THERA BANK

data analytics – credit card churn prediction

by Marcelo Moraes

DATA ANALYTICS FOCUS

- BRING AWARENESS ABOUT THE MAIN INDICATORS OF A POSSIBLE CHURN
- PROVIDE A BETTER UNDERSTANDING ABOUT CUSTOMER NEEDS
- BUILD A MODEL TO PREDICT CREDIT CARD CHURN

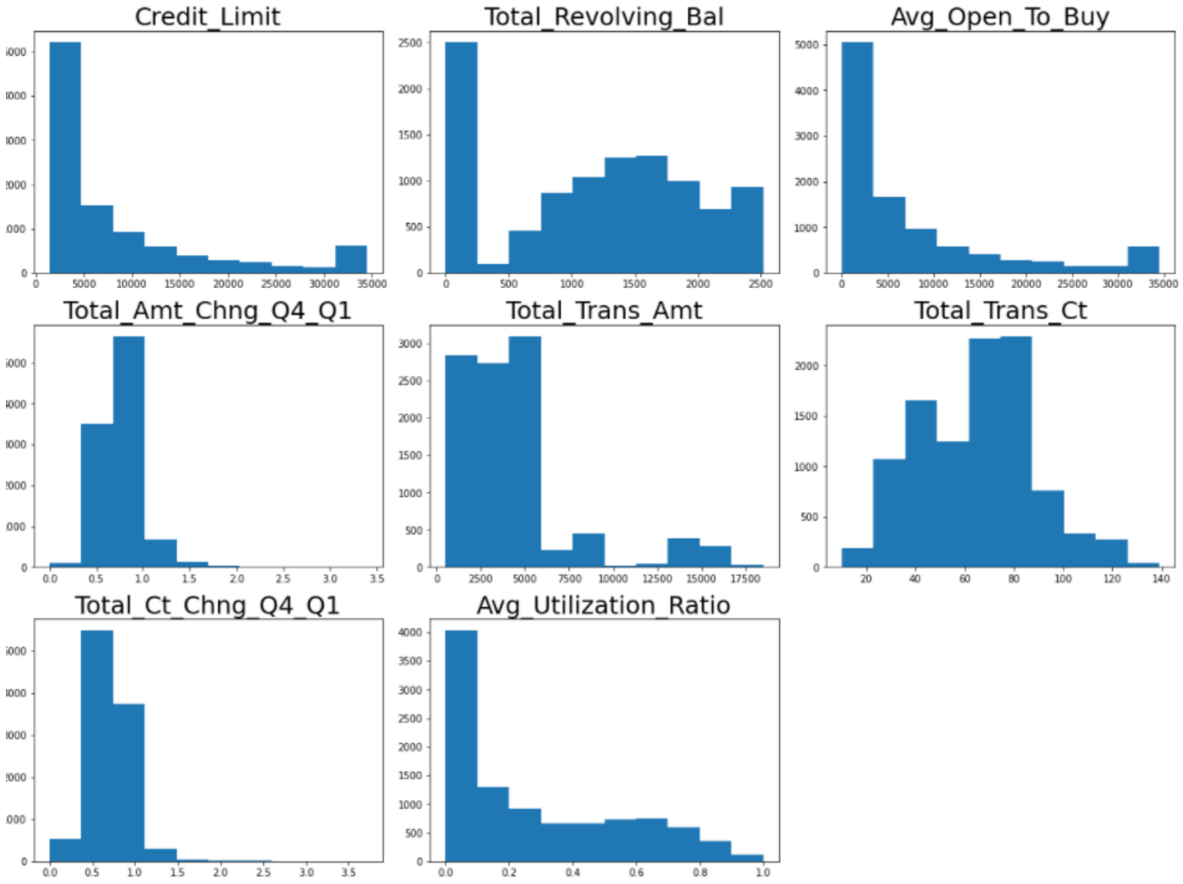
Variable	Description
Client Number	Unique identifier for the customer holding the account
Attrition_Flag	Account closed = 1 / Current customer =
Customer_Age	Age in Years
Gender	Gender of the account holder
Dependent_count	Number of dependents
Education_Level	Educational Qualification of the account holder
Marital_Status	Marital Status of the account holder
Income_Category	Annual Income Category of the account holder
Card_Category	Type of Card
Months_on_book	Period of relationship with the bank
Total_Relationship_Count	Total no. of products held by the customer
Months_Inactive_12_mon	No. of months inactive in the last 12 months
Contacts_Count_12_mon	No. of Contacts in the last 12 months
Credit_Limit	Credit Limit on the Credit Card
Total_Revolving_Bal	Total Revolving Balance on the Credit Card
Avg_Open_To_Buy	Open to Buy Credit Line (Average of last 12 months)
Total_Amt_Chng_Q4_Q1	Change in Transaction Amount (Q4 over Q1)
Total_Trans_Amt	Total Transaction Amount (Last 12 months)
Total_Trans_Ct	Total Transaction Count (Last 12 months)
Total_Ct_Chng_Q4_Q1	Change in Transaction Count (Q4 over Q1)
Avg_Utilization_Ratio	Average Card Utilization Ratio

DATA INFORMATION

(212667 data points)
(0% missing values)

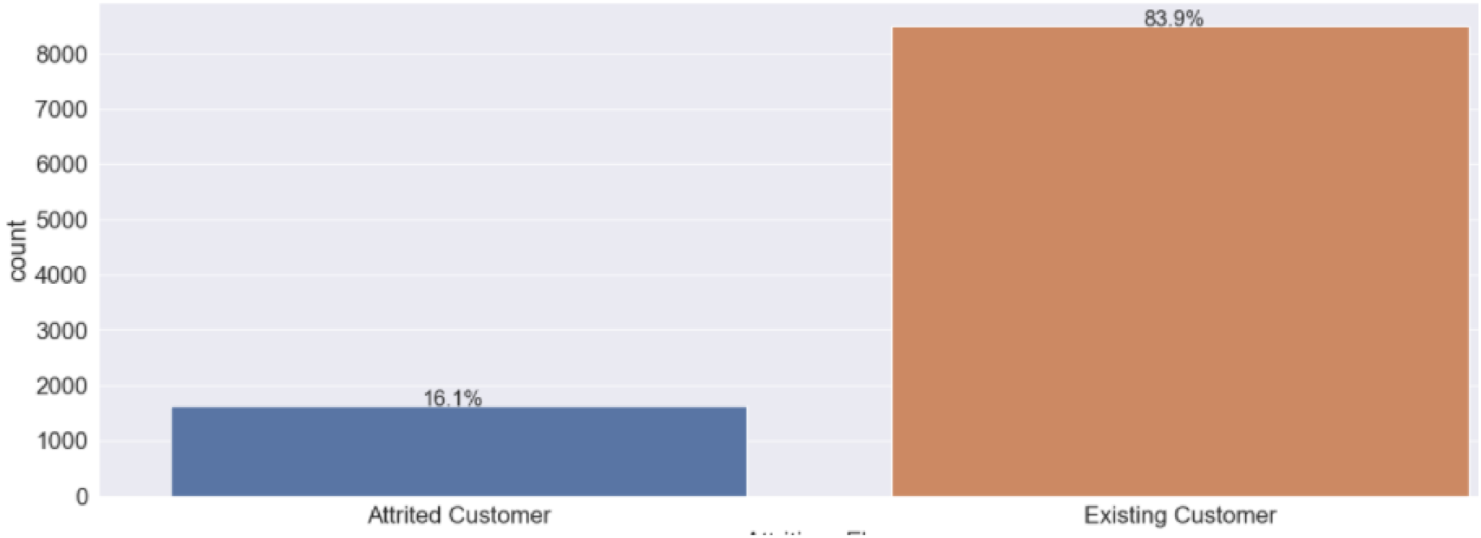
DATA OBSERVATION

- The dataset presents a lot of variables with outliers.
- The decision will be to leave outliers as is since this is a classification problem and most algorithms used in these study is not affected by outliers.



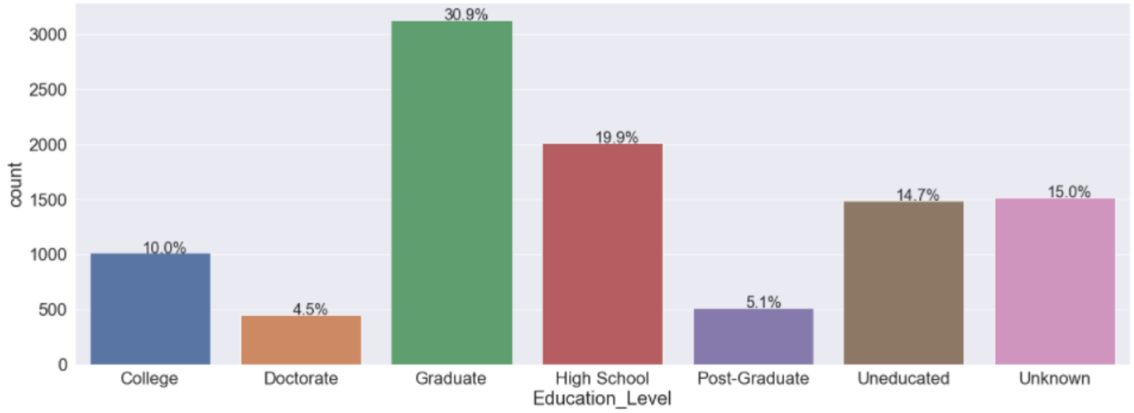
DATA OBSERVATION

- The representation of the class of customers with attrition represents only 16.1% of the data set making it unbalanced. There is high likelihood Machine Learning algorithm will be biased towards the bigger class since there are not much data for the underrepresented class to learn.
- Some technics were used to balance the data to lower the bias and increase variance.



- There is a bigger representation of customers with income less than $40k, the remaining income classes are more spread out.

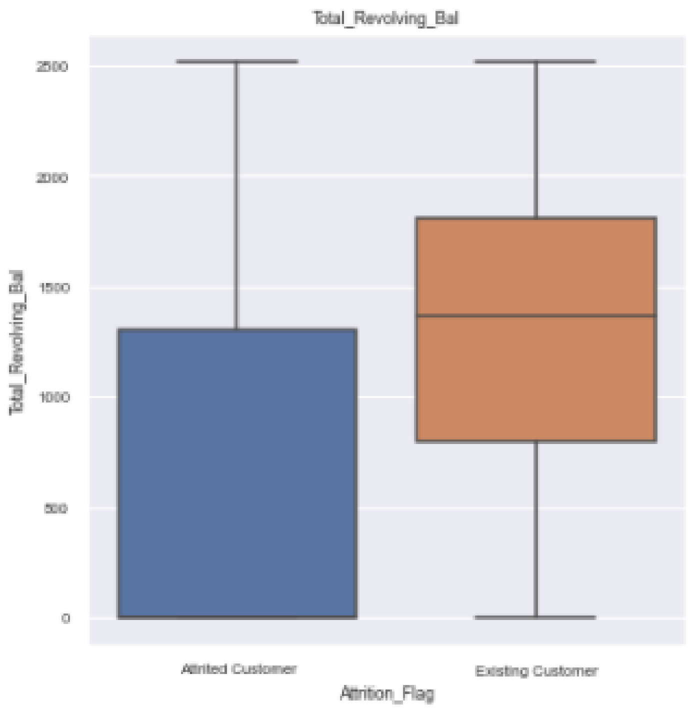
- Education with 30% of customers at graduate level

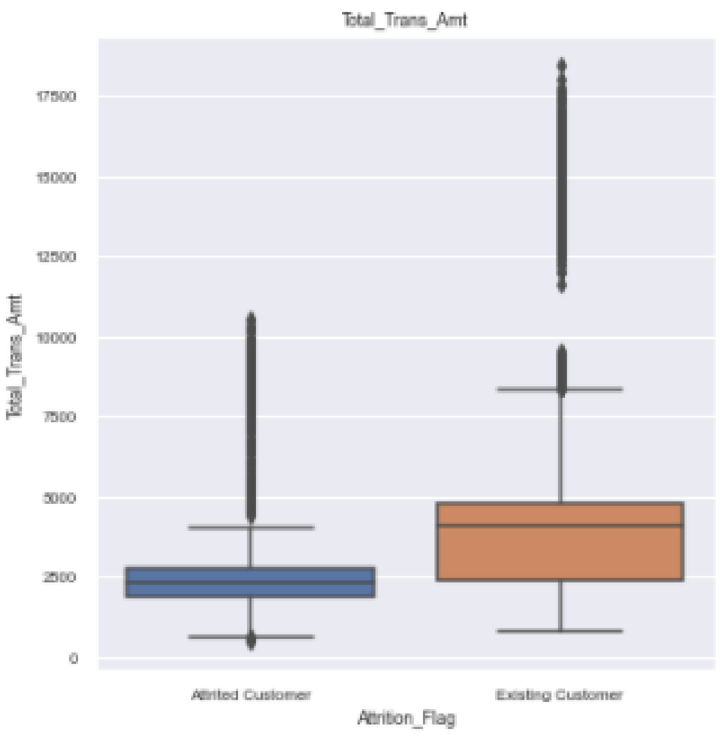


DATA OBSERVATION

DATA INSIGHTS

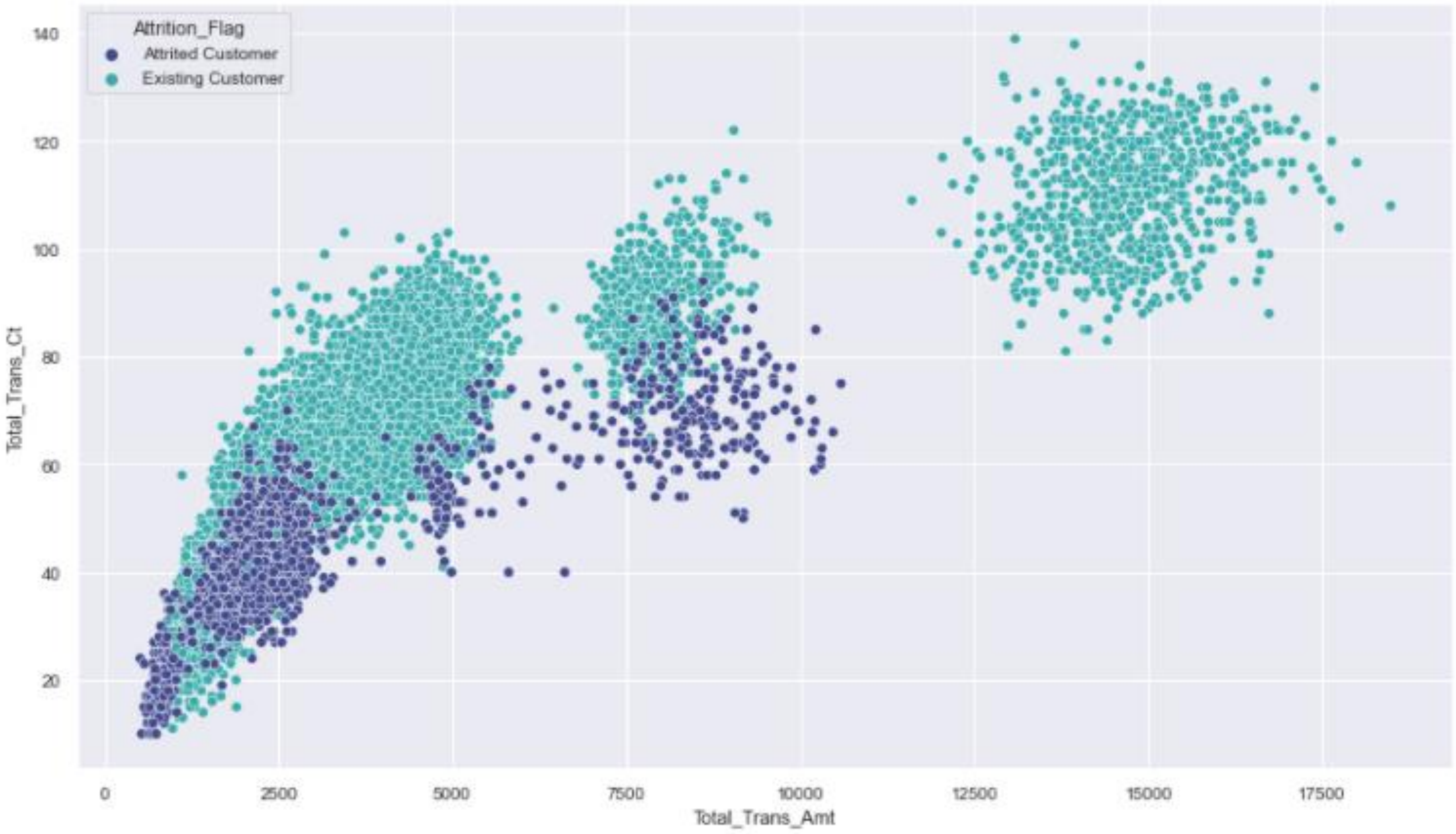
- Total revolving balance and Total Transaction Amount values show a big difference in average for attrited customer in comparison with existent customers.





* Red line showing sales trend

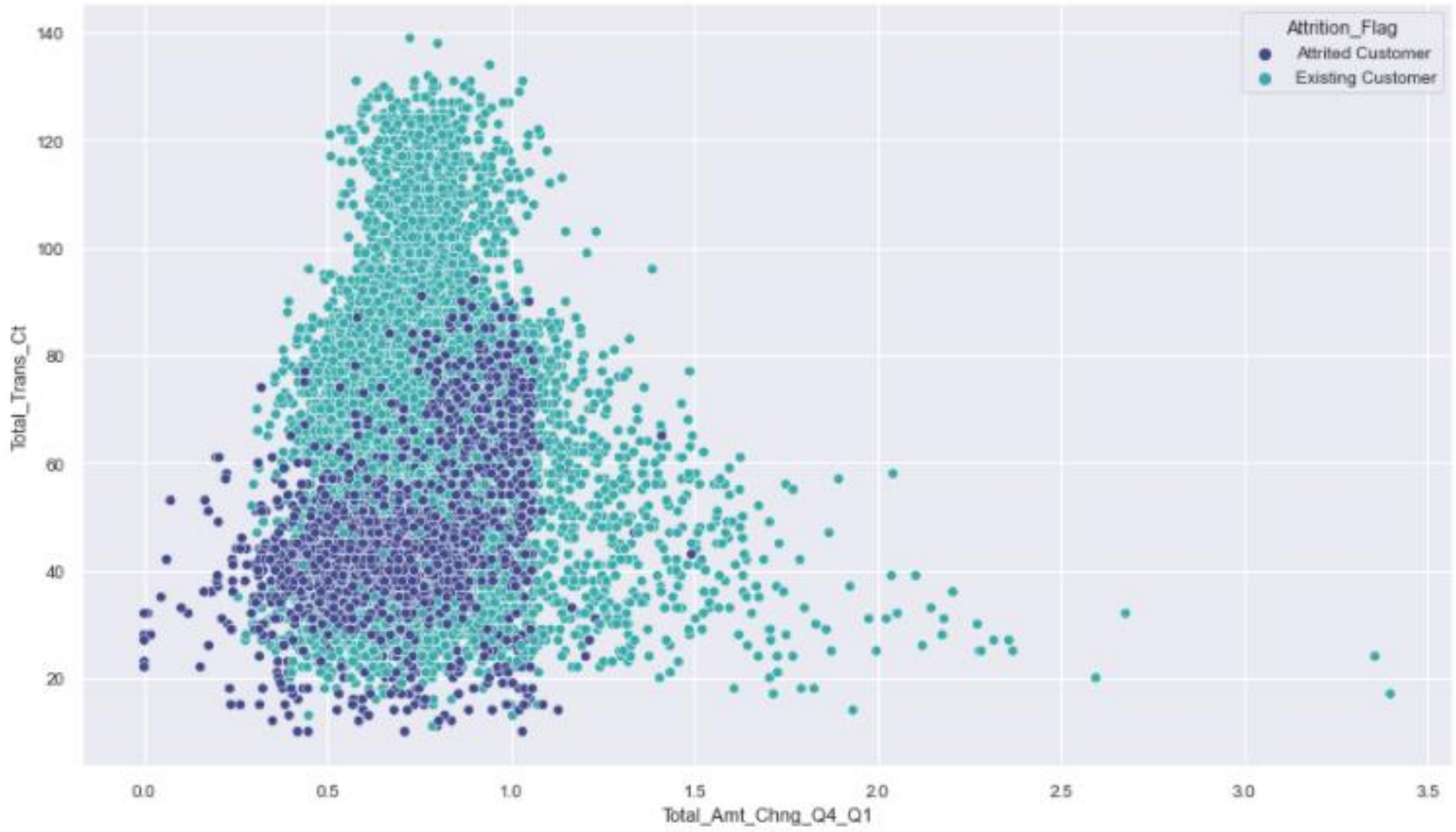
- Attrited customers have a lower transaction amount and lower quantity as well in comparison to current customers



* Red line showing sales trend

DATA INSIGHTS

- Attrited customers almost never shows a reduced ratio between quarters. In other words, customers on Q4 that lower their amount to Q1. However, customers which present this behavior is certain that will not attrite.



* Red line showing sales trend

DATA INSIGHTS

MODEL OVERVIEW

- 12 machine learning models were used in attempt to find the best predictor of customer churn.

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall
0	Logistic Regression	0.875000	0.875617	0.427568	0.422131
1	Logistic Regression on Oversampled data	0.832913	0.816387	0.824340	0.756148
2	Logistic Regression-Regularized (Oversampled d...	0.705076	0.804212	0.569003	0.551230
3	Logistic Regression on Undersampled data	0.782704	0.796315	0.788411	0.799180

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall
5	XGBoost with RandomizedSearchCV	0.944413	0.932544	0.991220	0.969262
4	XGBoost with GridSearchCV	0.970514	0.948996	1.000000	0.954918
6	GradientBoost with UnderSampling GridSearchCV	0.964870	0.948667	0.994732	0.954918
7	GradientBoost with UnderSampling- Manual adjust	0.914503	0.905232	0.950834	0.934426
0	AdaBoost with GridSearchCV	0.997884	0.967423	0.992976	0.870902
1	AdaBoost with RandomizedSearchCV	0.997884	0.967423	0.992976	0.870902
3	GradientBoost with RandomizedSearchCV	0.997884	0.971043	0.988586	0.870902
2	GradientBoost with GridSearchCV	0.990265	0.969727	0.956102	0.856557

*ridge and lasso reg. models were not included on the list since both presented a very low performance

MODEL METRICS

- Recall was the metric adopted to evaluate performance of the model:
    - Predict a customer will not leave their credit card services but they in fact DO leave (FN)
- The focus of the model was to maximize Recall, consequently lowering False Negative rates.
- By using a classification model with a very good performance in reduce false negatives Bank can improve its revenue and their services so that customers do no renounce their credit cards. It also bring awareness about the main indicators of a possible churn so that bank can proactively act towards those customers needs and revert the situation positively.

MODEL SELECTION

- The 3 models below were the ones which performed best among all others:

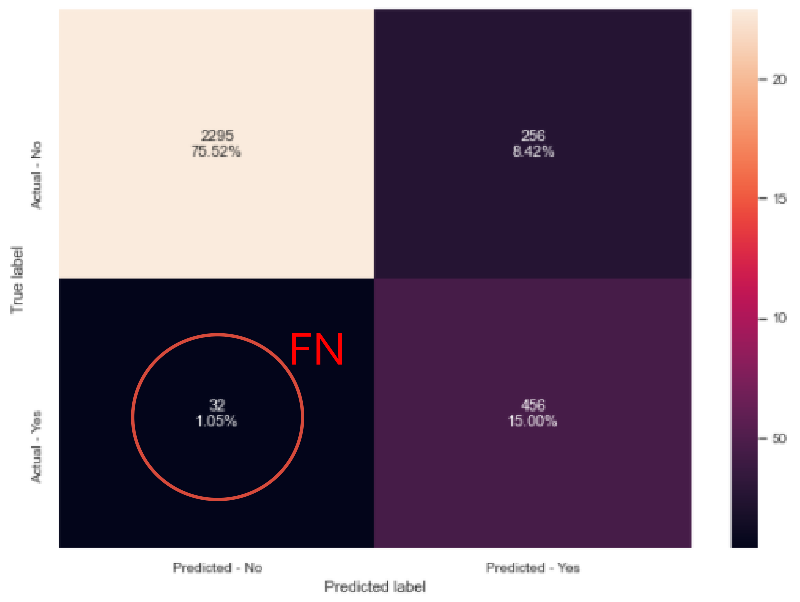
	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall
3	Logistic Regression on Undersampled data	0.782704	0.796315	0.788411	0.799180
7	GradientBoost with UnderSampling- Manual adjust	0.914503	0.905232	0.950834	0.934426
2	GradientBoost with GridSearchCV	0.990265	0.969727	0.956102	0.856557

- GradientBoost using under sampling technic was selected as the best predictor with the highest recall value on test data.

7	GradientBoost with UnderSampling- Manual adjust	0.914503	0.905232	0.950834	0.934426

- Main reasons for the model selection:
  - First, GBC tuned with down sampling had the highest recall .
  - Second, model presented lower chances to overfitting since the recall delta between train and test was very close and consistent among all 3 models.

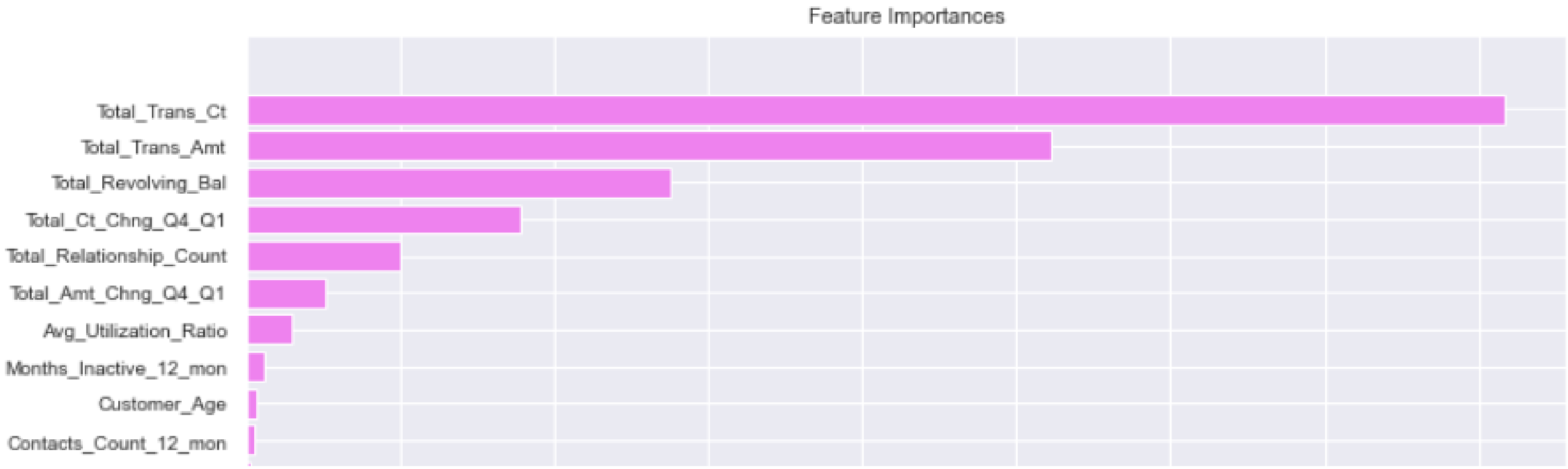
- Confusion Matrix, main points overview



- The winner model (GBC Tuned) conclusion:
  - Model was capable to minimize false negative error by 1.05%. In other words, the model was capable to correctly predict 90.52% of customers would attrite or not.
  - The remaining 8.42%(FP) model predicted a customer would attrite, but in fact they did not.

MODEL OVERVIEW

- Feature importance shows what were the most important features in ascending order the model used and get the obtained results.



- This is a list of features excluding the ones highly correlated: Total Trans Ct, Total Revolving Bal, Total Ct Changed Q4_Q1 and Total Relationship Count as the top 4 features to keep track of it.

MODEL OVERVIEW

RECOMMENDATIONS FOR THE BUSINESS

- Create visualization tools with thresholds to the upper and lower side based on most important features and business definition.
- Create an action plan to keep business prompt to react when customers are at risk of attrition.
- Create alerts for main stake holders in the corporation to implement action plan when required.
- Total revolving balance is a big factor for customer churn. Offerings how to support customers to pay their bills and split the balance and several payments may help customers and increase their satisfaction.
- Data provided does cover customer behavior but does not contain user experience information or customer's feedback. Such data aligned with information raised on these study can a powerful tool to improve service and customers satisfaction.
- As number of contacts in a year does have influence in attrition, implement surveys to identify the quality of the service might bring valuable information about the current service support.