Thera Bank

Credit Card Users Churn Prediction

AIML course - Bagging, Boosting, and Hyperparameter Tuning

Azin Faghihi
Role: Data Scientist
February 2025

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model performance summary for hyperparameter tuning.
- Appendix

Executive Summary

- Our modeling parameter, **Attrition_Flag** is an **imbalance class**. Hence, we will try oversampling and undersampling to address this issue.
- By conducting ***exploratory data analysis*** (univariate and bivariate) we explore relationships between all the variables (categorical & numerical)
    - We conclude that Attrition_Flag is affected by income, total transaction count, credit limit, ...
    - Due to the nature of the problem at hand, we chose **Recall** as the score to maximize.
- We conduct **hyperparameter tuning** to select a model that does not overfit and also has the highest Recall score
- The ***important features*** observed in our trained models match those of our observation in exploratory data analysis.

Business Problem Overview and Solution Approach

- ***Problem Overview:*** Thera Bank has experienced a decline in the number of credit card users. Since losing credit card customers leads to financial losses, the bank aims to identify the reasons behind customer attrition and address these issues.
- ***Objective:*** Based on the gathered data, we aim to predict customer attrition, enabling the bank to take proactive measures to retain customers.
- ***Solution Approach:*** Using exploratory data analysis and machine learning techniques (including multiple classification models, boosting, hyperparameter tuning, and sampling), we will identify the key factors that influence whether a customer renounces their credit card or retains it.

Data

- There are 10,127 (~10K) *rows* and 24 *columns* in the dataset.
- The *memory usage* is approximately 1899 KB.
- There are 3,380 null values, 15% in Education Level, 11% in Income Category, and 7% in Marital_Status
- There are no duplicated rows.

Memory Usage	1898.9 KB

#	
Rows	10127
Columns	24
Null Values	3380
Duplicated Rows	0

Data Dictionary

Column	Data Type	Description
CLIENTNUM	int64	Client number. Unique identifier for the customer holding the account
Attrition_Flag	object	Internal event (customer activity) variable - if the account is closed then "Attrited Customer" else "Existing Customer"
Customer_Age	int64	Age in Years
Gender	object	Gender of the account holder
Dependent_count	int64	Number of dependents
Education_Level	object	Educational Qualification of the account holder - Graduate, High School, Unknown, Uneducated, College(refers to college student), Post-Graduate, Doctorate
Marital_Status	object	Marital Status of the account holder
Income_Category	object	Annual Income Category of the account holder
Card_Category	object	Type of Card
Months_on_book	int64	Period of relationship with the bank (in months)
Total_Relationship_Count	int64	Total no. of products held by the customer
Months_Inactive_12_mon	int64	No. of months inactive in the last 12 months
Contacts_Count_12_mon	int64	No. of Contacts in the last 12 months
Credit_Limit	float64	Credit Limit on the Credit Card
Total_Revolving_Bal	int64	Total Revolving Balance on the Credit Card
Avg_Open_To_Buy	float64	Open to Buy Credit Line (Average of last 12 months)
Total_Amt_Chng_Q4_Q1	float64	Change in Transaction Amount (Q4 over Q1)
Total_Trans_Amt	int64	Total Transaction Amount (Last 12 months)
Total_Trans_Ct	int64	Total Transaction Count (Last 12 months)
Total_Ct_Chng_Q4_Q1	float64	Change in Transaction Count (Q4 over Q1)
Avg_Utilization_Ratio	float64	Average Card Utilization Ratio

Data

- The following tables show the summary information of our variables.
  - ***Categorical***: unique counts, most common value and its corresponding frequency
  - ***Numerical***: mean, median, standard deviation, minimum, maximum, outlier counts, ...

Object/Categorical Column	unique	top	freq
Attrition_Flag	2	Existing	8500
Gender	2	F	5358
Dependent_count	6	3	2732
Education_Level	6	Graduate	4647
Marital_Status	3	Married	5436
Income_Category	5	Less than $40K	4673
Card_Category	4	Blue	9436
Total_Relationship_Count	6	3	2305
Months_Inactive_12_mon	7	3	3846
Contacts_Count_12_mon	7	3	3380
Education_Level (missing)	6	Graduate	3128
Marital_Status (missing)	3	Married	4687
Income_Category (missing)	5	Less than $40K	3561

Numerical Column	mean	std	min	25%	50%	75%	max	IQR	# Outliers (Upper)	# Outliers (Lower)	# Outliers	Outliers %
Customer_Age	46.3	8.0	26.0	41.000	46.000	52.000	73.000	11.000	2	0	2	0.0
Months_on_book	35.9	8.0	13.0	31.000	36.000	40.000	56.000	9.000	198	188	386	3.8
Credit_Limit	8632.0	9088.8	1438.3	2555.000	4549.000	11067.500	34516.000	8512.500	984	0	984	9.7
Total_Revolving_Bal	1162.8	815.0	0.0	359.000	1276.000	1784.000	2517.000	1425.000	0	0	0	0.0
Avg_Open_To_Buy	7469.1	9090.7	3.0	1324.500	3474.000	9859.000	34516.000	8534.500	963	0	963	9.5
Total_Amt_Chng_Q4_Q1	0.8	0.2	0.0	0.631	0.736	0.859	3.397	0.228	350	48	398	3.9
Total_Trans_Amt	4404.1	3397.1	510.0	2155.500	3899.000	4741.000	18484.000	2585.500	896	0	896	8.8
Total_Trans_Ct	64.9	23.5	10.0	45.000	67.000	81.000	139.000	36.000	2	0	2	0.0
Total_Ct_Chng_Q4_Q1	0.7	0.2	0.0	0.582	0.702	0.818	3.714	0.236	300	96	396	3.9
Avg_Utilization_Ratio	0.3	0.3	0.0	0.023	0.176	0.503	0.999	0.480	0	0	0	0.0

EDA Results

Categorical

- The majority (~84%) of the customers are existing credit card customers. (**Attrition_Flag**)
- The most common (~93%) credit card type is Blue. (**Card_Category**)
- Most customers have 3 contacts per year. (**Contacts_Count_12_mon**)
- Most customers have 3 dependents. (**Dependent_count**)
- Most common education value is Graduate followed by High School. (**Education_Level**)
- Number of female customers (53%) is slightly higher than the male customers. (**Gender**)
- The most common income is Less than $40K followed by $40K-60K. (**Income_Category**)
- Most of the customers are married. (**Marital_Status**)
- About 38% of the customers are inactive 3 months in the last year. (**Months_Inactive_12_mon**)
- About 23% of the customers hold 3 bank products. (**Total_Relationship_Count**)

EDA Results

Numerical

- The average open-to-buy credit line is ~7.5K and is highly positively skewed. (**Avg_Open_To_Buy**)
- The average card utilization ratio is ~0.3. (**Avg_Utilization_Ratio**)
- The average credit card limit is ~8.3K and his highly positively skewed. (**Credit_Limit**)
- The average customer age is 46 years old. The customer age is slightly negatively distributed and is likely a normal distribution. (**Customer_Age**)
- The average period of relationship with the bank is ~36 months (3 years). This period is also like normally distributed. (**Months_on_book**)
- The average total change in transaction amount over Q4-Q1 is ~0.76. (**Total_Amt_Chng_Q4_Q1**)
- The average total change in transaction count over Q4-Q1 is ~0.71. (**Total_Ct_Chng_Q4_Q1**)
- The average total transaction amount is $4.4K. (**Total_Trans_Amt**)
- The average total transaction count in the past year is ~65. (**Total_Trans_Ct**)

EDA Results

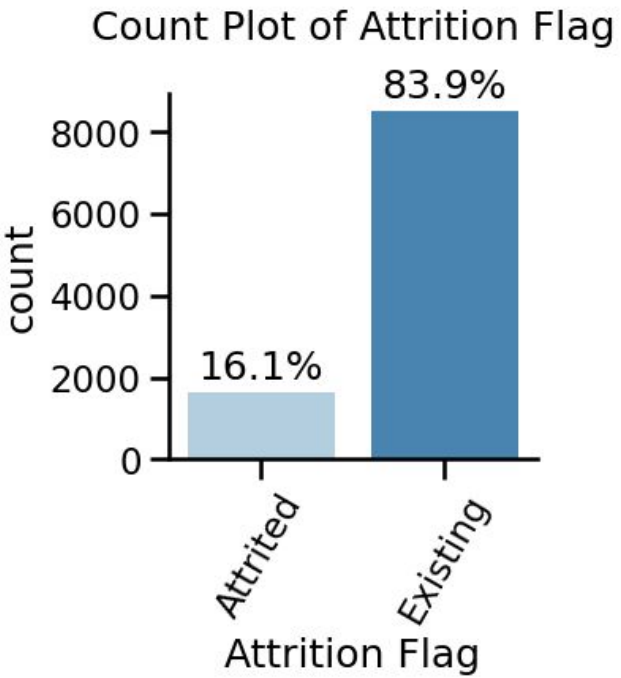
Correlations and Effects

- **Credit_Limit** and **Avg_Open_To_Buy** are highly correlated. This is also shown in their relationship with **Avg_Utilization_Ratio**: *(Avg_Open_To_Buy / Credit_Limit) + Avg_Utilization_Ratio = 1*
- There's also a high correlation between the **Total_Trans_Amt** and **Total_Trans_Ct** and also between **Customer_Age** and **Months_on_book**.
- **Gender, Dependent_count** each seems to have an effect on **Income_Category**. (p-value << 0.05)
- **Contacts_Count_12_mon**, **Months_Inactive_12_mon**, **Total_Relatinship_Count** each seem to have effect on the **Attrition_Flag**. (p-value << 0.05)
- **Income_Category**, **Gender**, **Marital_Status** and **Total_Relationship_Count** each have an effect on **Card_Category.** (p-value < 0.05)
- **Income_Category** has an effect on **Credit_Limit.** (One-Way ANOVA F-test p-value << 0.05)
- **Total_Trans_Ct/Total_Revolving_Bal/Total_Amt_Chng_Q4_Q1/Avg_Utilization_Ratio** has an effect on **Attrition_Flag.** (Two-Sample T-Test p-value << 0.05)

Link to Appendix slide on data background check

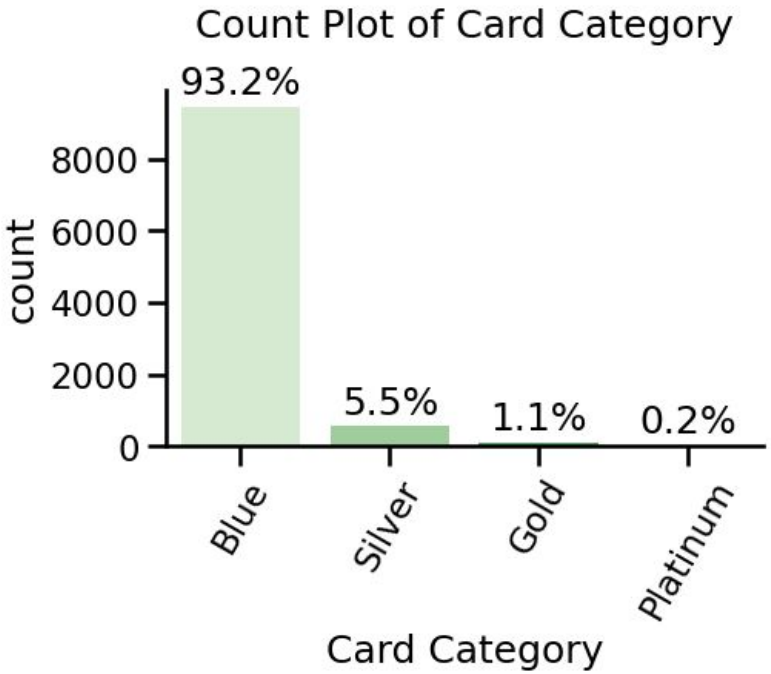
EDA - Univariate - *Attrition_Flag*

- The majority (~84%) of the customers **retain** their credit cards.



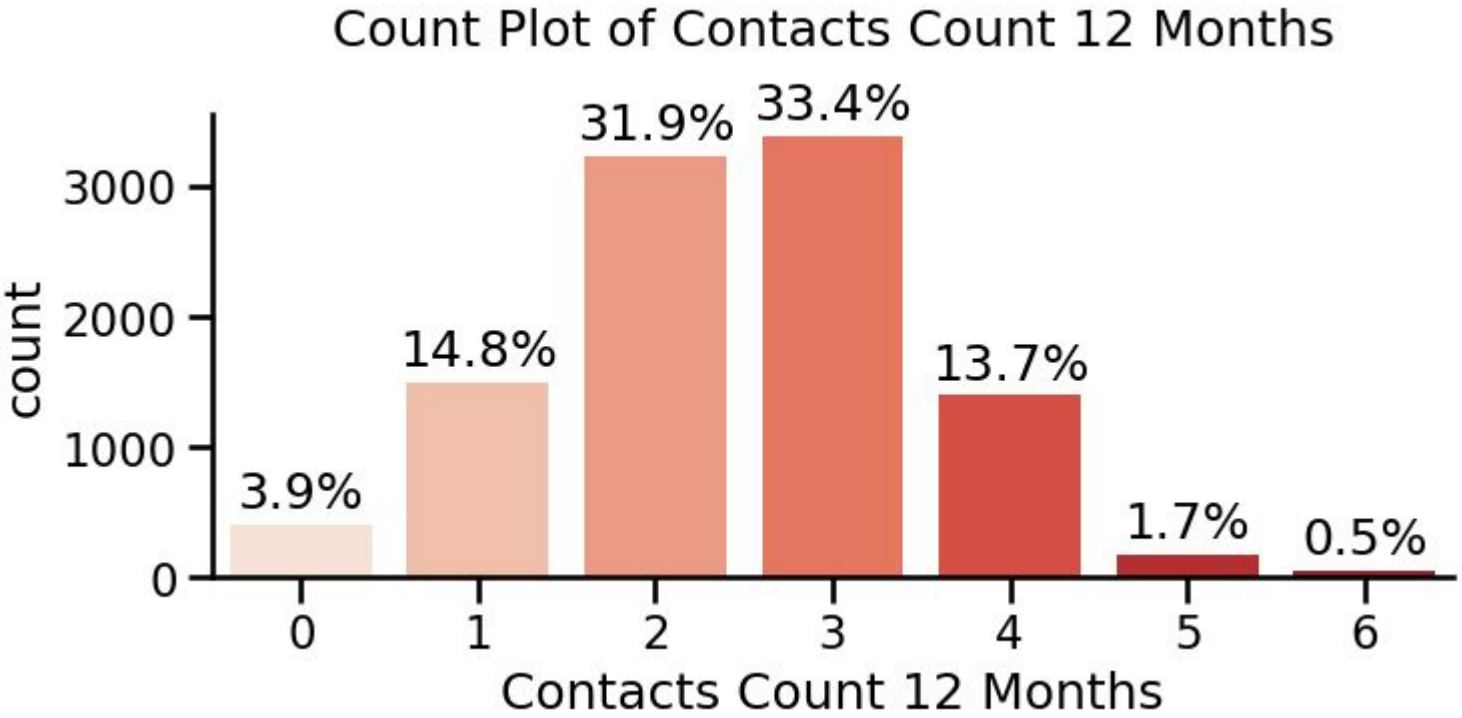
EDA - Univariate - *Card_Category*

- The majority (~93%) of the customers have **Blue** card.



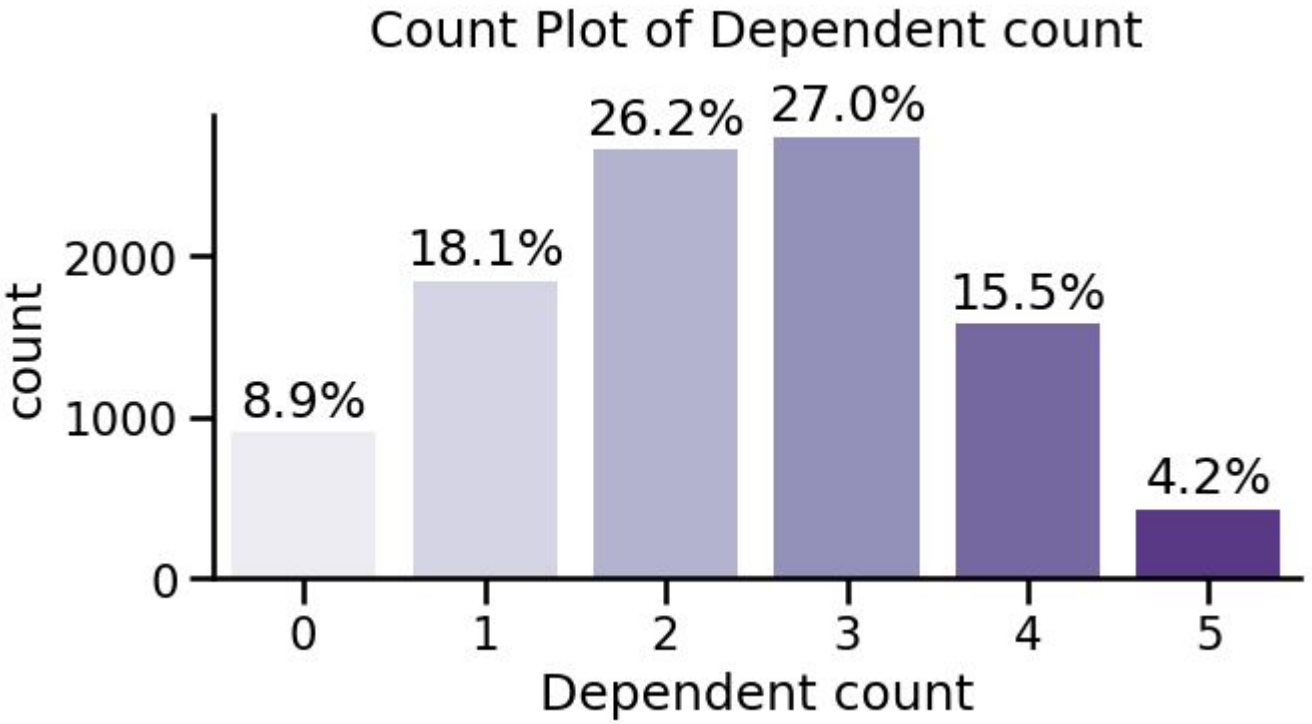
EDA - Univariate - *Contacts_Count_12_mon*

- The majority of the customers had 3 contacts and minority of them had 6 contacts.



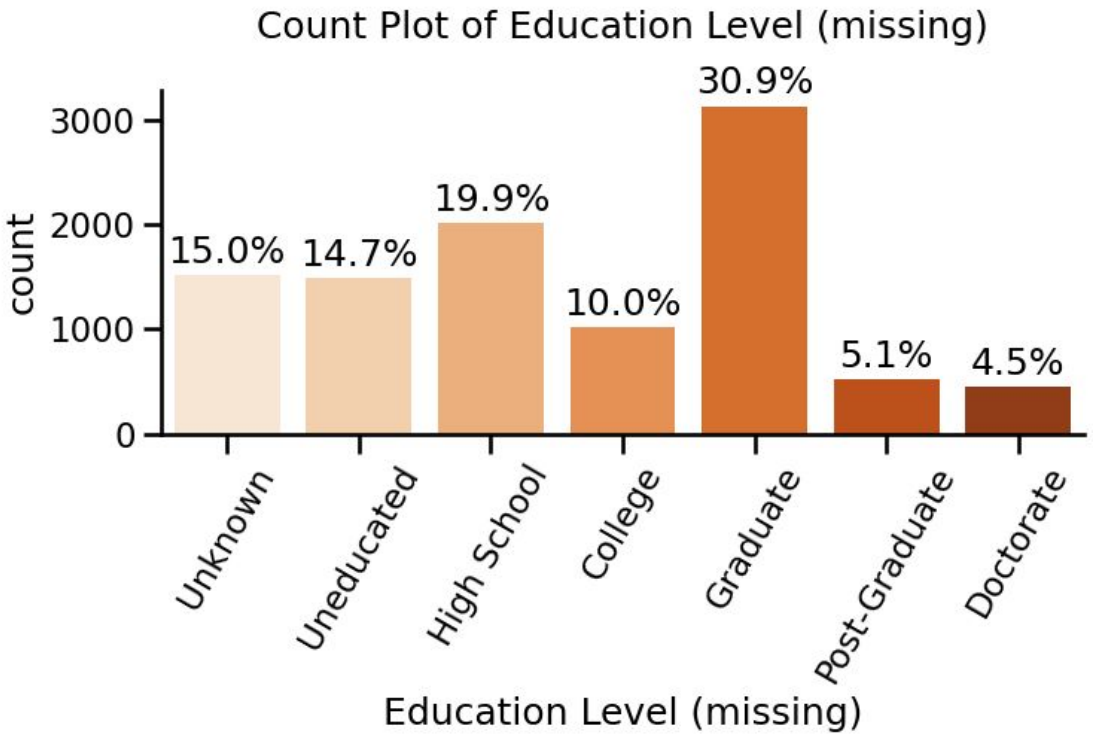
EDA - Univariate - *Dependent_count*

- About 80% of the customers have **less than 4** dependents.



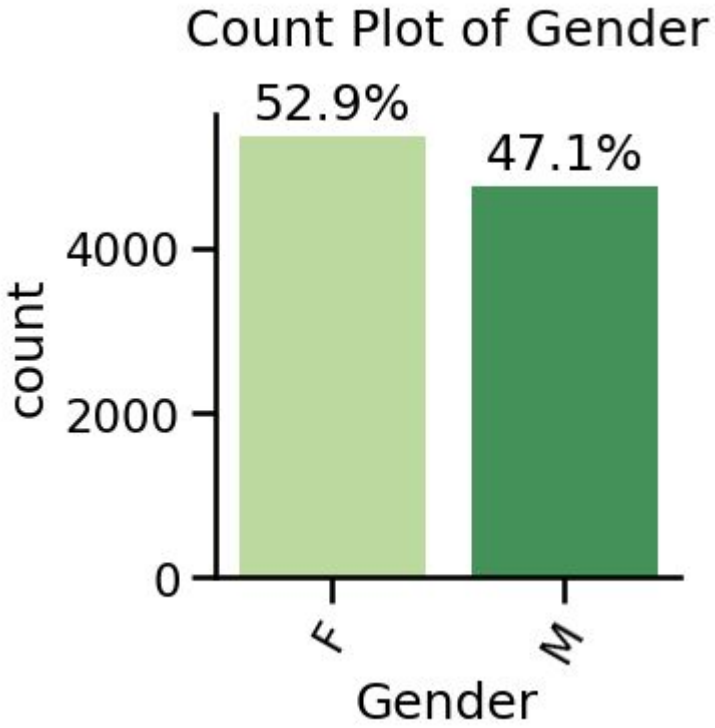
EDA - Univariate - *Education_Level*

- About 31% of the customers have **Graduate** degrees.



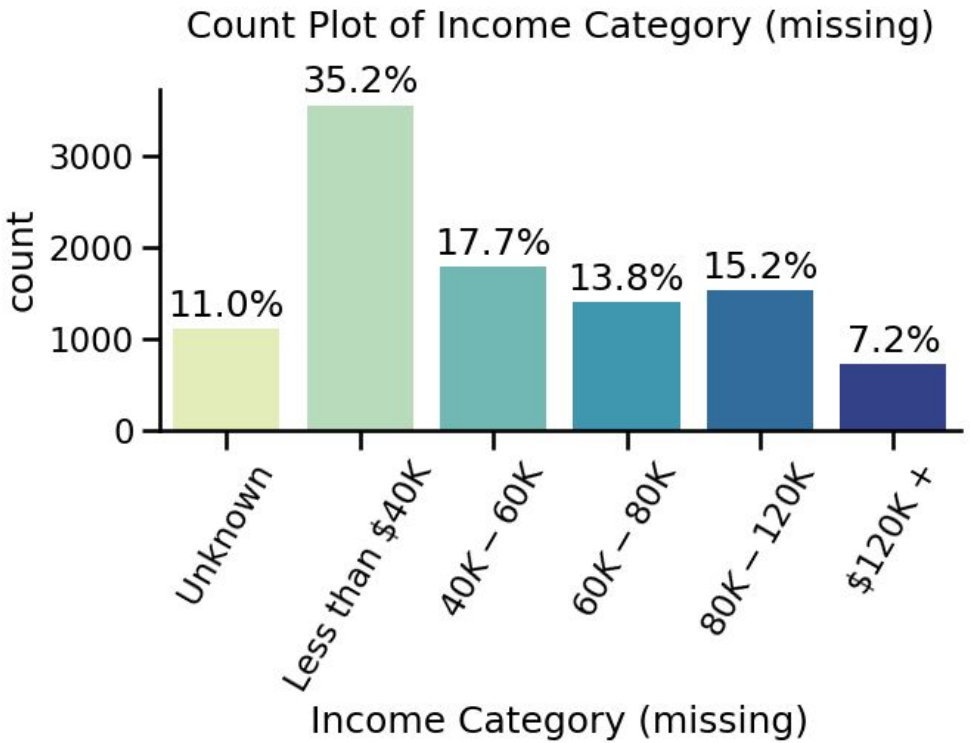
EDA - Univariate - *Gender*

- There are slightly more (53%) **Female** customers.



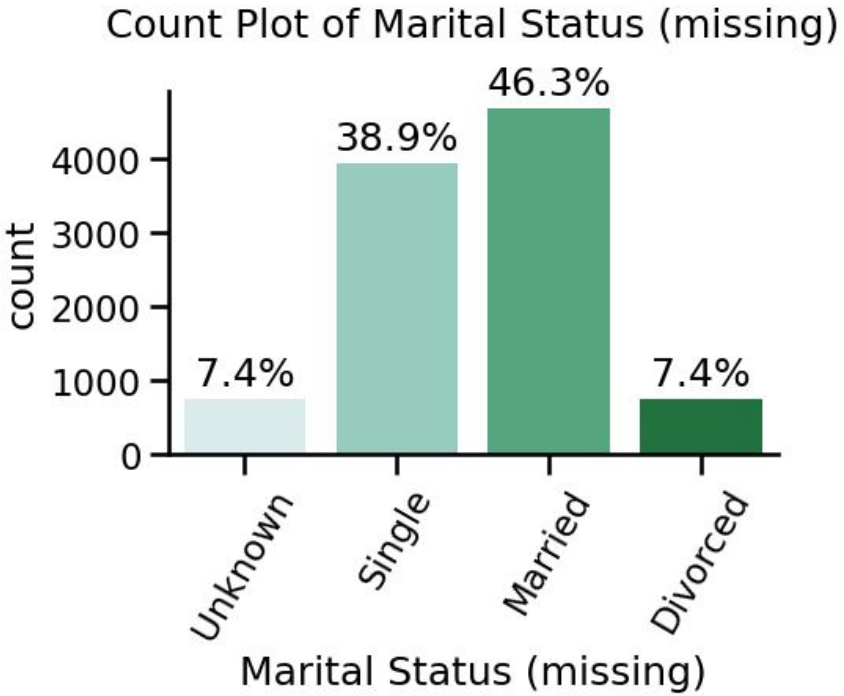
EDA - Univariate - *Income_Category*

- About 35% of the customers earn **less than $40K**.



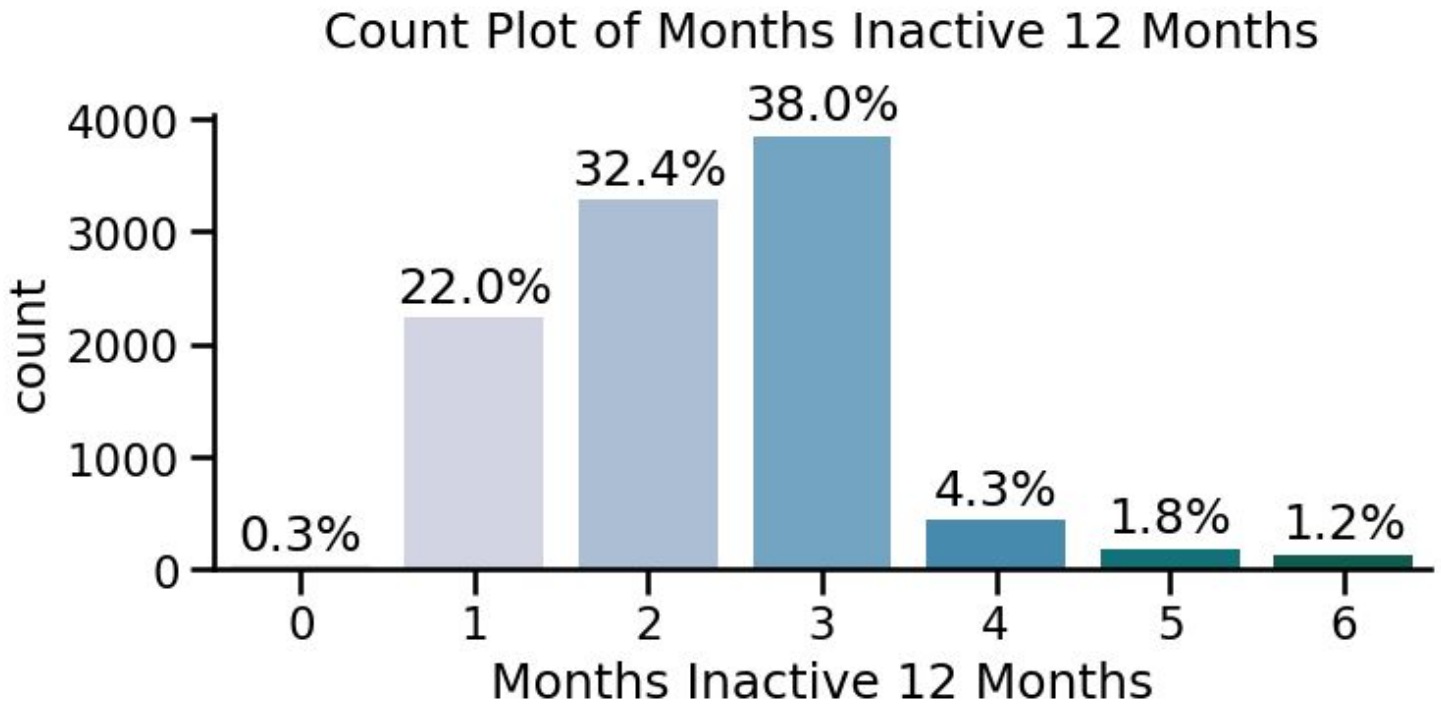
EDA - Univariate - *Marital_Status*

- About 7% of the customers are divorced and the majority (85%) are married or single.



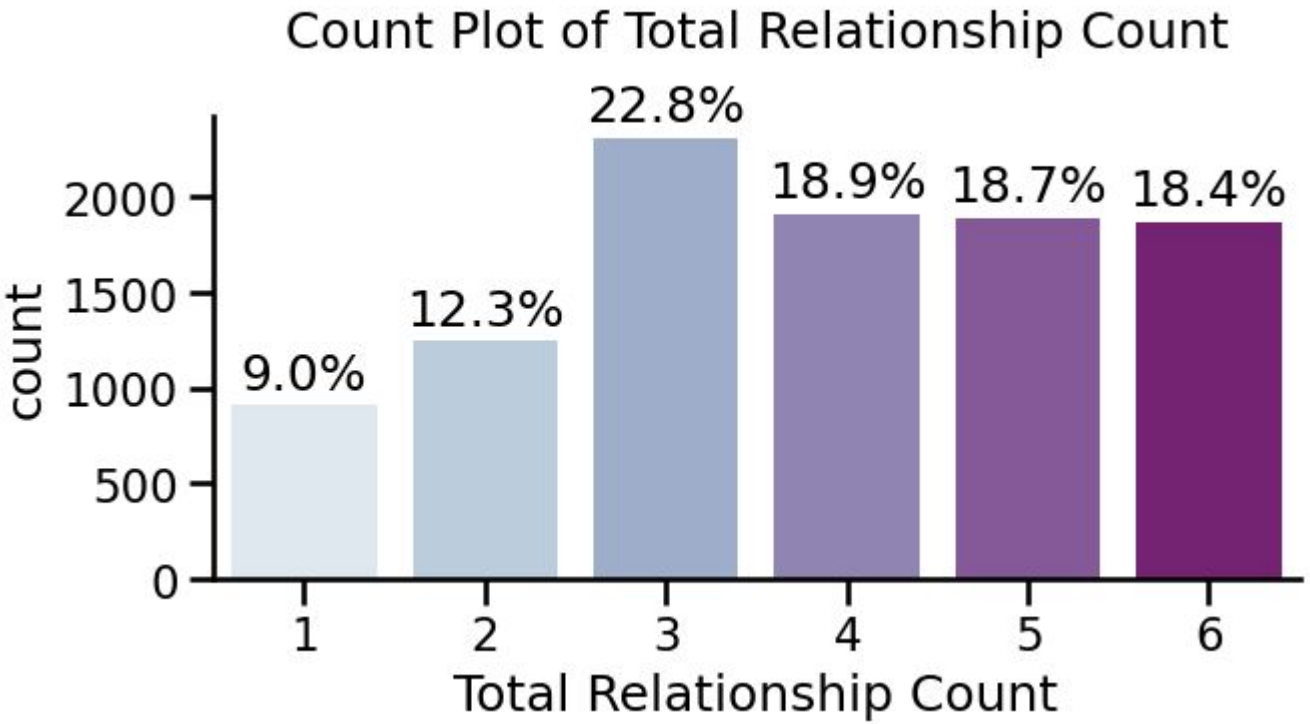
EDA - Univariate - *Months_Inactive_12_mon*

- Customers are rarely inactive for 5 or 6 months.



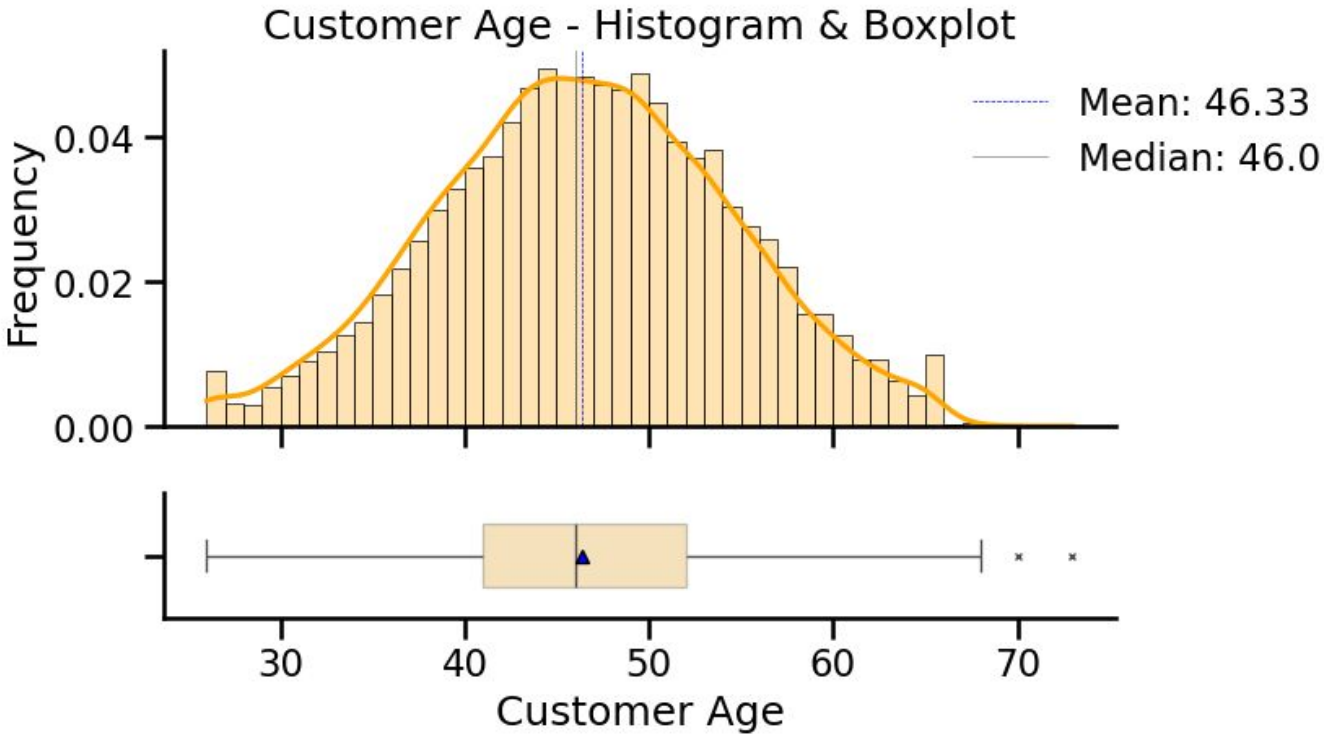
EDA - Univariate - *Total_Relationship_Count*

- Most of the customers are using more than 3 services from the bank.

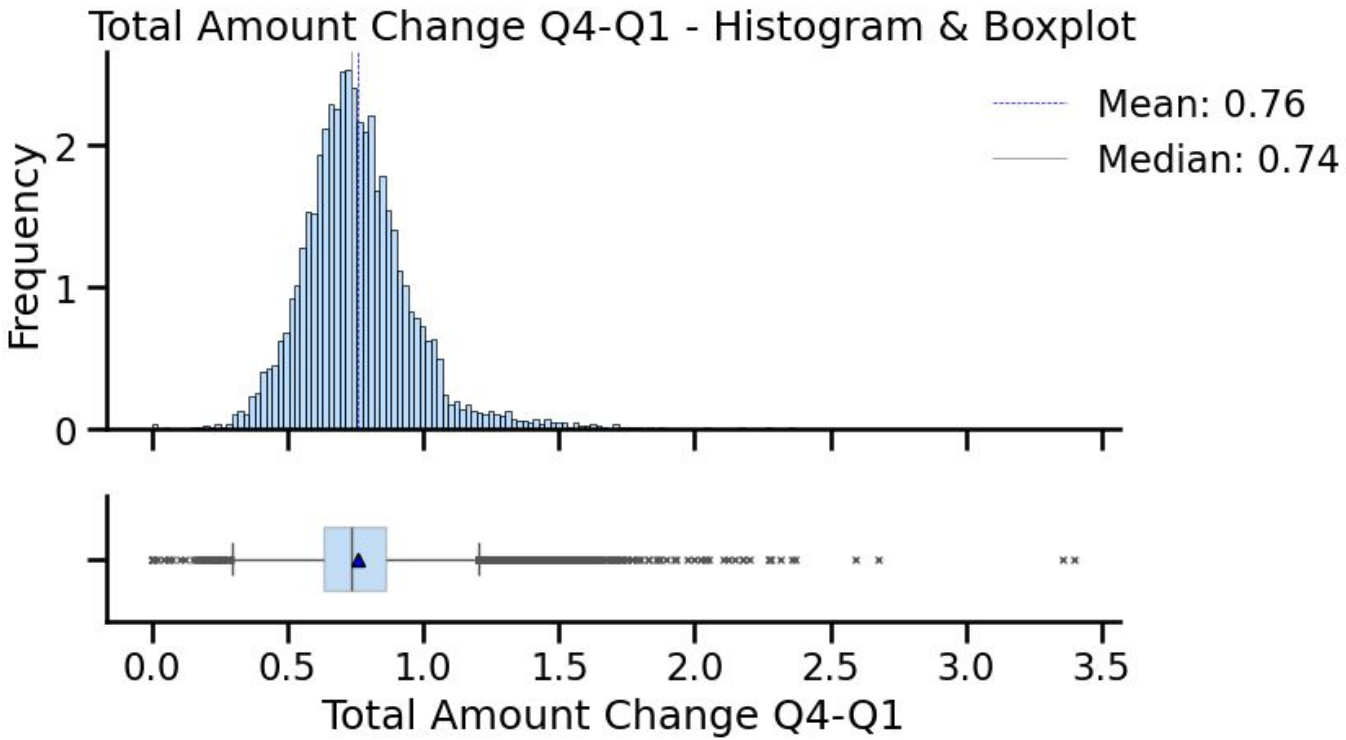


EDA - Univariate - *Customer Age*

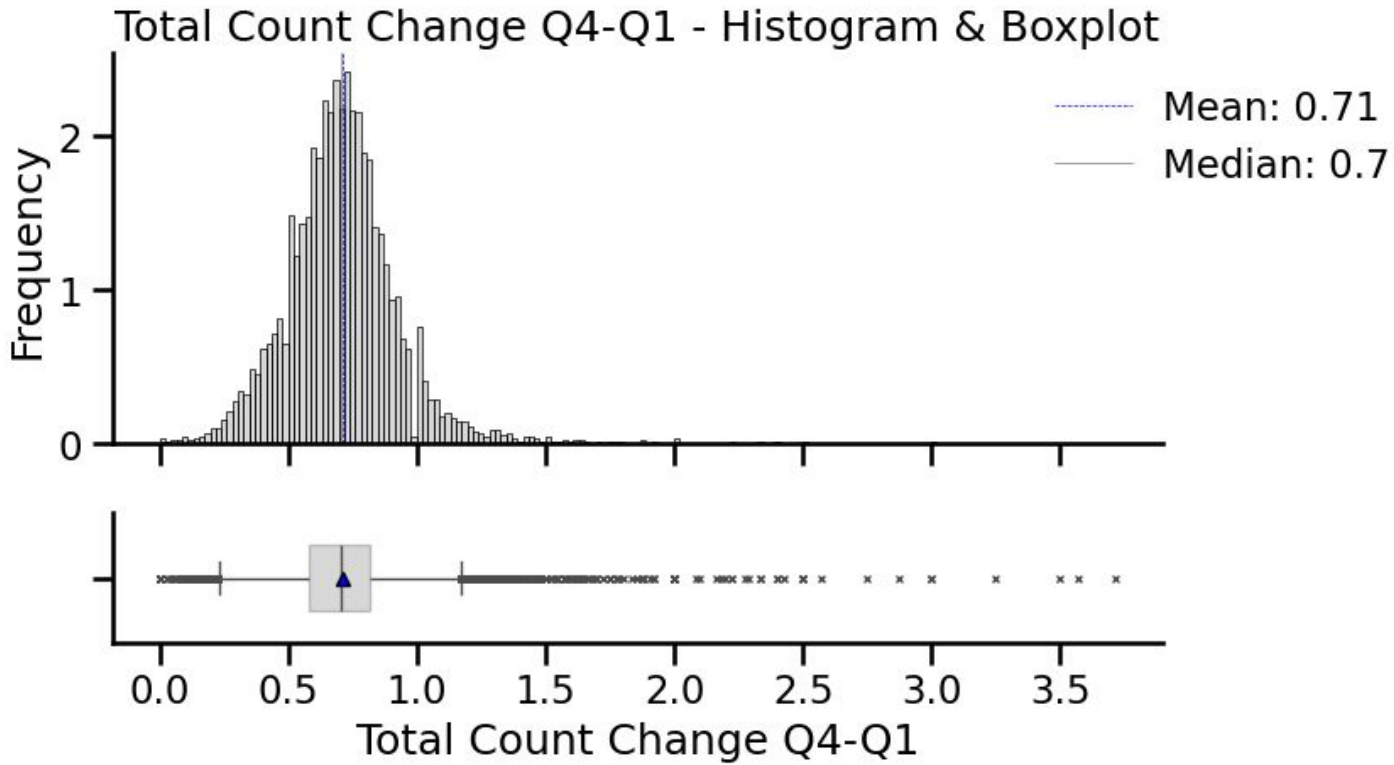
- The average age is 46 years old and the distribution is normal.



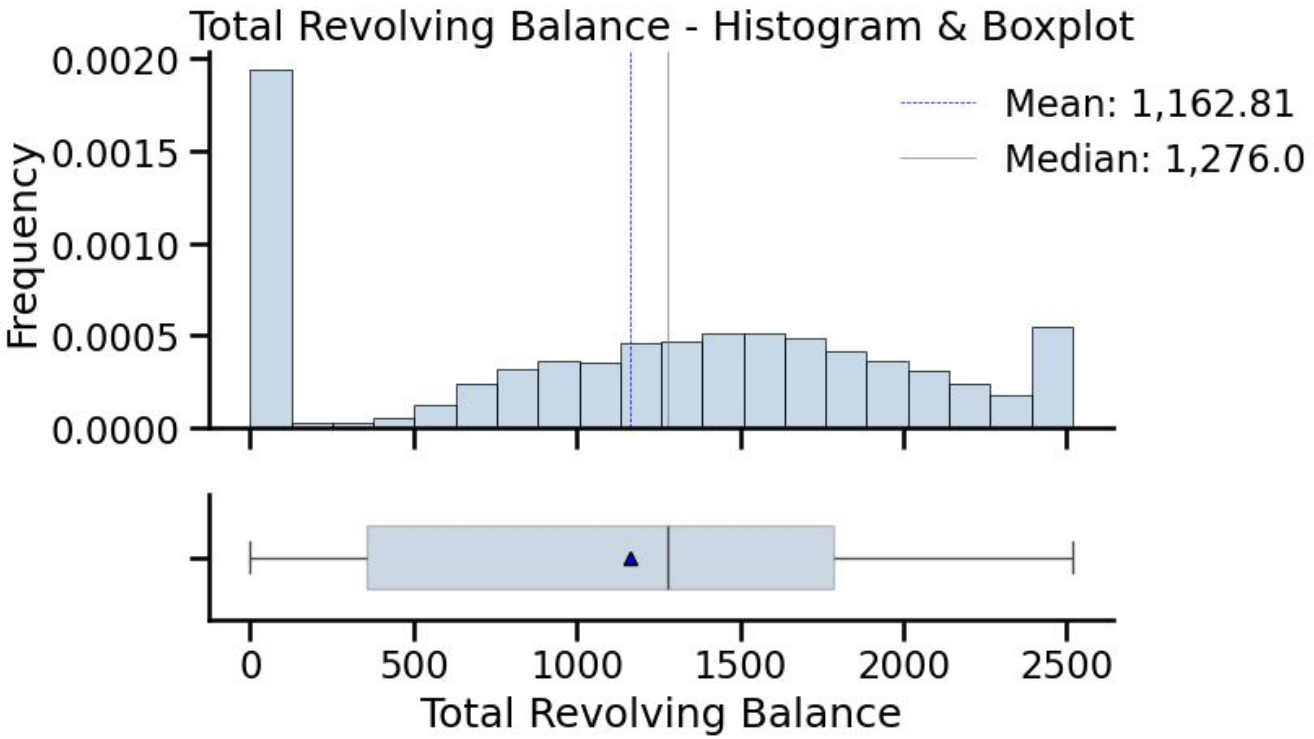
EDA - Univariate - *Total Amount Change Q4-Q1*



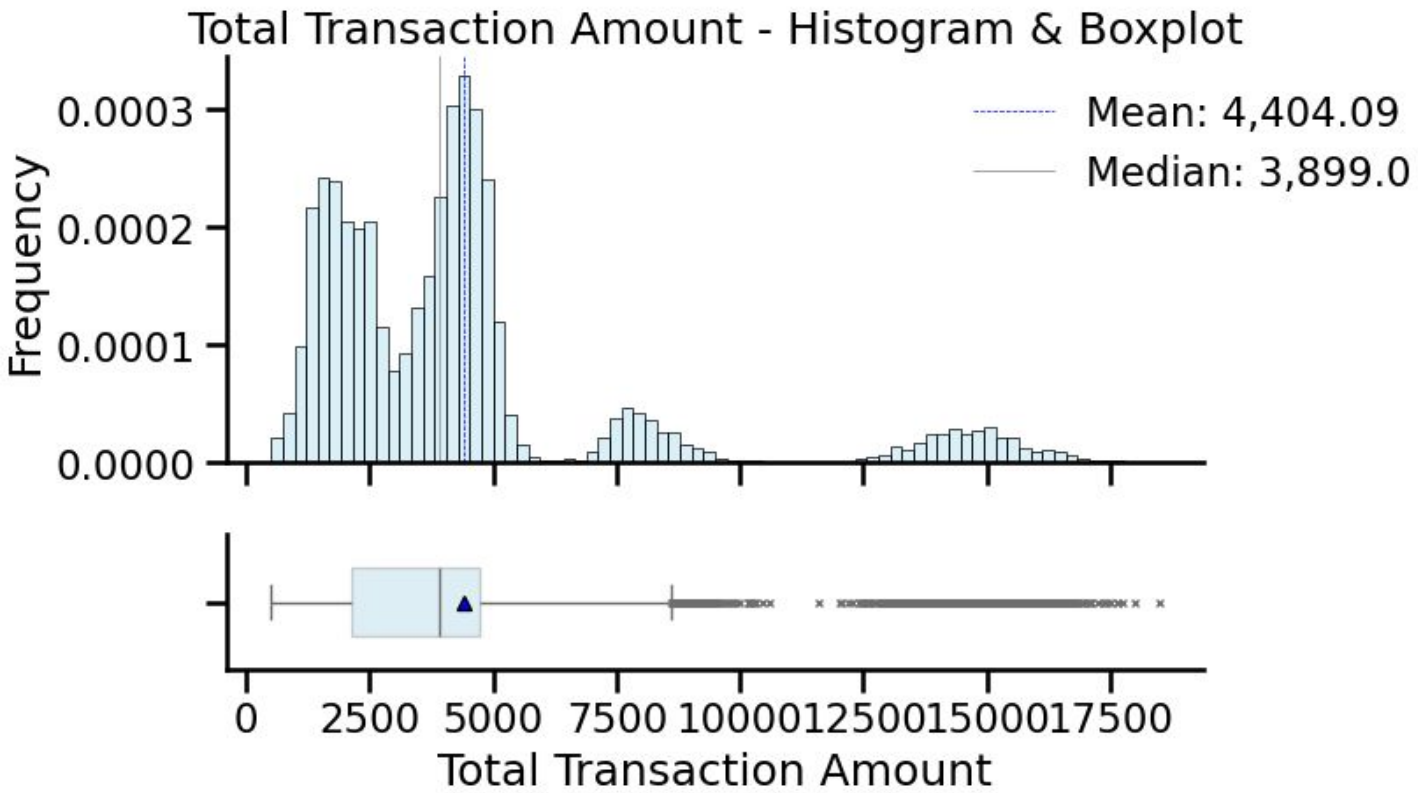
EDA - Univariate - *Total Amount Count Q4-Q1*



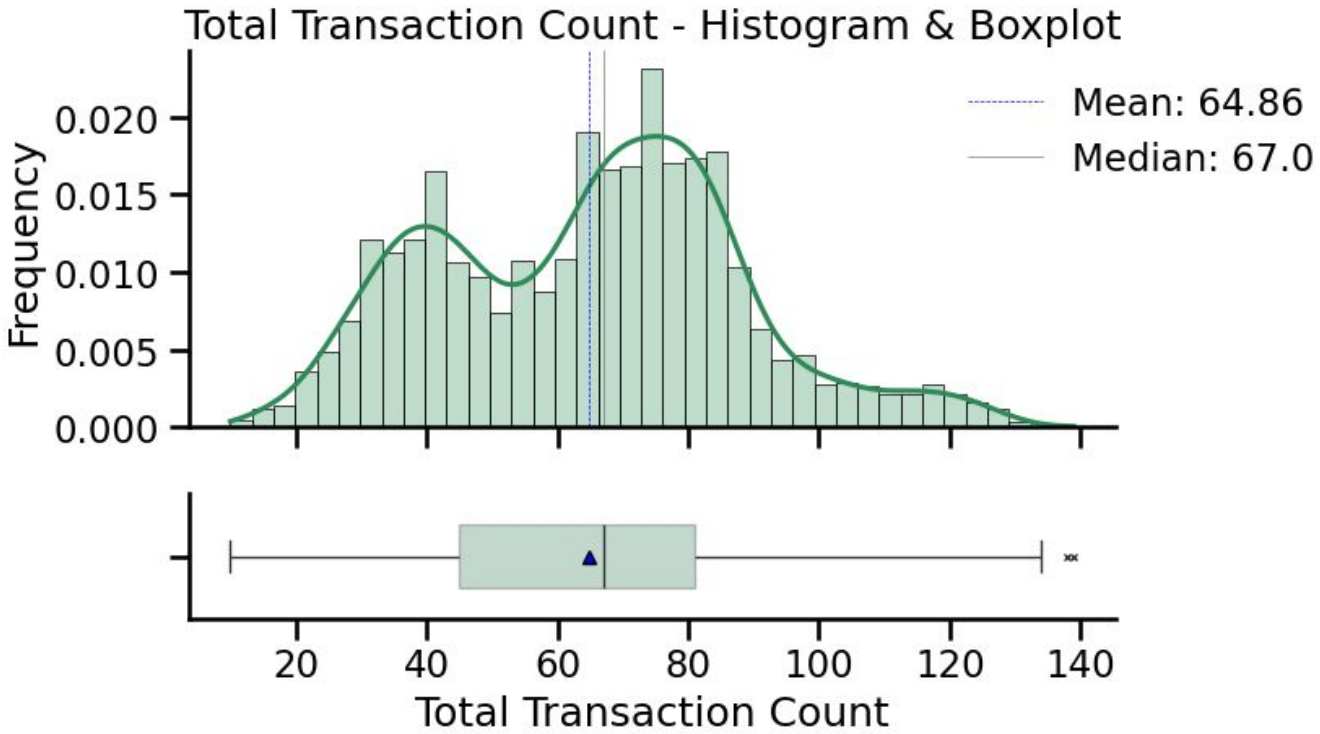
EDA - Univariate - *Revolving Balance*



EDA - Univariate - *Total Transaction Amount*

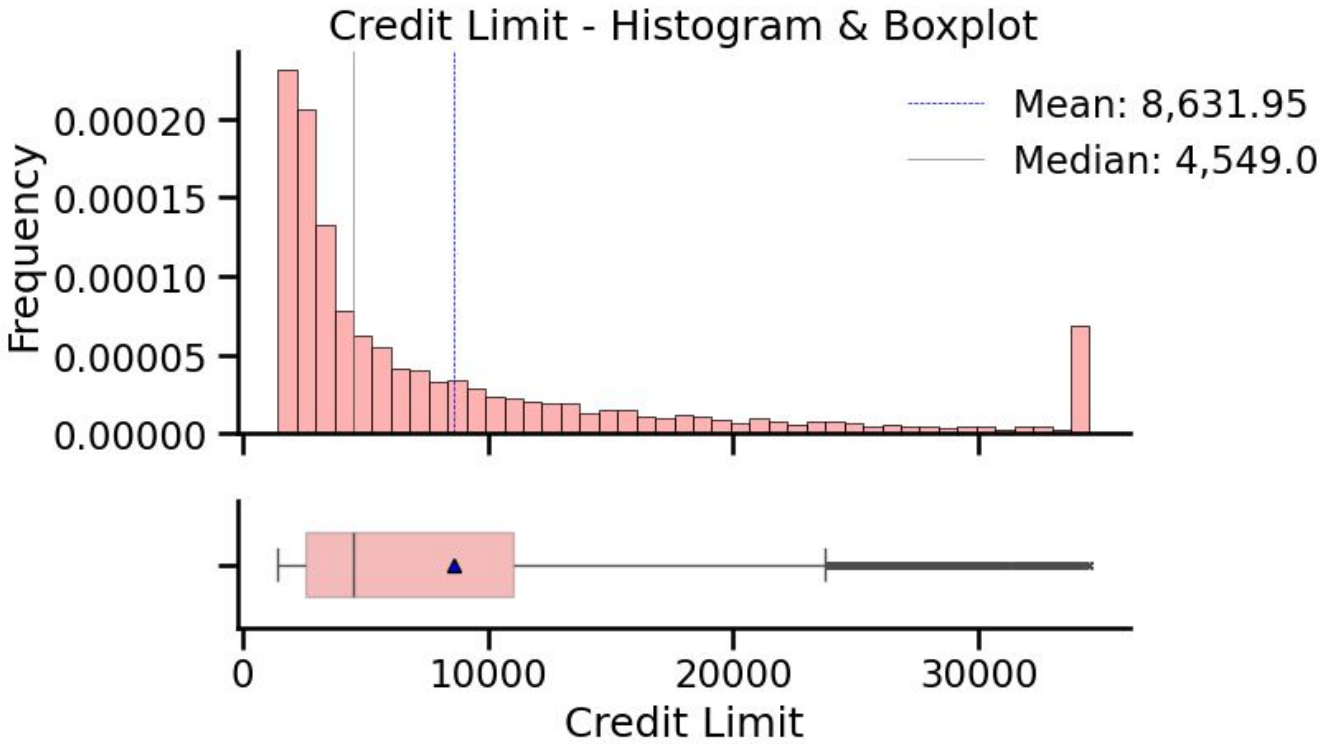


EDA - Univariate - *Total_Trans_Ct*



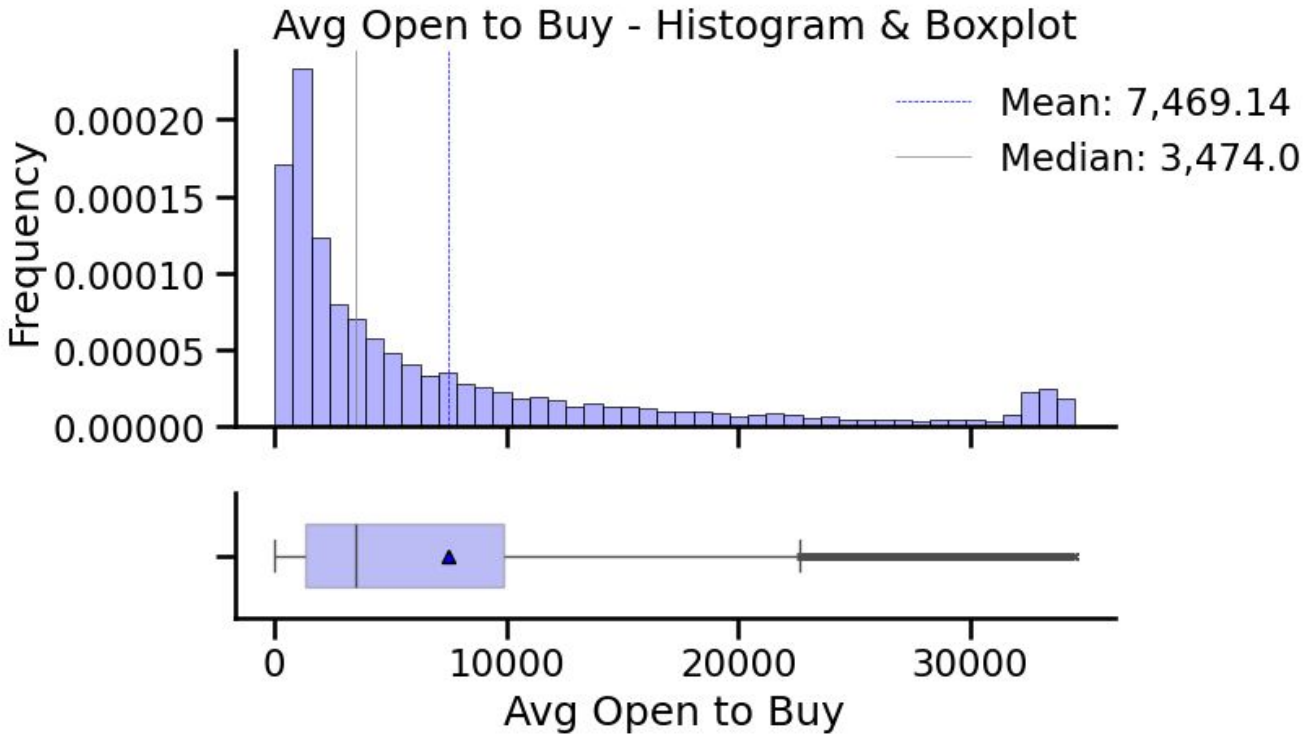
EDA - Univariate - *Credit Limit*

- The average credit limit is $8.6K. The outliers need to be treated.



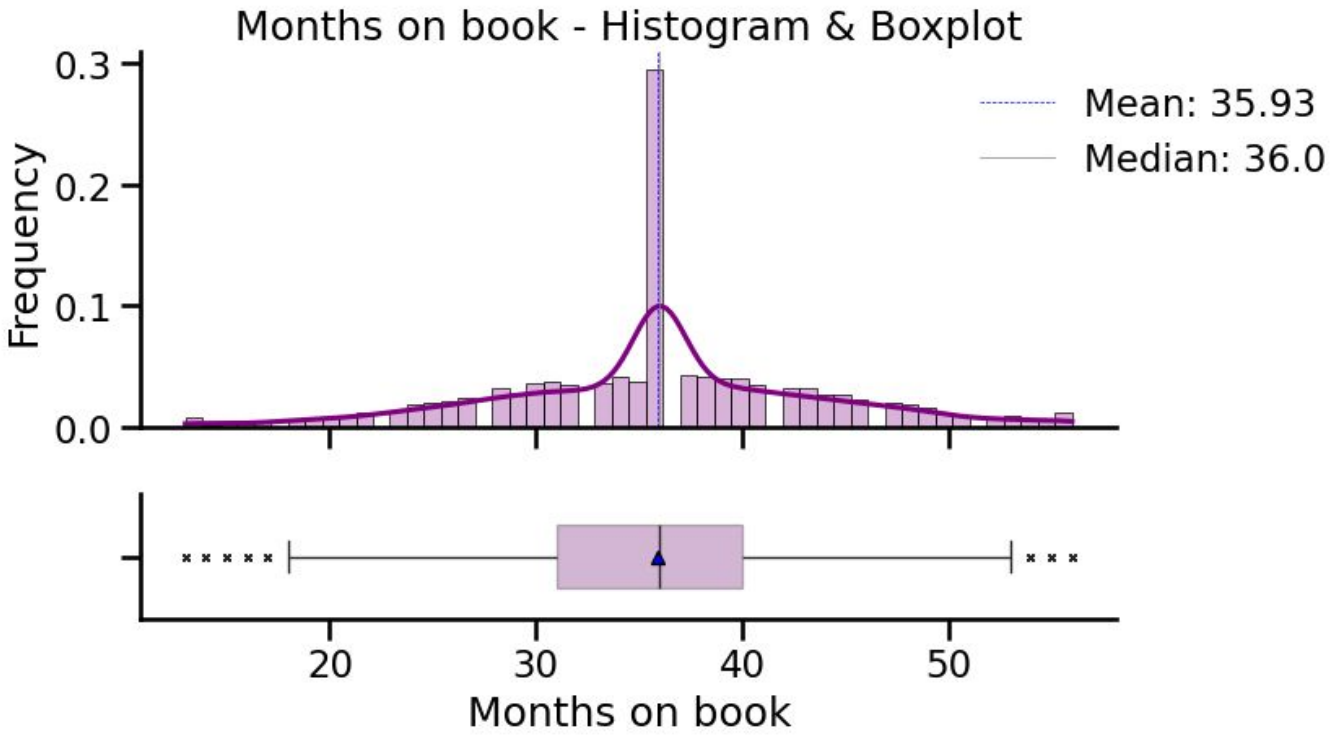
EDA - Univariate - *Avg_Open_to_Buy*

- The average credit limit is $8.4K. The outliers need to be treated.



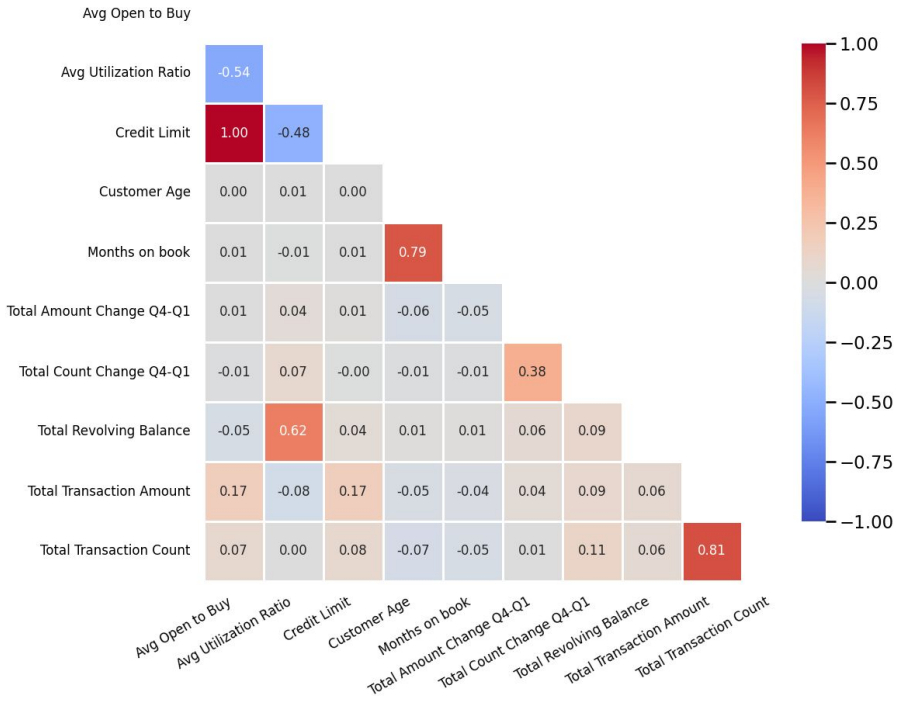
EDA - Univariate - *Months_on_book*

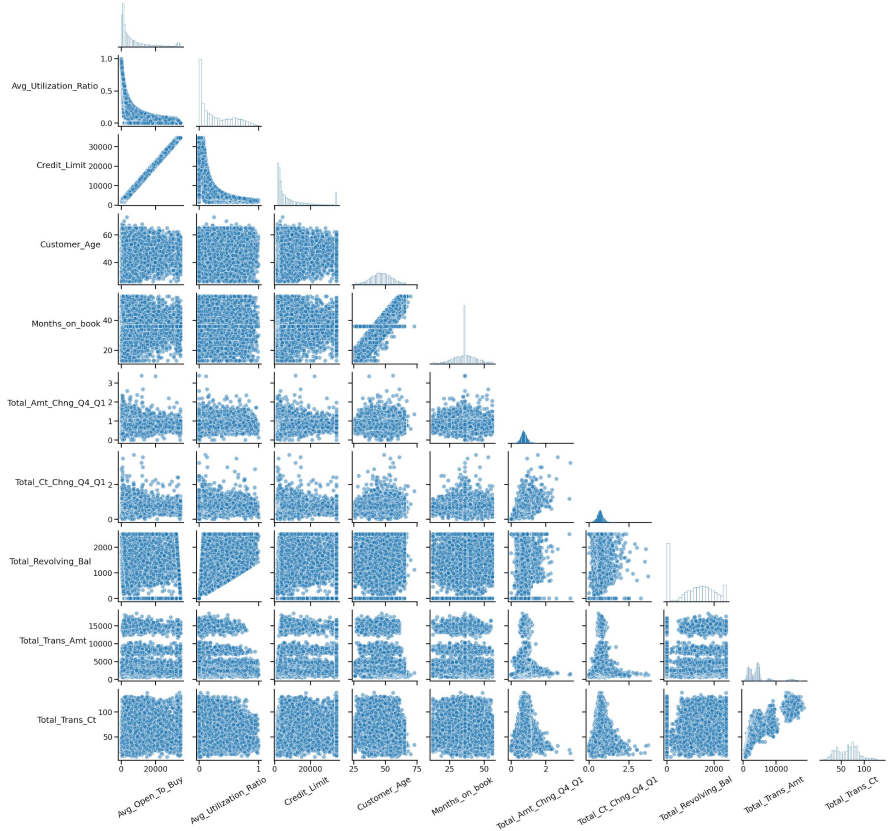
- The median months on book is about 36 months.



EDA - Bivariate - *Numerical Variables*

- Credit Limit and Avg open-to-buy are highly correlated.

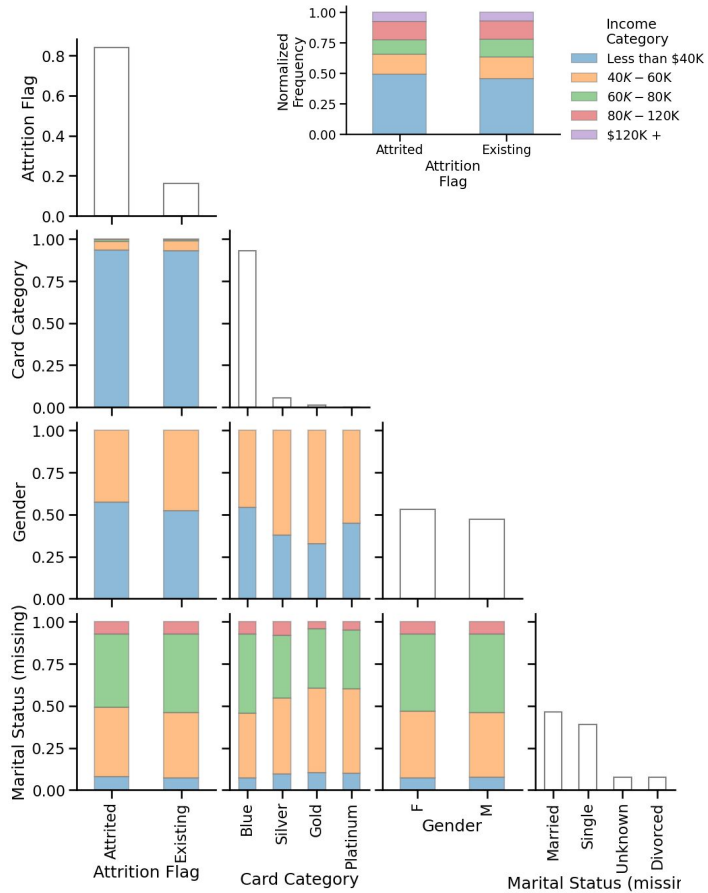




EDA - Bivariate - *Categorical Variables*

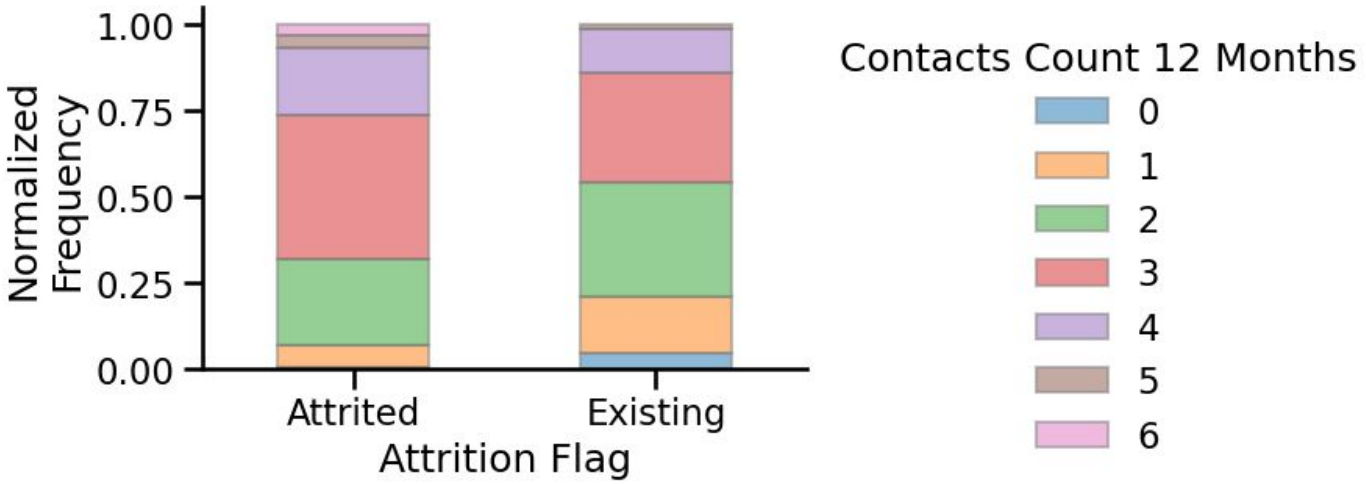
- By conducting χ^2 test of independence we observe that the following variables have effects on each other:

Category 1	Category 2	p-value
Gender	Income_Category (missing)	0.0e+00
Attrition_Flag	Contacts_Count_12_mon	1.8e-123
Attrition_Flag	Months_Inactive_12_mon	1.6e-82
Attrition_Flag	Total_Relationship_Count	2.7e-59
Contacts_Count_12_mon	Total_Relationship_Count	3.5e-51
Card_Category	Total_Relationship_Count	1.6e-24
Card_Category	Gender	3.6e-16
Card_Category	Income_Category (missing)	1.2e-14
Dependent_count	Income_Category (missing)	2.4e-14
Contacts_Count_12_mon	Gender	2.6e-07



EDA - Bivariate - *Attrition Flag vs Contacts Count*

- There seems to be a different contacts count distribution among the customers who retain their credit cards and those who do not.



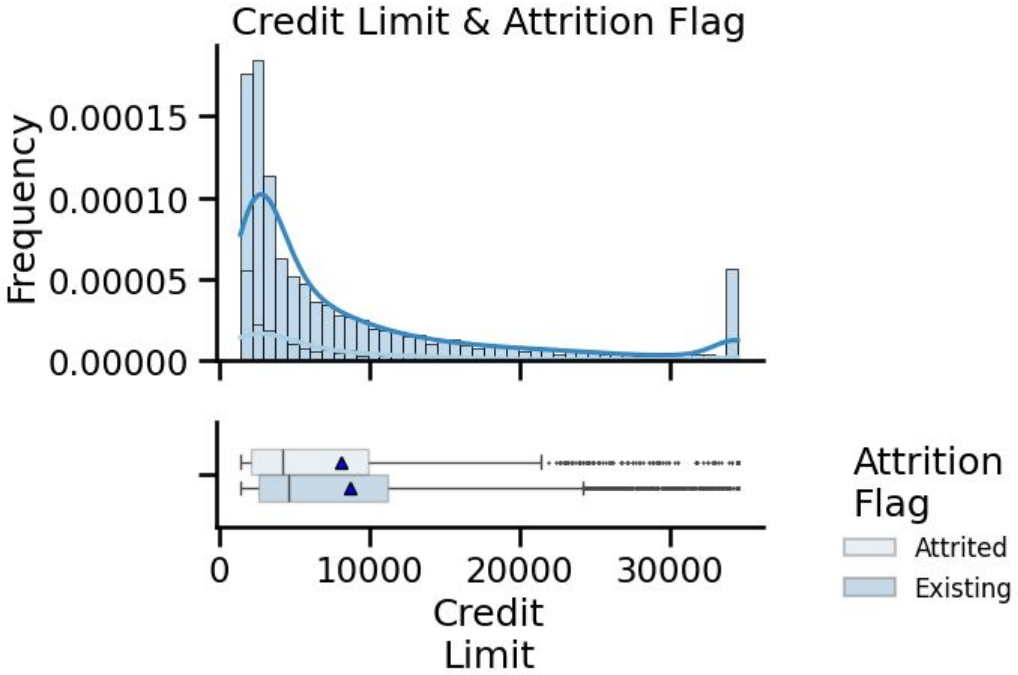
EDA - Bivariate - *Categorical-Numerical Variables*

- By conducting **One-Way ANOVA F-test** we observe that the following variables have effects on each other:
    - In the following slides, we take a look at some of these relationships.

Category	Numerical	p-value
Income_Category (missing)	Credit_Limit	0.0e+00
Income_Category (missing)	Avg_Open_To_Buy	0.0e+00
Total_Relationship_Count	Total_Trans_Amt	0.0e+00
Gender	Avg_Open_To_Buy	0.0e+00
Attrition_Flag	Total_Trans_Ct	0.0e+00
...	...	...
Months_Inactive_12_mon	Avg_Utilization_Ratio	1.9e-02
Contacts_Count_12_mon	Avg_Open_To_Buy	3.8e-02
Education_Level (missing)	Customer_Age	3.8e-02
Income_Category (missing)	Total_Amt_Chng_Q4_Q1	4.4e-02
Card_Category	Total_Revolving_Bal	4.6e-02

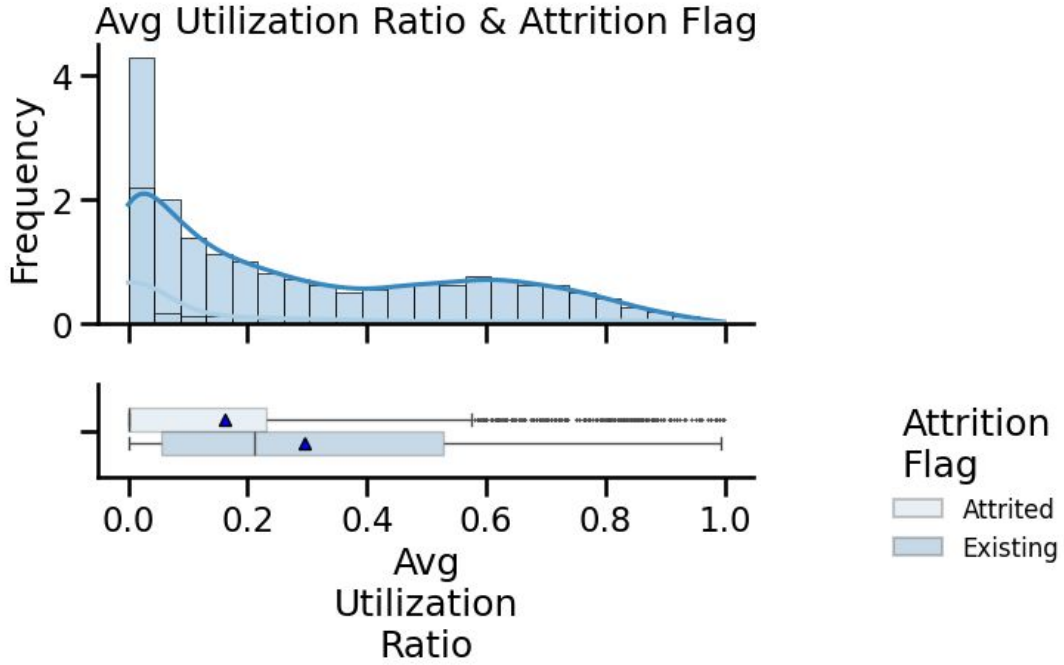
EDA - Bivariate - *Attrition Flag vs Credit Limit*

- We observe that the average income of customers who attrited is lower than those who do not.



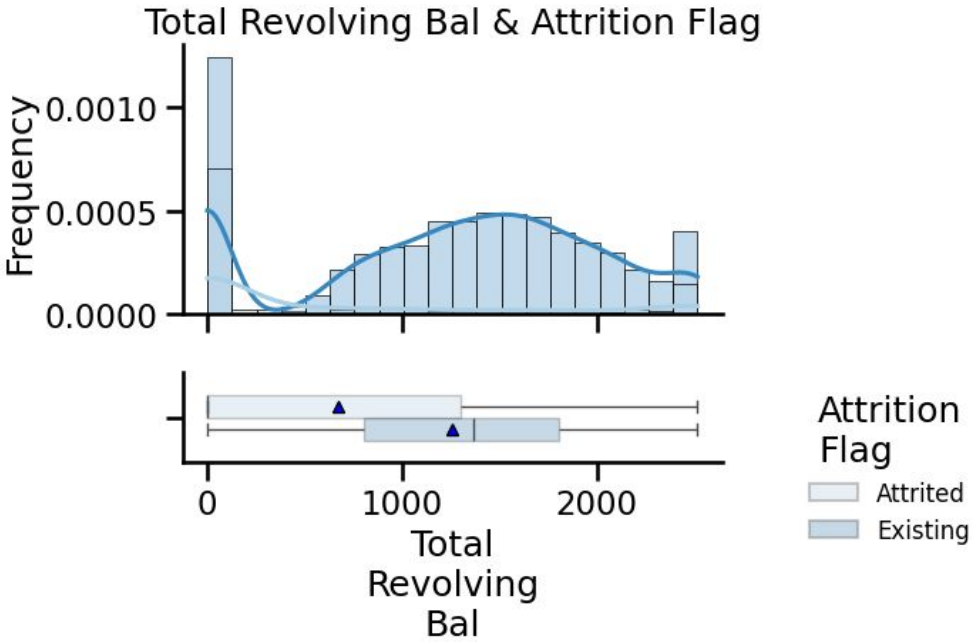
EDA - Bivariate - *Attrition Flag vs Avg Utilization Ratio*

- We observe that the average utilization of the customers who retain their card is significantly higher than those who do not.



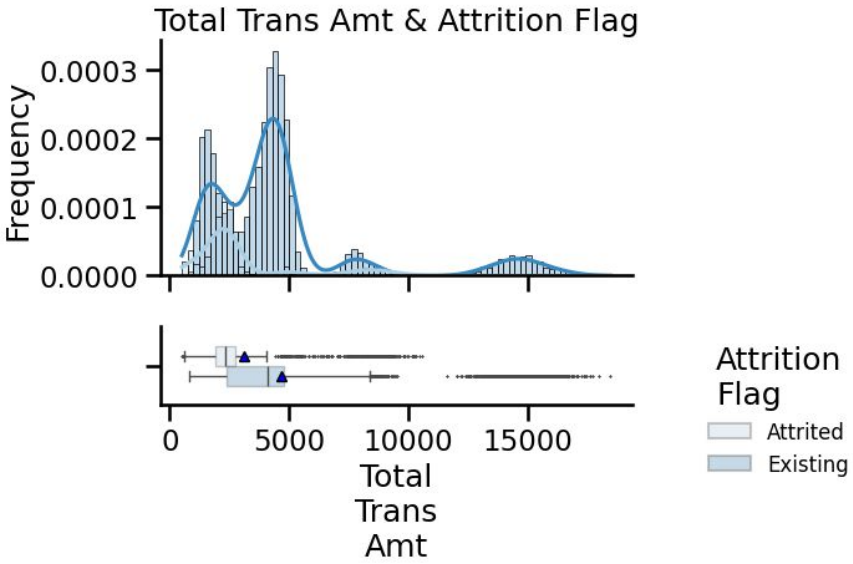
EDA - Bivariate - *Attrition Flag vs Total Revolving Balance*

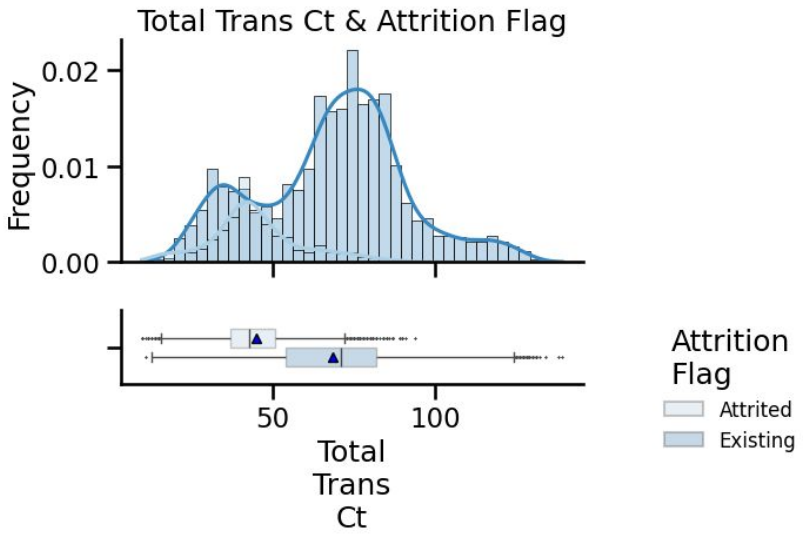
- We observe that the median total revolving balance is much lower in the customers who attrited.



EDA - Bivariate - *Attrition Flag vs Total Transaction Amount & Count*

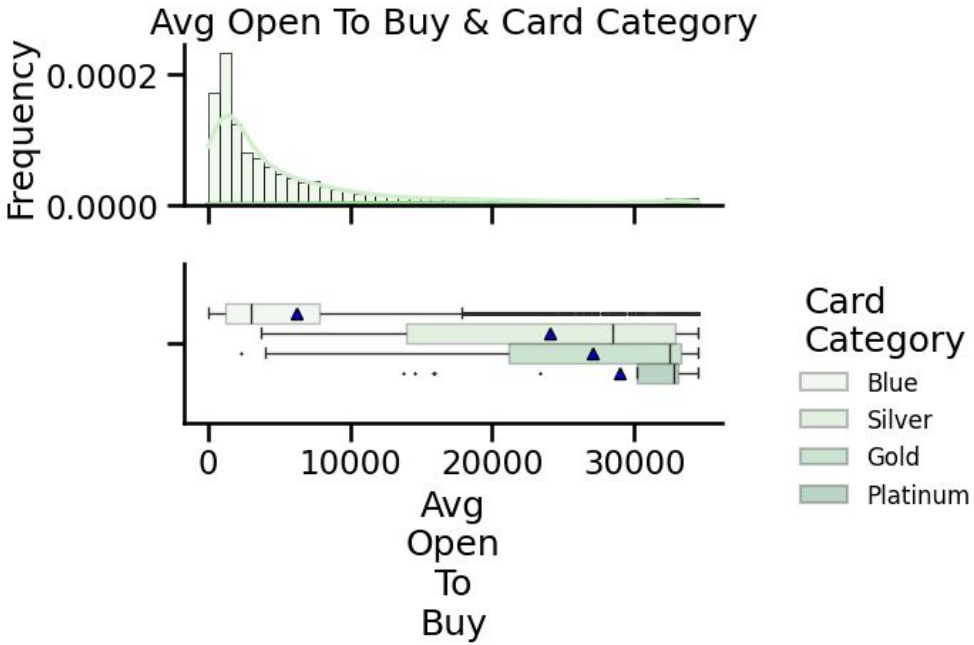
- We observe that the total transaction amount/count on average is higher amount the customers who retain their credit cards.





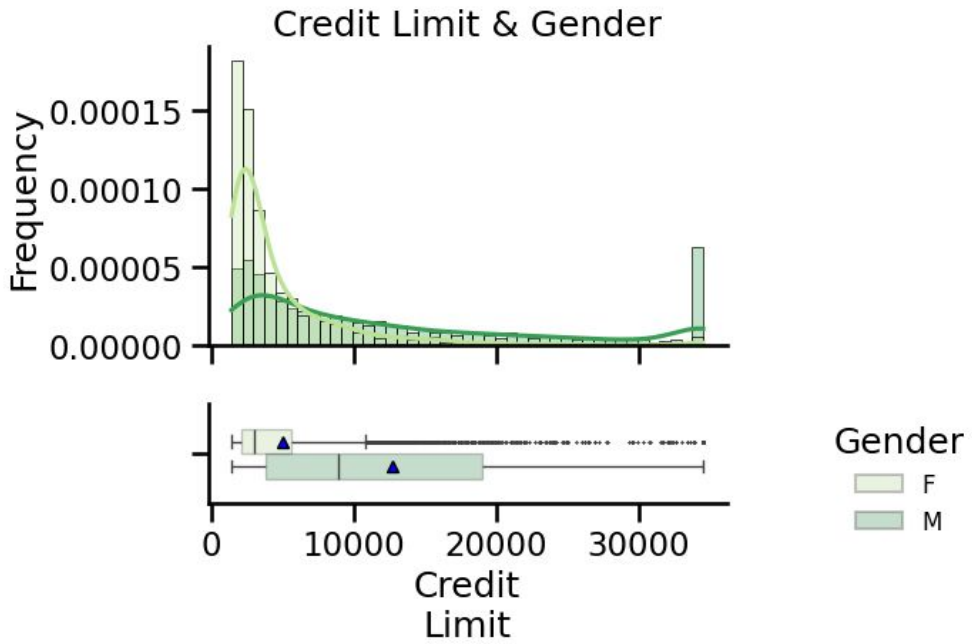
EDA - Bivariate - *Avg Open to Buy vs Card Category*

- We observe that those customers who hold the Blue card have less average open-to-buy limit.



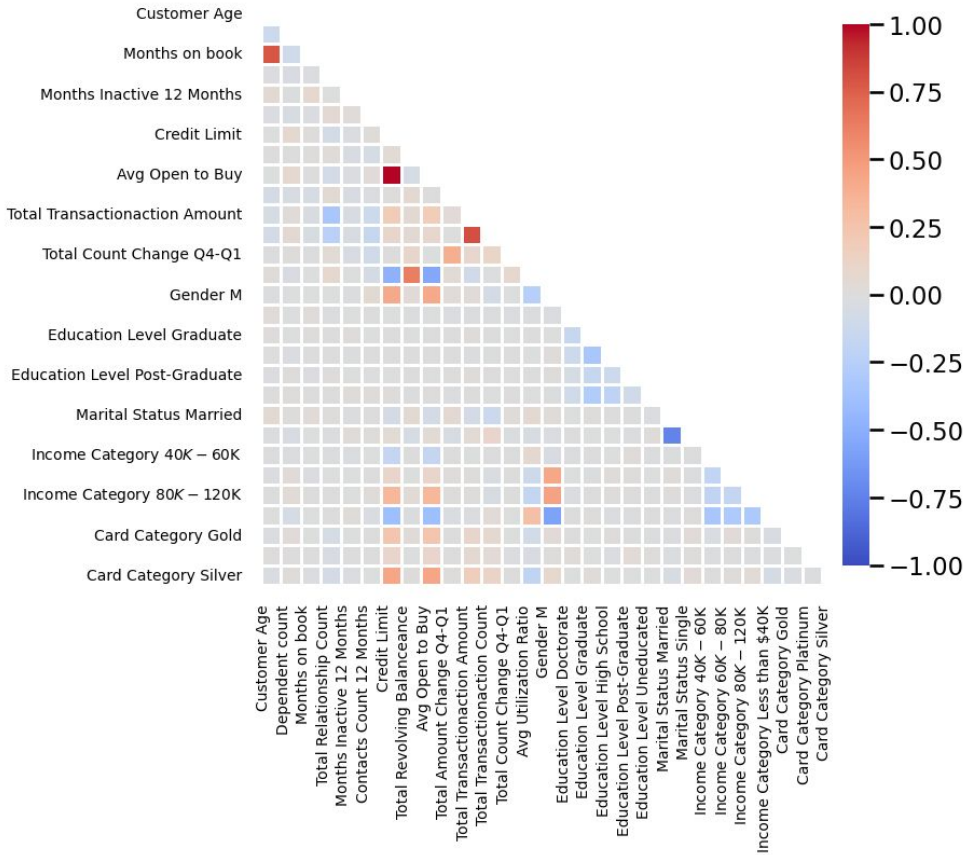
EDA - Bivariate - *Credit Limit vs Gender*

- We observe that the male customers on average have higher credit limits.



EDA - Bivariate - *All the variables*

- We can convert the categorical variables into dummy variables and construct the correlation matrix for all the columns
- This confirms our previous observations.



Data Preprocessing

- There are **no duplicates** in the original data.
- **Missing values**: There are 3,380 null values, 15% in Education Level, 11% in Income Category, and 7% in Marital_Status. These values are imputed using the ***most_frequent*** strategy.
- **Outliers:** The credit limit, average open-to-buy, total transaction amount have the most number of outliers.
- **Feature engineering:** There's a collinearity between credit limit and average open-to-buy. We need to drop one of these columns before modeling. The id column also needs to be dropped. The customer age and months on book are also correlated. We might need to drop one of these columns.

Data Preprocessing outliers

- Here are the percentage of the outliers in the columns.
- Treatment:
  - We drop the Credit Limit since it has collinearity with average open-to-buy variable.
  - We will try dropping months_on_book since it has outliers and also has a high correlation with the customer age.

Column	Outlier %
Credit_Limit	9.717
Avg_Open_To_Buy	9.509
Total_Trans_Amt	8.848
Contacts_Count_12_mon	6.211
Total_Amt_Chng_Q4_Q1	3.910
Total_Ct_Chng_Q4_Q1	3.891
Months_on_book	3.812
Months_Inactive_12_mon	3.268
Customer_Age	0.020
Total_Trans_Ct	0.020

Model Building

Train, Validation, Test split

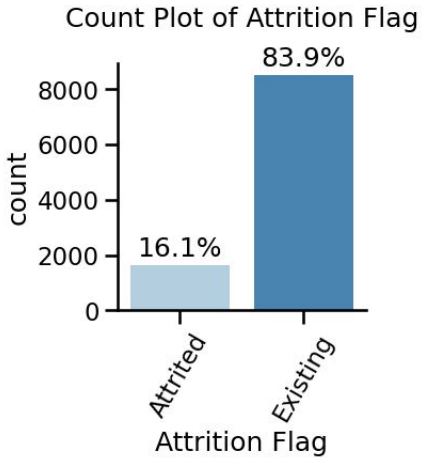
- The data is split into **train, validation and test** sets**.**
- We also conduct hyperparameter tuning to avoid overfitting the models.
- Here are the dimensions of the data partitions.

Data	Rows	Columns	Proportion %
X Train	8,101	27	80
X Validation	507	27	5
X Test	1,519	27	15

Model Building
Sampling

- We also conduct modeling on **oversampled** and **undersampled** data since we are dealing with an **imbalanced** class (Attrition_Flag).
- Here are the proportions of our target class in the original, oversampled and undersampled training set.

Attrition_Flag	Original %	Over %	Under %
0	83	50	50
1	16	50	50



Model Performance Summary

Model Evaluation Criterion

Our model can make wrong predictions in two ways:

- ***False Positive***: Predicting that a customer will not retain their credit card when they will.
  - If minimized, it improves the precision.
- ***False Negative***: Predicting that a customer will retain their credit card while they would attrit.
  - If minimized, it improves the recall.

In our problem, we are more interested in reducing the false negative and thus minimizing the **Recall**.

Model Performance Summary

- Here are some the trained models with the highest validation recall scores.
- The model parameters are also listed.
- We also want to have a small score difference between train and validation.
- Hence, we select the AdaBoost model, trained on oversampled data.

Model	Sampling	Train Recall	Validation Recall	Score Diff	learning_rate	estimator	init	n_estimators	subsample
GradientBoosting	Over	0.983	0.979	0.004	0.1			100	1.0
AdaBoost	Over	0.965	0.967	0.002	0.05	DecisionTreeClassifier		50	
GradientBoosting	Over	0.967	0.965	0.002	0.01		AdaBoostClassifier	100	0.7
AdaBoost	Under	0.970	0.963	0.007	0.1	DecisionTreeClassifier		100	
RandomForest	Over	0.996	0.955	0.041		DecisionTreeClassifier		110	
GradientBoosting	Under	0.984	0.951	0.033	0.1		AdaBoostClassifier	100	0.9
AdaBoost	Over	0.953	0.948	0.005	1.0			50	

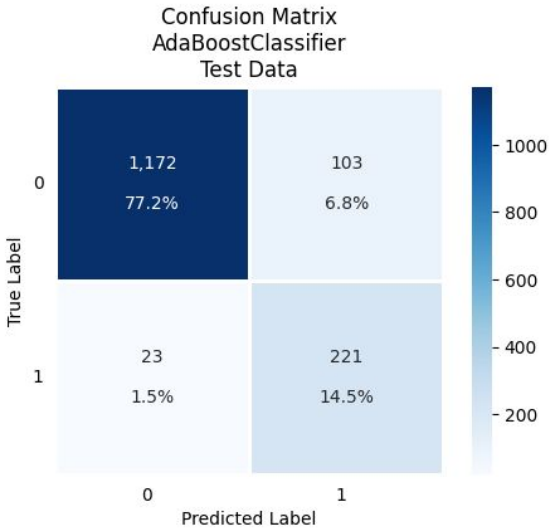
Model Performance Summary
Results on Test Data

- For the selected model, here are more train and validation details:

Model	Data	Sampling	Accuracy	Recall	Precision	F1	n_estimators	learning_rate	estimator	Best CV score
AdaBoost	Train	Over	0.948	0.965	0.934	0.949	50	0.05	DecisionTreeClassifier	0.958672
AdaBoost	Validation	Over	0.952	0.967	0.938	0.953	50	0.05	DecisionTreeClassifier	0.958672

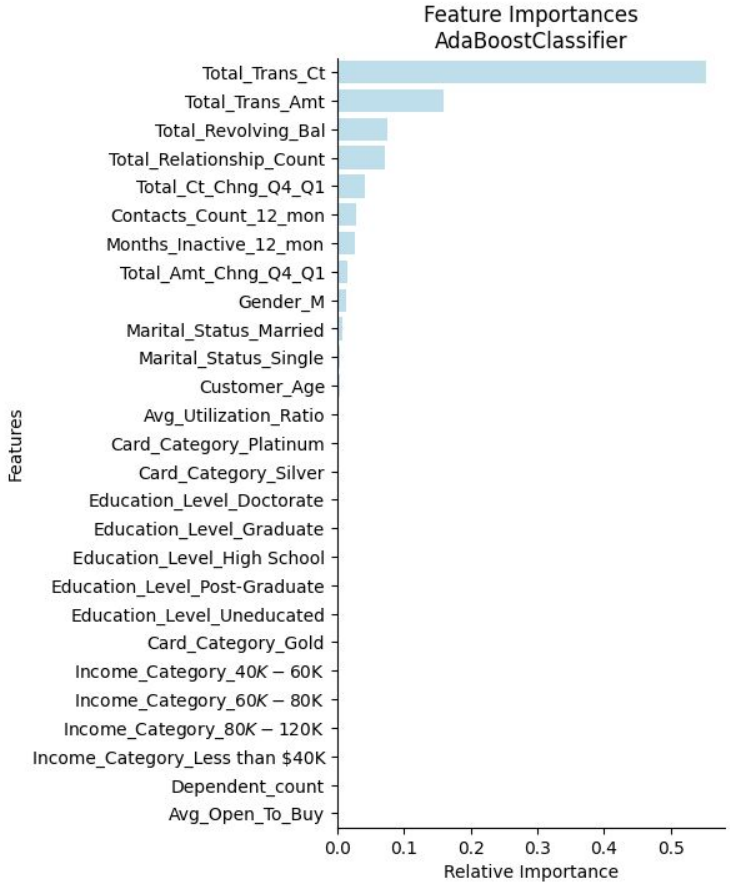
- Here are the results on the test data:

Test Scores	Accuracy	Recall	Precision	F1
AdaBoostClassifier	0.917	0.906	0.682	0.778



Model Performance Summary
Important Features

- Here are the important features detected by our model which match our observations from EDA.



APPENDIX

Model Performance Summary (original data)

- Here are the results summary for the original data.

Model	Sampling	Train Recall	Validation Recall	Score Diff	learning_rate	estimator	init	n_estimators	subsample
AdaBoost	Original	0.839	0.852	0.013	0.1	DecisionTreeClassifier		100	
Bagging	Original	0.998	0.852	0.146				70	
GradientBoosting	Original	0.891	0.840	0.051	0.1		AdaBoostClassifier	100	0.9
GradientBoosting	Original	0.889	0.827	0.062	0.1			100	1.0
Bagging	Original	0.979	0.778	0.201				10	
AdaBoost	Original	0.784	0.765	0.019	1.0			50	
RandomForest	Original	0.977	0.741	0.236		DecisionTreeClassifier		25	

Link to Appendix slide on model assumptions

Model Performance Summary (oversampled data)

- `SMOTE` from `imblearn.over_sampling` is used for undersampling the data.
- The GradientBoosting model works well on the oversampled data.

Model	Sampling	Train Recall	Validation Recall	Score Diff	learning_rate	estimator	init	n_estimators	subsample
GradientBoosting	Over	0.983	0.979	0.004	0.1			100	1.0
AdaBoost	Over	0.965	0.967	0.002	0.05	DecisionTreeClassifier		50	
GradientBoosting	Over	0.967	0.965	0.002	0.01		AdaBoostClassifier	100	0.7
RandomForest	Over	0.996	0.955	0.041		DecisionTreeClassifier		110	
AdaBoost	Over	0.953	0.948	0.005	1.0			50	
Bagging	Over	0.997	0.944	0.053				10	

Model Performance Summary (undersampled data)

- `RandomUnderSampler` from `imblearn.under_sampling` is used for undersampling the data.
- Here are the models with the highest recall scores and small train-validation score difference.

Model	Sampling	Train Recall	Validation Recall	Score Diff	learning_rate	estimator	init	n_estimators	subsample
AdaBoost	Under	0.970	0.963	0.007	0.1	DecisionTreeClassifier		100	
GradientBoosting	Under	0.984	0.951	0.033	0.1		AdaBoostClassifier	100	0.9
GradientBoosting	Under	0.983	0.938	0.045	0.1			100	1.0
RandomForest	Under	0.998	0.938	0.060		DecisionTreeClassifier		25	
AdaBoost	Under	0.937	0.914	0.023	1.0			50	
Bagging	Under	0.993	0.901	0.092				10	

Great Learning

Happy Learning !