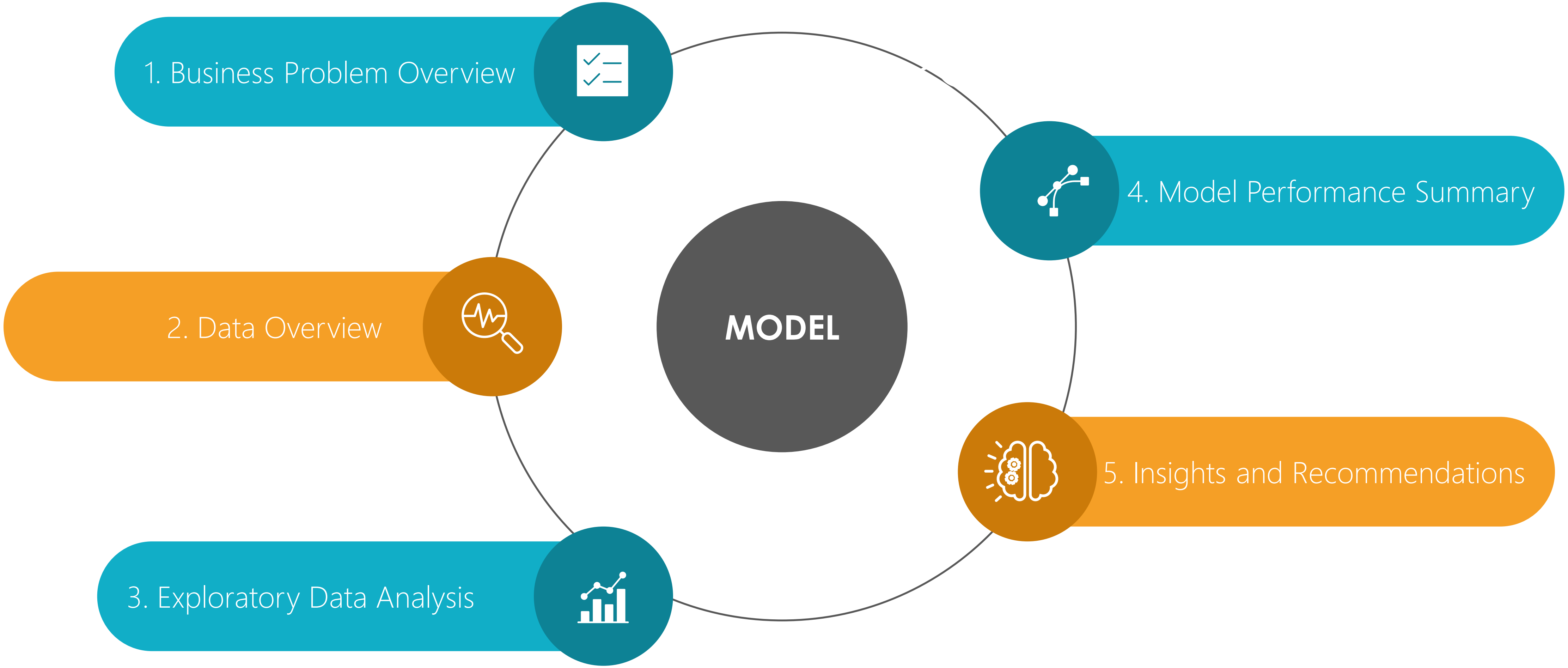
Credit Card Users Churn Prediction

Thera Bank

Contents



MODEL

1. Business Problem Overview

2. Data Overview

3. Exploratory Data Analysis

4. Model Performance Summary

5. Insights and Recommendations

Business Problem Overview

The Thera bank recently saw a steep decline in the number of users of their credit card.

Credit cards are a good source of income for banks because of different kinds of fees charged by the banks like annual fees, balance transfer fees, and cash advance fees, late payment fees, foreign transaction fees, and others.

Some fees are charged to every user irrespective of usage, while others are charged under specified circumstances.

Customers' leaving credit cards services would lead bank to loss, so the bank wants to analyze the data of customers and identify the customers who will leave their credit card services and reason for same – so that bank could improve upon those areas

Data Overview

DATA SHAPE

The data contains 21 information about 20,127 customers. Our data is imbalanced: 83.9% of our customers are Existing Customers on the other hand 16.1% are Attrited Customer.

COLUMNS INFORMATIONS

The data include personal information like Age, Income, Gender, Marital Status, Month on Book, Credit Limit, etc.

COLUMNS DROPPED

We will drop the CLIENTNUM variable from the data as it is unique for each customer and AVG_OPEN_TO_BUY which is perfect correlated to Credit Limit will not add value to our analysis.

FINANCIAL ANALYSIS

The data also includes some financial analysis like ratio of the total transaction count in 4th quarter and the total transaction count in 1st quarter.

Column: Total_Ct_Chng_Q4_Q1

DATA ERROR INPUT

Some observations in the Income_Category variable have "abc" as a value. This must be a data missing, we will replace with NaN and treat it as Missing Value.

MISSING VALUES

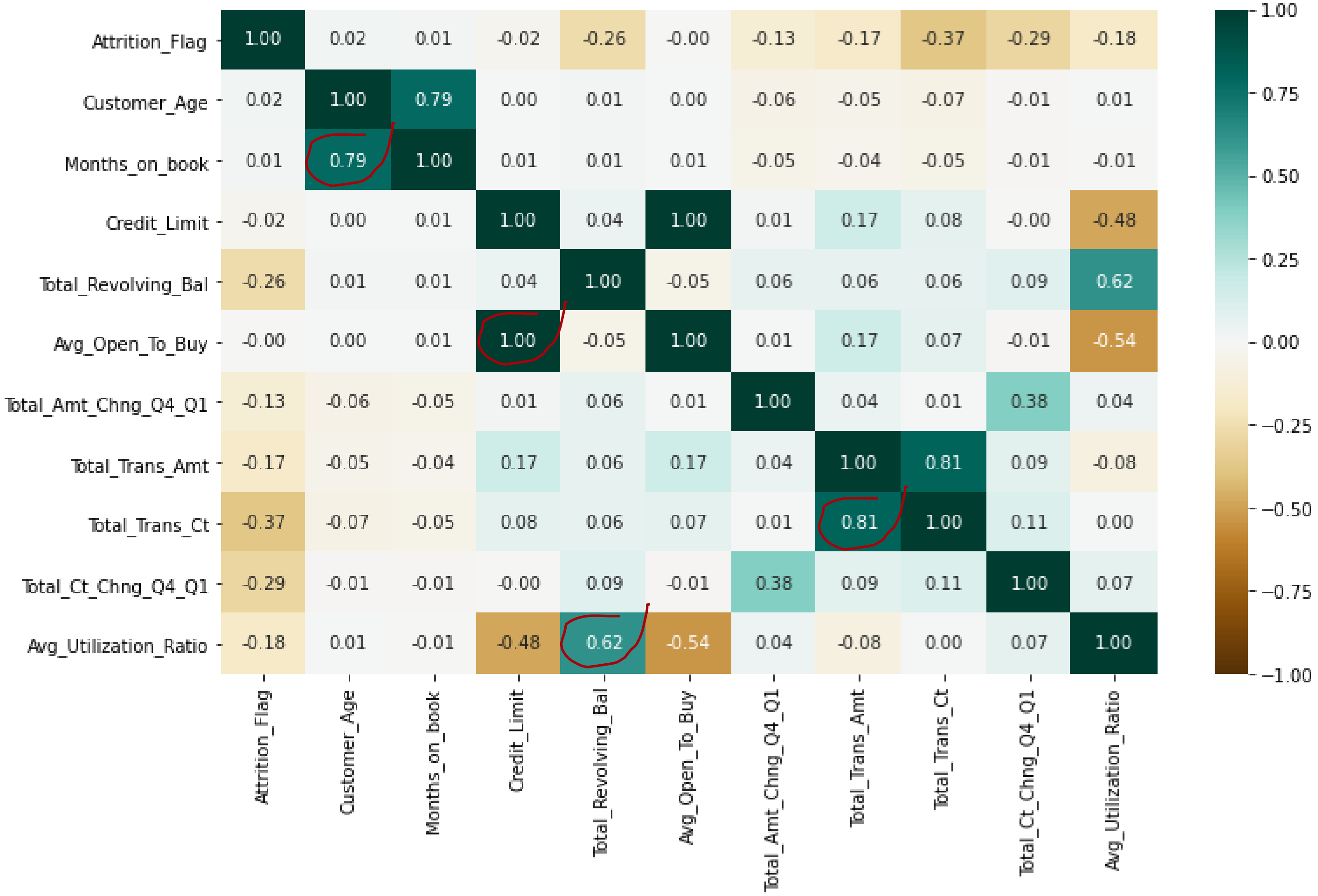
There are 3 columns that contains missing values. We will impute these missing values before building the model using Simple Imputer.

Columns: Income_Category, Marital_Status, Education_Level

OUTLIERS

Some observations have a huge range of values showing potentials outliers, after analyzing them, we understood that in real-life, those are values that will happen, so we will not treat them.

Exploratory Data Analysis



	Attrition_Flag	Customer_Age	Months_on_book	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
Attrition_Flag	1.00	0.02	0.01	-0.02	-0.26	-0.00	-0.13	-0.17	-0.37	-0.29	-0.18
Customer_Age	0.02	1.00	0.79	0.00	0.01	0.00	-0.06	-0.05	-0.07	-0.01	0.01
Months_on_book	0.01	0.79	1.00	0.01	0.01	0.01	-0.05	-0.04	-0.05	-0.01	-0.01
Credit_Limit	-0.02	0.00	0.01	1.00	0.04	1.00	0.01	0.17	0.08	-0.00	-0.48
Total_Revolving_Bal	-0.26	0.01	0.01	0.04	1.00	-0.05	0.06	0.06	0.06	0.09	0.62
Avg_Open_To_Buy	-0.00	0.00	0.01	1.00	-0.05	1.00	0.01	0.17	0.07	-0.01	-0.54
Total_Amt_Chng_Q4_Q1	-0.13	-0.06	-0.05	0.01	0.06	0.01	1.00	0.04	0.01	0.38	0.04
Total_Trans_Amt	-0.17	-0.05	-0.04	0.17	0.06	0.17	0.04	1.00	0.81	0.09	-0.08
Total_Trans_Ct	-0.37	-0.07	-0.05	0.08	0.06	0.07	0.01	0.81	1.00	0.11	0.00
Total_Ct_Chng_Q4_Q1	-0.29	-0.01	-0.01	-0.00	0.09	-0.01	0.38	0.09	0.11	1.00	0.07
Avg_Utilization_Ratio	-0.18	0.01	-0.01	-0.48	0.62	-0.54	0.04	-0.08	0.00	0.07	1.00

As expected, Months_on_book and Customer_Age have a High Positive Correlation, We will do more analysis to see how does it influence on other variables.

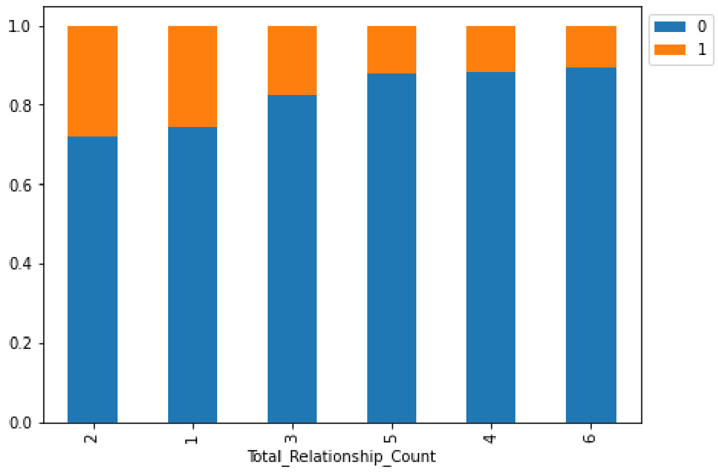
We can drop one of the columns: Avg Open to Buy or Credit Limit as they are Perfectly Correlated and will not add value to our analysis.

Total Trans Ct is positively correlated with Total Trans Amt which can be expected as customers with higher number of transition might spend more than customers with lower number of transition.

Customers that carries over from one month to the next (Total Revolving Bal) have spent a high value of the available credit (Avg Utilization Ratio), this explain the positive correlation between them.

Exploratory Data Analysis

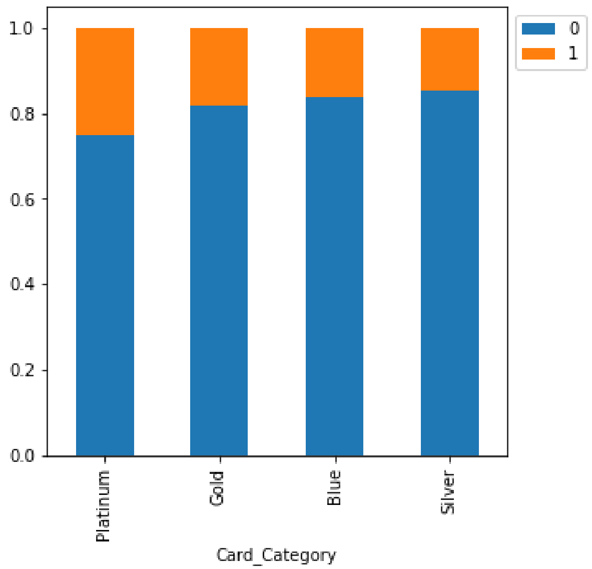
Total Relationship Count:



28% of customers that held 2 products churn, followed by 25% of 1 product held.

The company should work in make customers hold more products keeping them loyal.

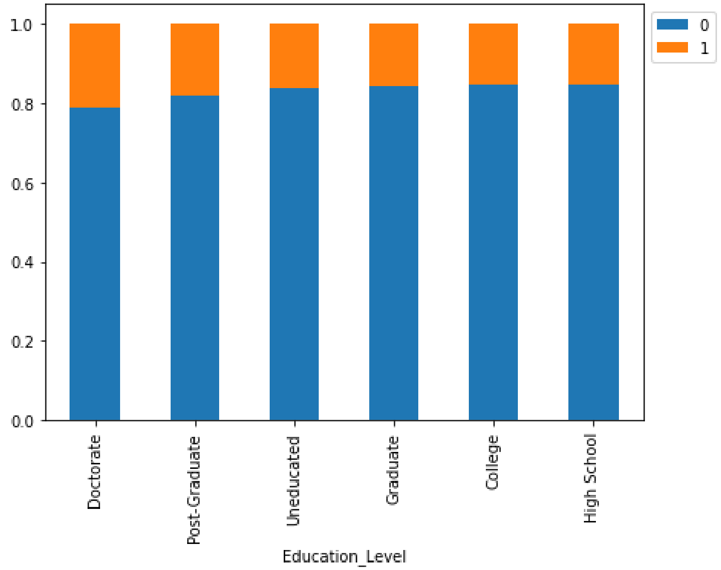
Card Category:



25% of Platinum customers churn, followed by 18% of Gold customers.

The company should look closer this customer profile to understand why they are churn.

Education Level:

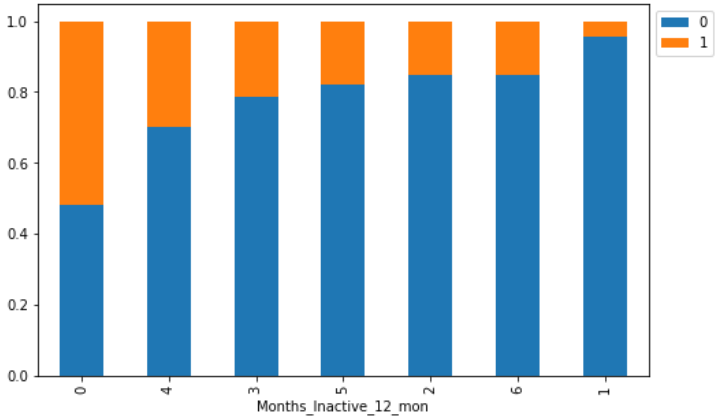


21% of Doctorate customers churn, followed by 18% of Post-Graduate.

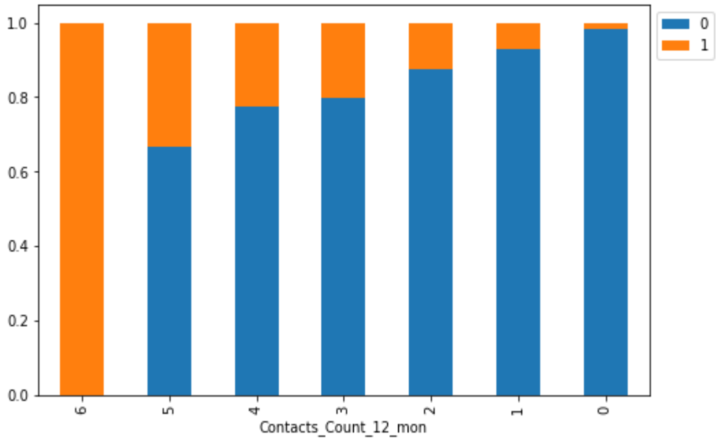
High education level, high income(Platinum/Gold cards) are more likehood to churn.

Exploratory Data Analysis

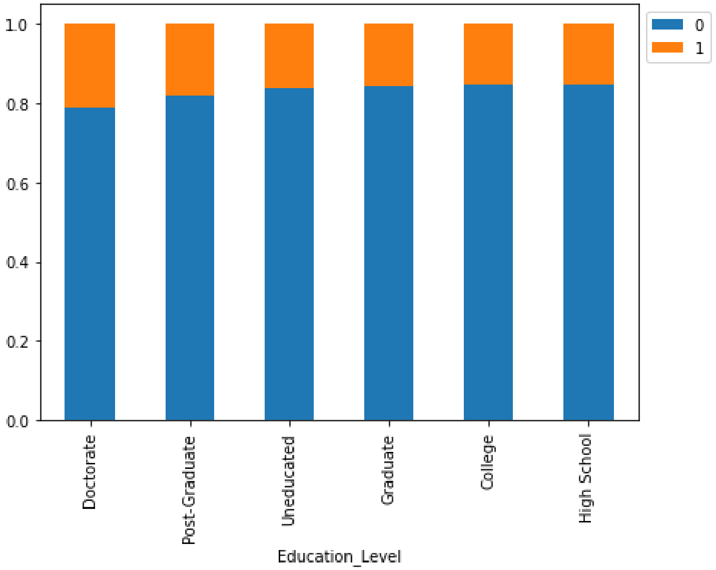
Months Inactive 12 mon:



Contact Count 12 mon:



Education Level:



52% of customers inactive for 0 months surprisingly churn, followed for 4 and 3 months. We need further analysis to understand this behavior.

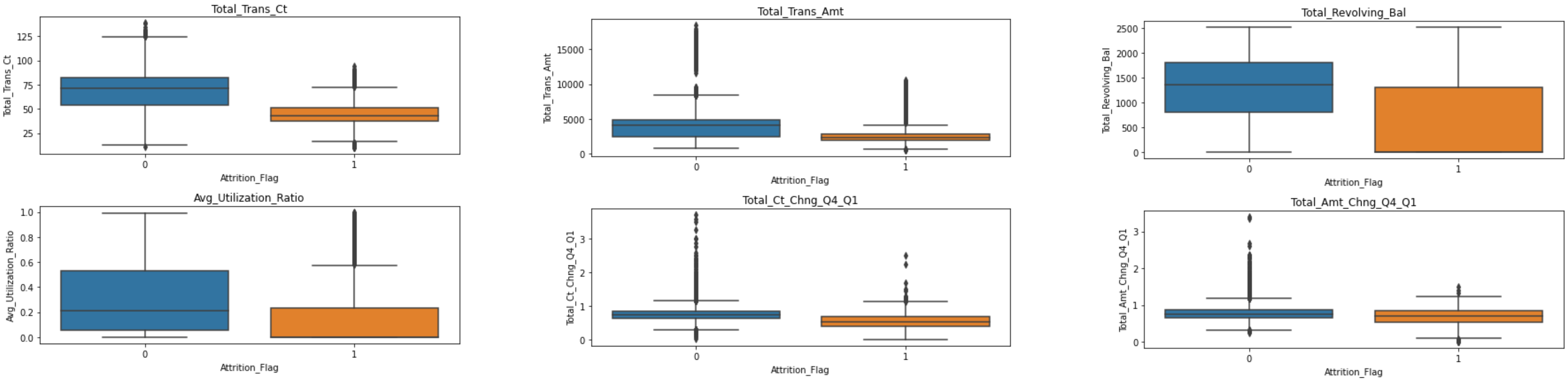
100% of customers with 6 contact on the last 12 months churn, followed 5 and 4 contacts.

Mapping the reason for the contact, and cancelation can help understand it.

21% of Doctorate customers churn, followed by 18% of Post-Graduate.

High education level, high income(Platinum/Gold cards) are more like hood to churn.

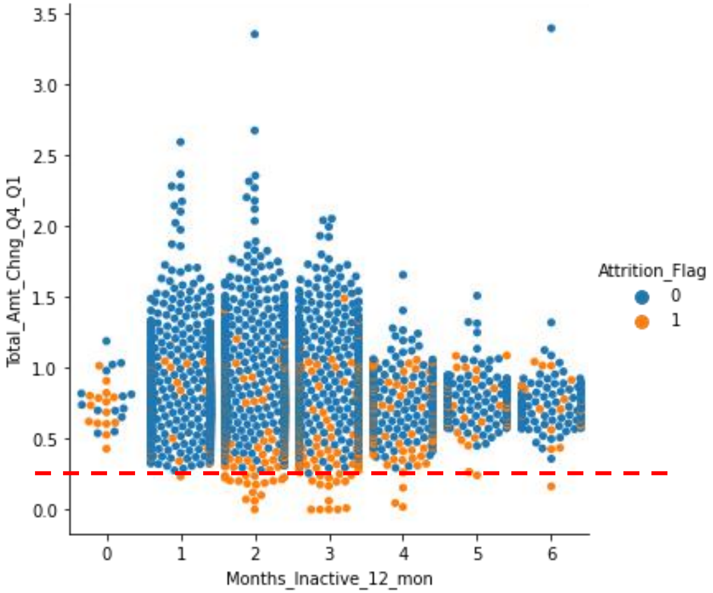
Exploratory Data Analysis



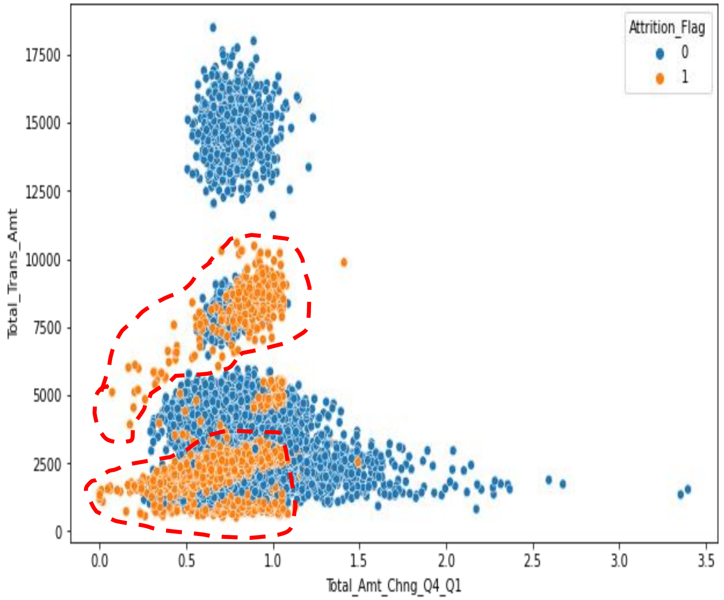
We can see that Customers with lowest Total_Revolving_Bal, Total_Trans_Amt, Total_Trans_Ct, Total_Ct_Chng_Q4_Q1, Total_Amt_Chng_Q4_Q1 and Avg_Utilization_Ratio greater are the chances of the customer to churn.

Exploratory Data Analysis

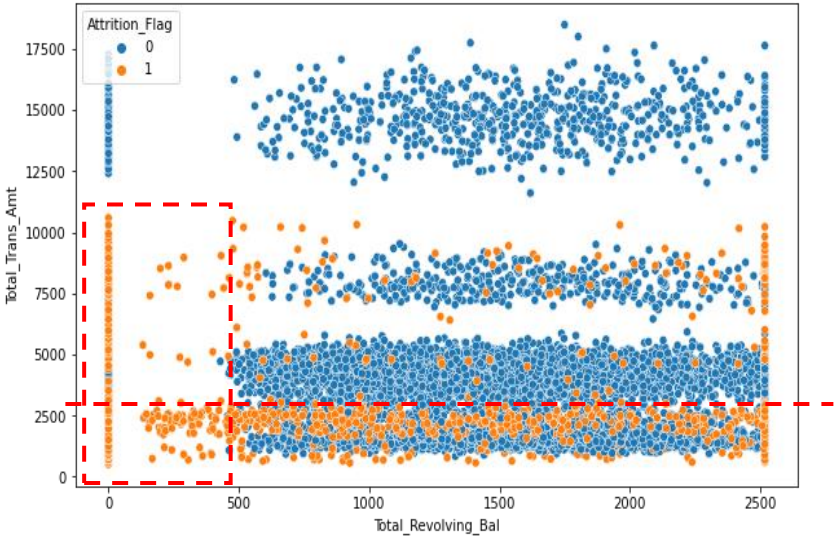
Months Inactive X Total AmtQ4Q1:



Contact Count 12 mon:



Education Level:



Customers that spent less than 0.3 on Q4 compared to Q1 are more like hood to churn.

Increase the usage of Credit Card is a way to keep customers.

We can see some pattern that needs further analyses. Customers with Q4/Q1 less than 1 and Total Trans less 3000 or Customers with Total Trans between 6000 and 11000 with Q4/Q1 less than 1 are more like hood to churn.

Customers with Total Trans Amt less than 3000 are more like hood to churn and the odds increase as the total decreases.

Total Revolving Bal less than 500 increases chances of churn.

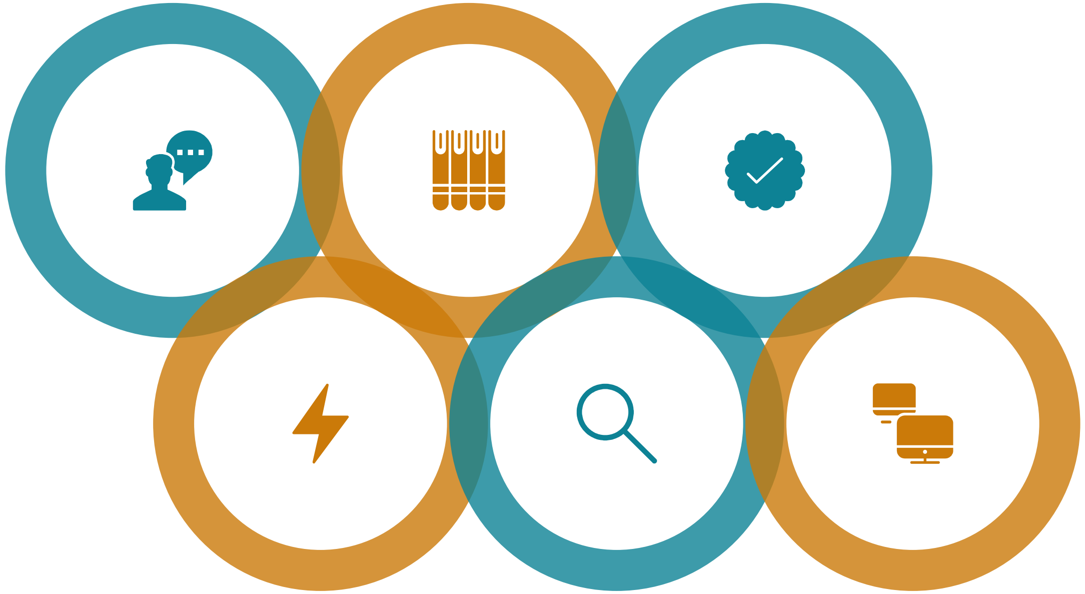
Model Performance Summary

I. We want to predict if the customer is going to churn or not.

II. We will use Recall as the metric to evaluate our model and try to minimize the number of false negatives.

III. Incorrectly predicting that the customer is not going to churn will result in losing on a potential source of income for the company because that customer will not be targeted by the marketing team when it should be targeted.

IV. We will build different models (Bagging, Random Forest, Gradient Boosting, AdaBoost, XGBoost and Decision Tree.



V. We will build different models using Oversampled and Undersampled data.

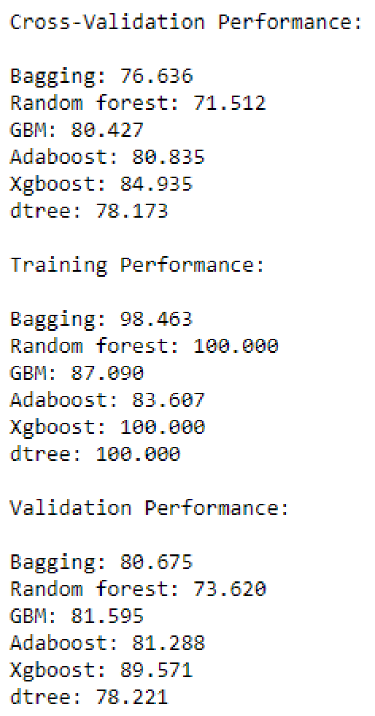
VI. We will choose 3 model that might perform better after tuning and we will do Hyperparameter tuning using Random Search.

VII. The best model with Recall >0.95 and Precision >70 we will apply on our test set.

VIII. We will create a final model using Pipelines.

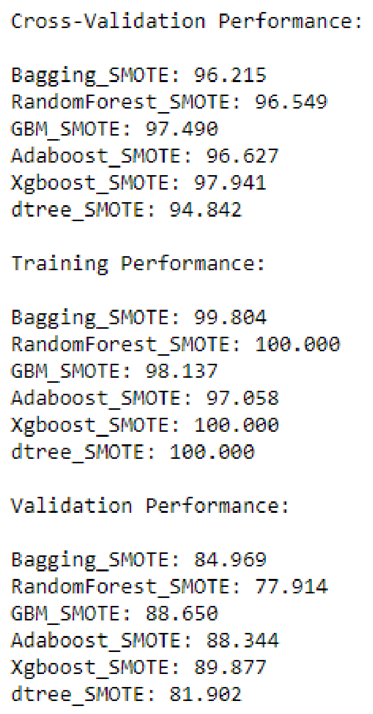
Exploratory Data Analysis

Model Building: Default parameters:



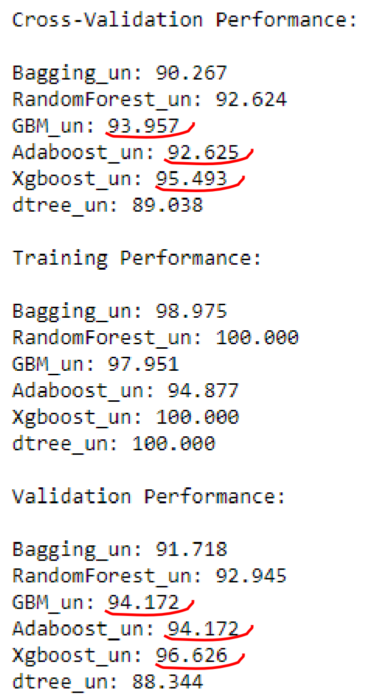
```
Cross-Validation Performance:

Bagging: 76.636
Random forest: 71.512
GBM: 80.427
Adaboost: 80.835
Xgboost: 84.935
dtree: 78.173

Training Performance:

Bagging: 98.463
Random forest: 100.000
GBM: 87.090
Adaboost: 83.607
Xgboost: 100.000
dtree: 100.000

Validation Performance:

Bagging: 80.675
Random forest: 73.620
GBM: 81.595
Adaboost: 81.288
Xgboost: 89.571
dtree: 78.221
```

Model building - Oversampled data



```
Cross-Validation Performance:

Bagging_SMOTE: 96.215
RandomForest_SMOTE: 96.549
GBM_SMOTE: 97.490
Adaboost_SMOTE: 96.627
Xgboost_SMOTE: 97.941
dtree_SMOTE: 94.842

Training Performance:

Bagging_SMOTE: 99.804
RandomForest_SMOTE: 100.000
GBM_SMOTE: 98.137
Adaboost_SMOTE: 97.058
Xgboost_SMOTE: 100.000
dtree_SMOTE: 100.000

Validation Performance:

Bagging_SMOTE: 84.969
RandomForest_SMOTE: 77.914
GBM_SMOTE: 88.650
Adaboost_SMOTE: 88.344
Xgboost_SMOTE: 89.877
dtree_SMOTE: 81.902
```

Model building - Undersampled data



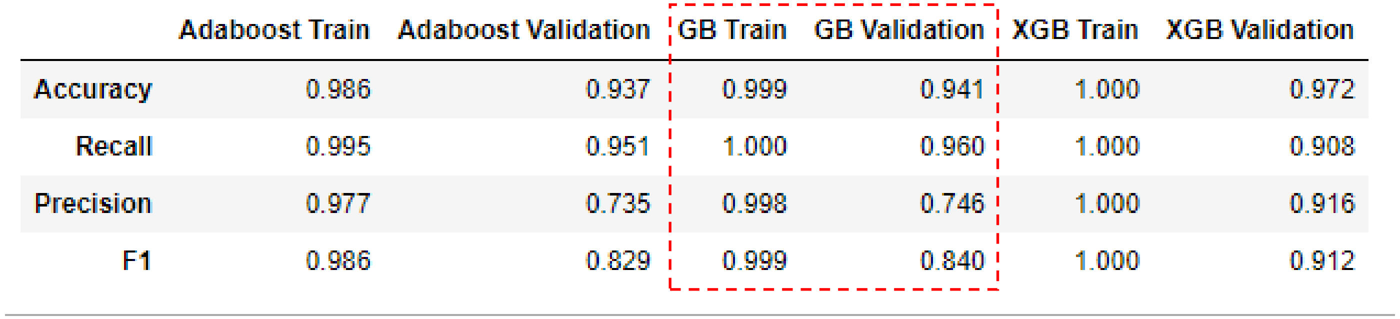
```
Cross-Validation Performance:

Bagging_un: 90.267
RandomForest_un: 92.624
GBM_un: 93.957
Adaboost_un: 92.625
Xgboost_un: 95.493
dtree_un: 89.038

Training Performance:

Bagging_un: 98.975
RandomForest_un: 100.000
GBM_un: 97.951
Adaboost_un: 94.877
Xgboost_un: 100.000
dtree_un: 100.000

Validation Performance:

Bagging_un: 91.718
RandomForest_un: 92.945
GBM_un: 94.172
Adaboost_un: 94.172
Xgboost_un: 96.626
dtree_un: 88.344
```

- The best Cross Validation performance happened using Oversampling, on the other hand Undersampling is showing more consistency between Cross Validation and Validation Set.
- Our top 3 better expected performance on unseen data considering Cross Validation and Performance on Validation set happens with UNDERSAMPLING data set on the following models: XGBoost, Gradient Boosting and Adaboost, all showing consistence and 2 outliers on GBM and no outliers on the others.
- We will proceed with Hyperparameters tuning using random search to try to find the best set of hyperparameters and this method also takes less time than GridSearch.

Model Performance Summary

- The Gradient Boosting model tuned using Random Search is giving the best validation recall of 0.954 and it has a precision value greater than 70 on under sampling data.
- XGB is giving a good Recall and Precision around 90, but our metric of interest is Recall.

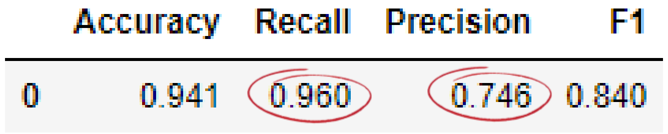
Training X Validation performance on Random Search::



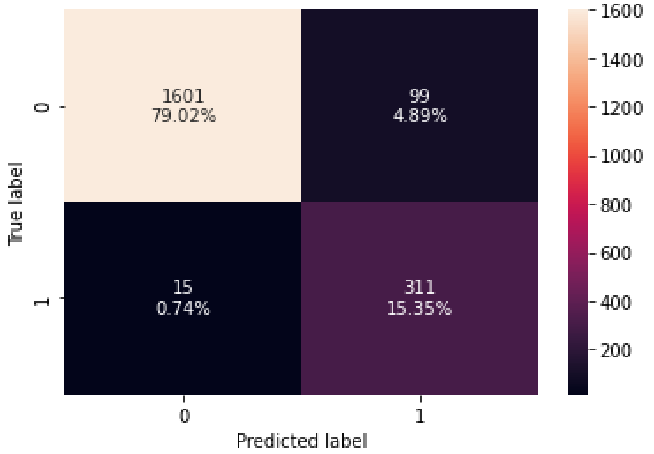
	Adaboost Train	Adaboost Validation	GB Train	GB Validation	XGB Train	XGB Validation
Accuracy	0.986	0.937	0.999	0.941	1.000	0.972
Recall	0.995	0.951	1.000	0.960	1.000	0.908
Precision	0.977	0.735	0.998	0.746	1.000	0.916
F1	0.986	0.829	0.999	0.840	1.000	0.912

Test data is performing well, with a really good Recall score of 0.96 and a good Precision 0.746.

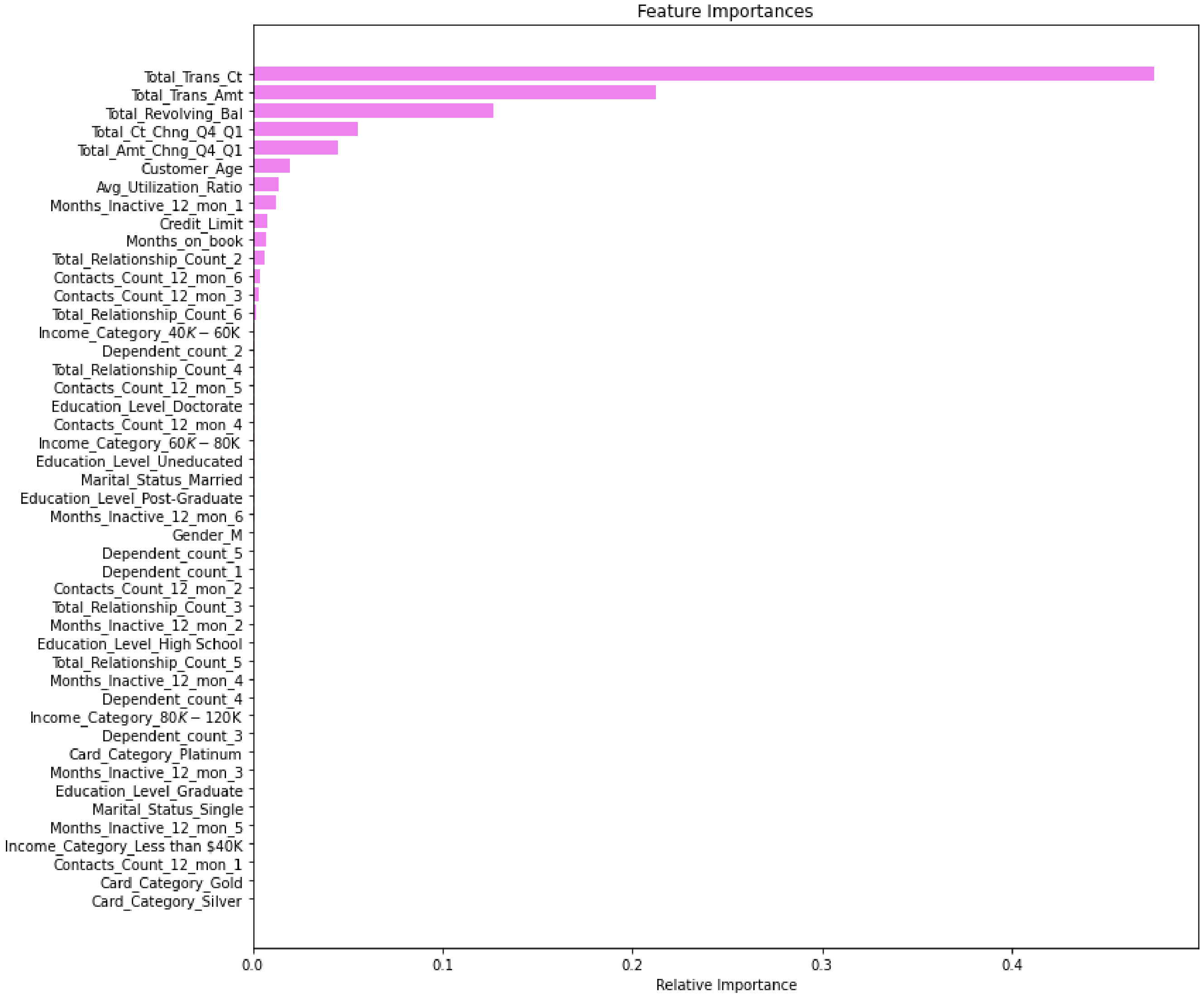
Performance on the test set:



	Accuracy	Recall	Precision	F1
0	0.941	0.960	0.746	0.840



Model Performance Summary



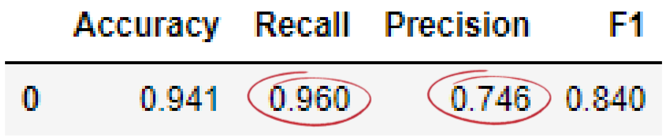
Total Trans Ct is the most important feature, followed by Total Trans Amt, Total Revolving Bal, Total CT Chng Q4 Q1 and Total Amt Chng Q4 Q1, as per the tuned undersampling data set on Gradient Boosting model.

Pipelines for productionizing the model

- We will create 2 different pipelines, one for numerical columns and one for categorical columns
- For numerical columns, we will do missing value imputation as pre-processing
- For categorical columns, we will do one hot encoding and missing value imputation as pre-processing
- We are doing missing value imputation for the whole data, so that if there is any missing value in the data in future that can be taken care of.

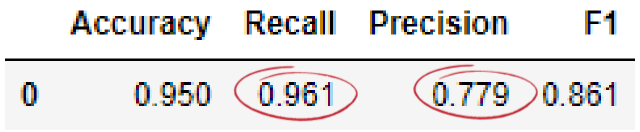
- Our Pipeline model is performing well on our data, with 0.961 Recall and 0.779 Precision.
- We did Simple Imputer for missing values considering median for numerical and most frequent for categorical.
- We used undersampling data set and applied the Hyperparameter tuning with Random Search on Gradient Boosting model.

Performance on the test set:



	Accuracy	Recall	Precision	F1
0	0.941	0.960	0.746	0.840

Performance with Pipeline:



	Accuracy	Recall	Precision	F1
0	0.950	0.961	0.779	0.861

Insights and Recommendations

MODEL PERFORMANCE

- The best model is Gradient Boosting on undersampling set, that score on Recall is greater than 95 and Precision greater than 70. This means that the model is good at identifying churn customers.
- The model performance still really good on test set, Recall 96 and Precision 75 and applying Pipeline it still giving a great performance 96 and 78 Recall and Precision respectively, over all we can see that our model is showing consistence.
- Total Trans Ct, Total Trans Amt, Total Revolving Bal, Total CT Chng Q4 Q1 and Total Amt Chng Q4 Q1, are the important variables in determining if the customer will churn or not, less they use the credit card, greater is the chance of canceling it.

OPPORTUNITY

- We saw in our analysis that customers with high usage of the Credit Card are more likely to not churn.
- How can we increase the usage of the Credit Card? Promotions, Points, Cashback, Free Fee.
- Understand what customers (Soliciting Customers feedback via surveys) wants and what others Credit Card are offering can help us develop the best marketing campaign.
- Understand how satisfied are the customers, using our call center to tabulate the data capturing customers concerns, monitor complains and develop dashboards to detect addressable patterns can help us to keep our customers and increase our customer base.

Thank You

Amanda Mendonca