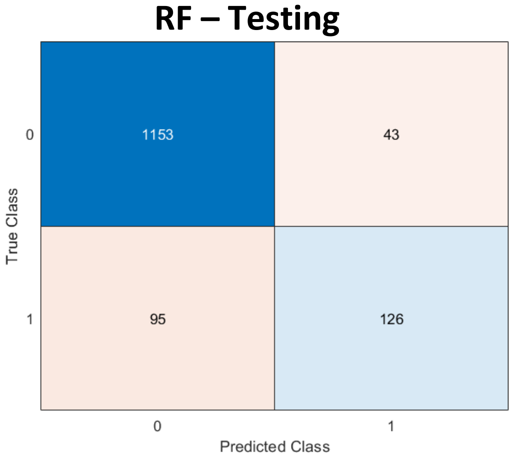
Confusion Matrices of RF and NB Model on Training and Testing Set



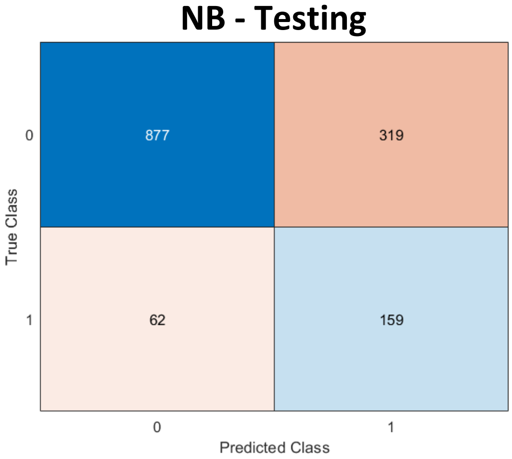
RF – Training



NB - Training



RF – Testing



NB - Testing

Implementation Details

Dataset – Data Preparation and Processing

- The "CLIENTNUM" feature represents customer ID's and was checked for any duplicate values. Given that there were no duplicate values, it was removed from the training and testing set as it should not impact the classification of customers.
- In addition, the last 2 features (i.e. "Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1" and "Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2") was also excluded from the training and testing set because it was not part of original dataset and was included by the individual who posted it on Kaggle.
- SMOTE was adopted so that the minority class is oversampled such that it makes up 50% of the majority class (resulting in a 66.66% to 33.33% split of majority to minority class over the dataset respectively) from the original distribution of around 18.69% of the majority class (resulting in a 84.25% to 15.75% split of majority to minority class over the dataset respectively).
- Correlation heatmap and VIF values were computed and plotted to identify any dependence between features and as mentioned in the intermediate results, the Avg_Open_To_Buy feature was omitted instead of the Total_Revolving_Bal as one is the systemic difference between the Credit_Limit and the Total_Revolving_Bal while the other is idiosyncratic and differs from customer to customer.

Model – Model Training and Hyperparameter Tuning

- The selected values of hyperparameters that were iterated over was chosen by striking a compromise between the ability of observing the impact of the hyperparameters on model performance and the computational intensity.
- When determining the number of bins as one of the hyperparameters, the number of bins was applied to all non-categorical features uniformly (e.g. if 5 bins was considered, all non-categorical features would be binned into 5 bins and there would not be a case where non-categorical feature A is binned into 3 bins while non-categorical feature B is binned into 6 bins). Despite simplifying the binning process significantly, varying the prior distribution in tandem allowed for evident visualization on the impact of binning and prior distribution on the NB model performance.
- In addition, if the number of bins for each non-categorical feature was varied non-uniformly, whilst still varying the prior distribution and employing k-fold cross validation, the computational intensity would increase exponentially. This is due to the fact that by binning uniformly (i.e. all non-categorical features binned by the same number of bins), the number of combinations is independent of the number of features.