Capstone Presentation (Predicting customer churn)

Jerome Tan

Contents

Background

Customer churn, also known as customer attrition rate, is when a customer chooses to stop using a particular products or services.

Every year, companies lose
<u>$ 1.6 trillion</u>
in revenue due to customer churn.

The cost of acquiring a new customer is
<u>5 times</u>
the cost of retaining an existing customer.

In financial services, **<u>a 5% increase</u>** in customer retention produces more than **<u>25% increase</u>** in profits.

Therefore, it would be beneficial to understand and reduce the churn rate as this will impact the revenue of a business.

Problem Statement

In this project, we will be focusing on churn for bank's credit card customers. We have been engaged by a bank to analyze data collected from existing customers to predict which customer is likely to churn so that the bank can focus their marketing efforts towards such customers and try to provide better services so as to be able to retain existing customers.

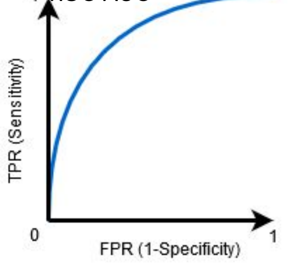
Project Goals

Classification model → Model Deployment

- False Negatives
- False Positives

Metrics

ROC-AUC



$$F_1\text{−score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

Streamlit

Datasets

There are 3 datasets that we are using for this project, attrition, customer and card. Attrition consists data on the attrition status of the customers, customer consists mainly of the customer data and card consists mainly of the customer's card data.

Attrition

Client Number
Attrition Flag

Customer

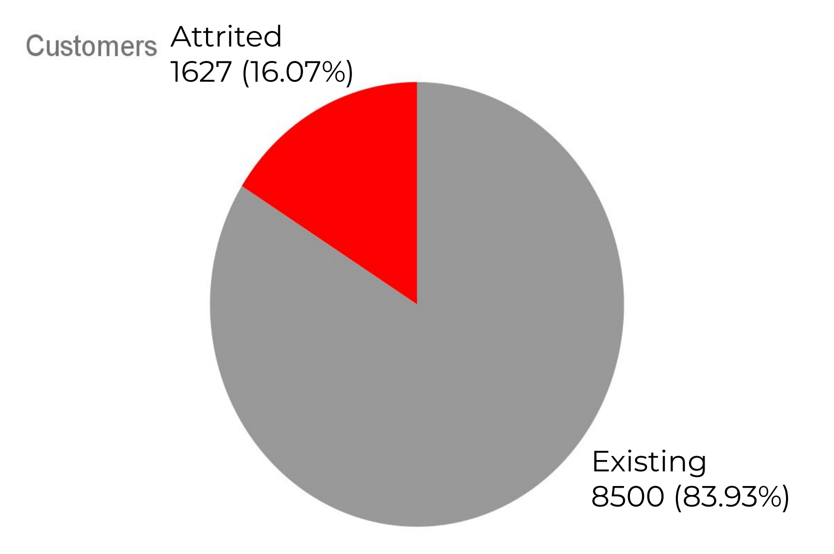
Client Number
Customer Age
Gender
Dependent Count
Education Level
Marital Status
Income Category

Card

Client Number
Card Category
Months on Book
Total Relationship Count
No. of months inactive
No. of contacts
Credit Limit
Total Revolving Balance
Average Open to Buy Credit Line
Transaction Amount Change (Q4 over Q1)
Total Transaction Amount
Total Transaction Count
Transaction Count Change (Q4 over Q1)

SQL

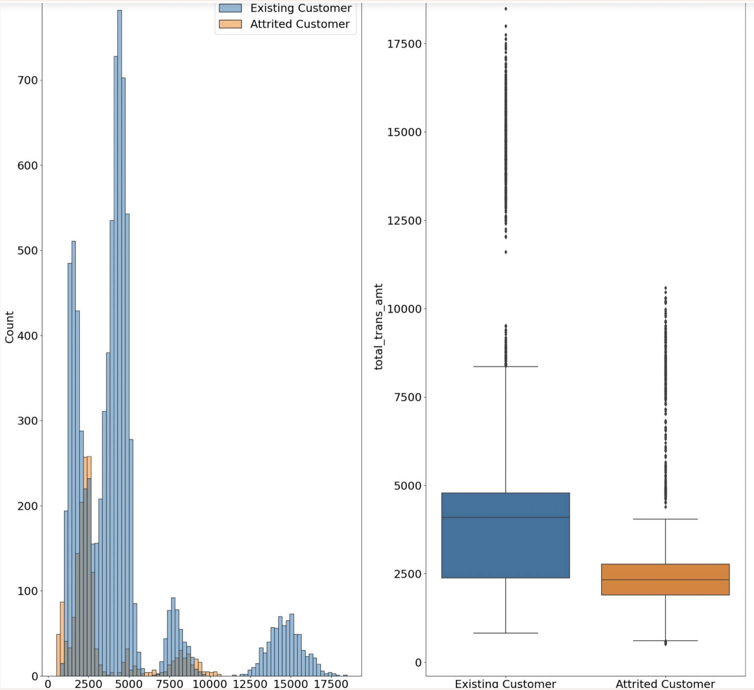
- 10127 rows of data
- Merge/Join the 3 datasets into a combined dataset



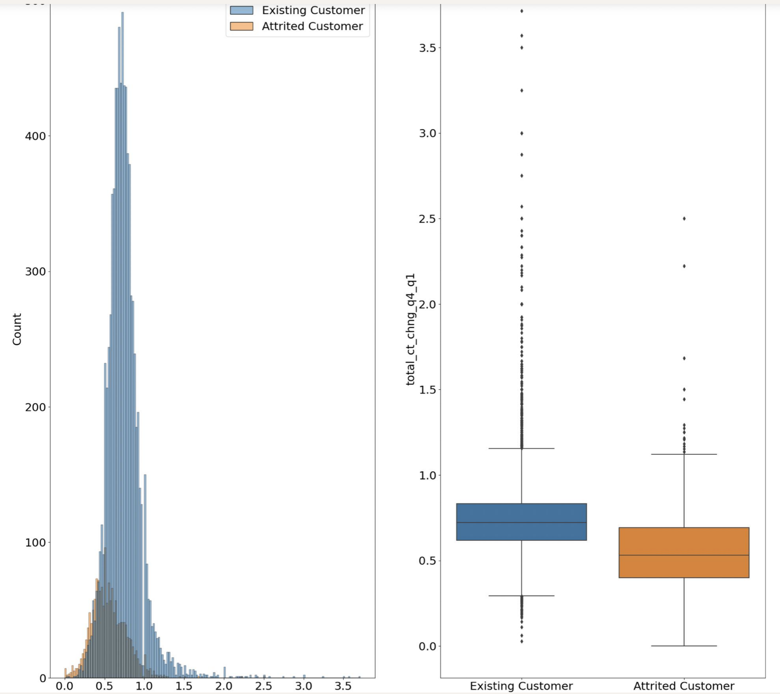
Card Category	Average Credit Limit	Average Total Transaction Amount
Blue	$ 7363	$ 4225
Silver	$ 25277	$ 6590
Gold	$ 28416	$ 7685
Platinum	$ 30283	$ 8999

EDA (Numerical Variables)

Total Transaction Amount



Total Transaction Count Change (Q4 over Q1)



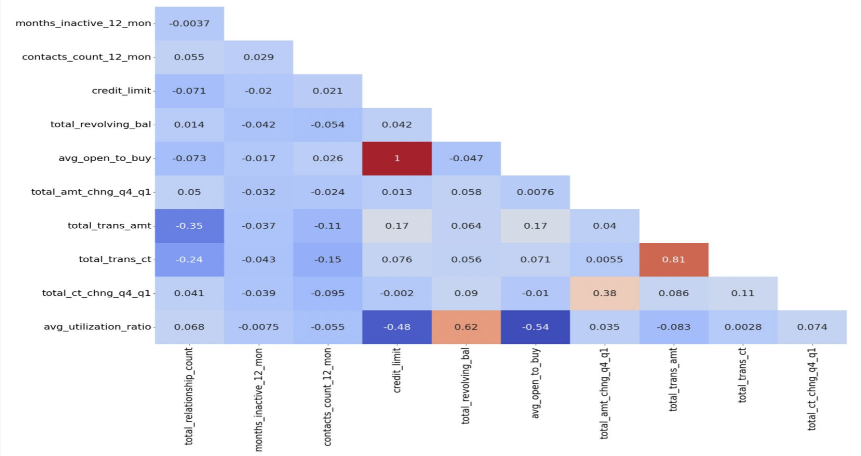
EDA (Numerical Variables)

Dependent Count

Statistical tests

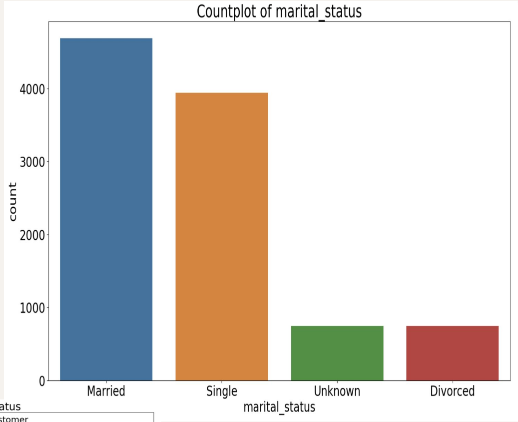
```
customer_age has non-normal distribution: running KS test. result: not statistically significant
dependent_count has non-normal distribution: running KS test. result: not statistically significant
months_on_book has non-normal distribution: running KS test. result: not statistically significant
total_relationship_count has non-normal distribution: running KS test. result: statistically significant
months_inactive_12_mon has non-normal distribution: running KS test. result: statistically significant
contacts_count_12_mon has normal distribution: running Anova test. result: statistically significant
credit_limit has non-normal distribution: running KS test. result: statistically significant
total_revolving_bal has non-normal distribution: running KS test. result: statistically significant
avg_open_to_buy has non-normal distribution: running KS test. result: statistically significant
total_amt_chng_q4_q1 has non-normal distribution: running KS test. result: statistically significant
total_trans_amt has non-normal distribution: running KS test. result: statistically significant
total_trans_ct has non-normal distribution: running KS test. result: statistically significant
total_ct_chng_q4_q1 has non-normal distribution: running KS test. result: statistically significant
avg_utilization_ratio has non-normal distribution: running KS test. result: statistically significant
```

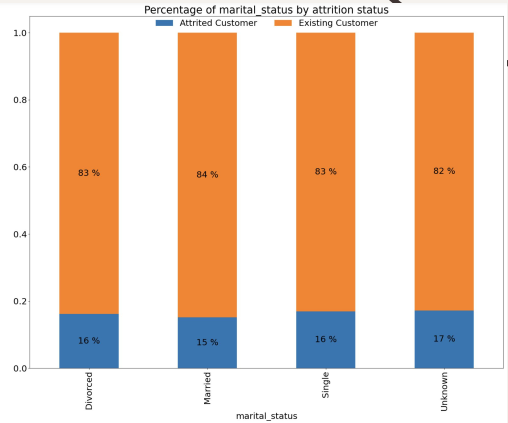
Correlation Heatmap

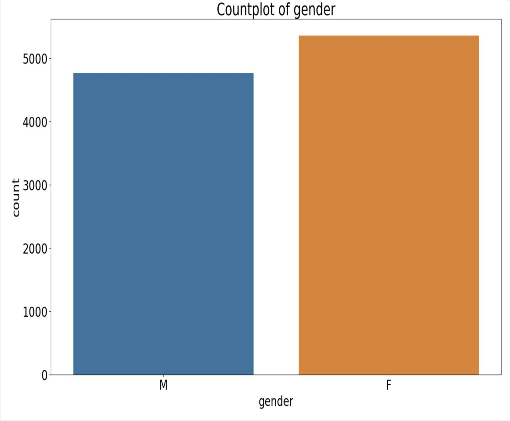


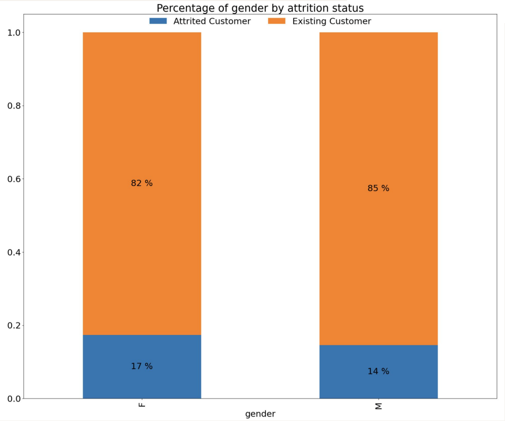
EDA (Categorical Variables)

Marital Status









Gender

EDA (Categorical Variables)

```
education_level     0.560273
card_category       0.320692
marital_status      0.253711
income_category     0.115657
gender              0.006368
dtype: float64
```

In the case of classification problems where input variables are also categorical, we can use statistical tests to determine whether the output variable is dependent or independent of the input variables. If independent, then the input variable is a candidate for a feature that may be irrelevant to the problem and removed from the dataset.

The chi-squared statistical hypothesis is an example of a test for independence between categorical variables. We apply the threshold of a p-value below 0.05 to be considered significant.

Preprocessing

Original dataset
Attrition flag - 0 for existing, 1 for attrited
Gender - 0 for Male and 1 for Female

For the other categorical variables, we one-hot encode and converted it to dummy variables.

Train-test split with test size of 0.2 and stratify
Standardization of data

Filtered dataset
Attrition flag - 0 for existing, 1 for attrited
Gender - 0 for Male and 1 for Female

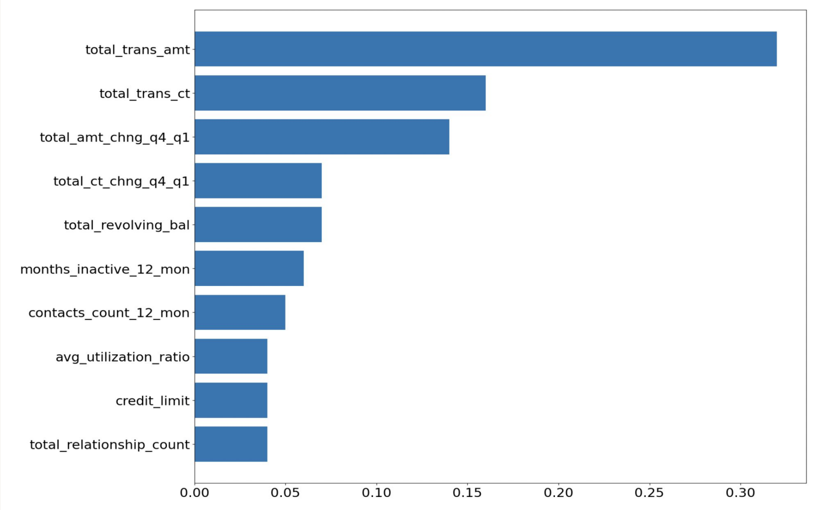
Train-test split with test size of 0.2 and stratify
Standardization of data

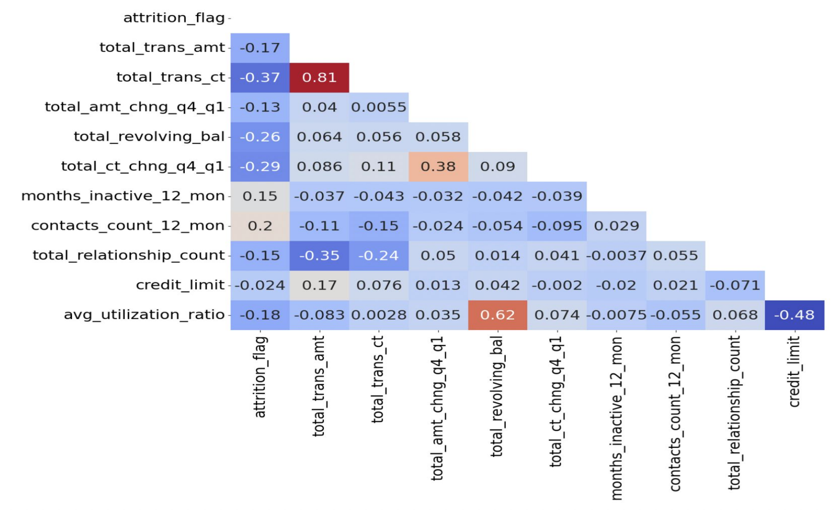
Modeling and Evaluation

Model	Dataset	Accuracy (training set)	Accuracy (test set)	F1 score (training set)	F1 score (test set)	ROC AUC
K-nearest neighbor	Original	0.897	0.876	0.554	0.419	0.856
K-nearest neighbor	Filtered	0.941	0.926	0.795	0.739	0.944
Bagging Classifier	Original	1.0	0.906	1.0	0.586	0.983
Bagging Classifier	Filtered	1.0	0.955	1.0	0.854	0.985
Random Forest	Original	0.999	0.956	0.997	0.847	0.985
Random Forest	Filtered	0.999	0.956	0.998	0.856	0.985
Ada Boost Classifier	Original	0.972	0.957	0.912	0.861	0.984
Ada Boost Classifier	Filtered	0.964	0.956	0.887	0.856	0.983
Support Vector Model	Original	0.981	0.915	0.938	0.709	0.940
Support Vector Model	Filtered	0.965	0.940	0.886	0.799	0.965

Feature Importance

Based on the selected model (Ada Boost Classifier with filtered dataset), we obtain the top 10 important (significant) features that are being utilized by the model in predicting the customer churn. We can observe the feature importance for our best selected model.





Feature Importance (continued)

Number of features (Threshold)	Accuracy (training set)	Accuracy (test set)	F1 score (training set)	F1 score (test set)	ROC AUC
6 (>0.05)	0.957	0.942	0.861	0.808	0.980
7 (>0.04)	0.957	0.947	0.862	0.828	0.982
10 (>0.01)	0.964	0.955	0.886	0.855	0.983
11 (no threshold)	0.964	0.956	0.887	0.856	0.983

Conclusion

A classification model has been developed to predict the churn for the bank credit card customers.

With our classification model, the bank will be able to predict which are the customers which are likely to attrite and the bank will be able to take action and reach out to those customers to discuss on what can be done so as to be able to retain those customers with the Bank.

Lastly, we have deployed our classification model onto streamlit app to allow bank staff to utilize to obtain predictions of the attrition status for a particular customer or a batch of customers.

Model deployment on streamlit app