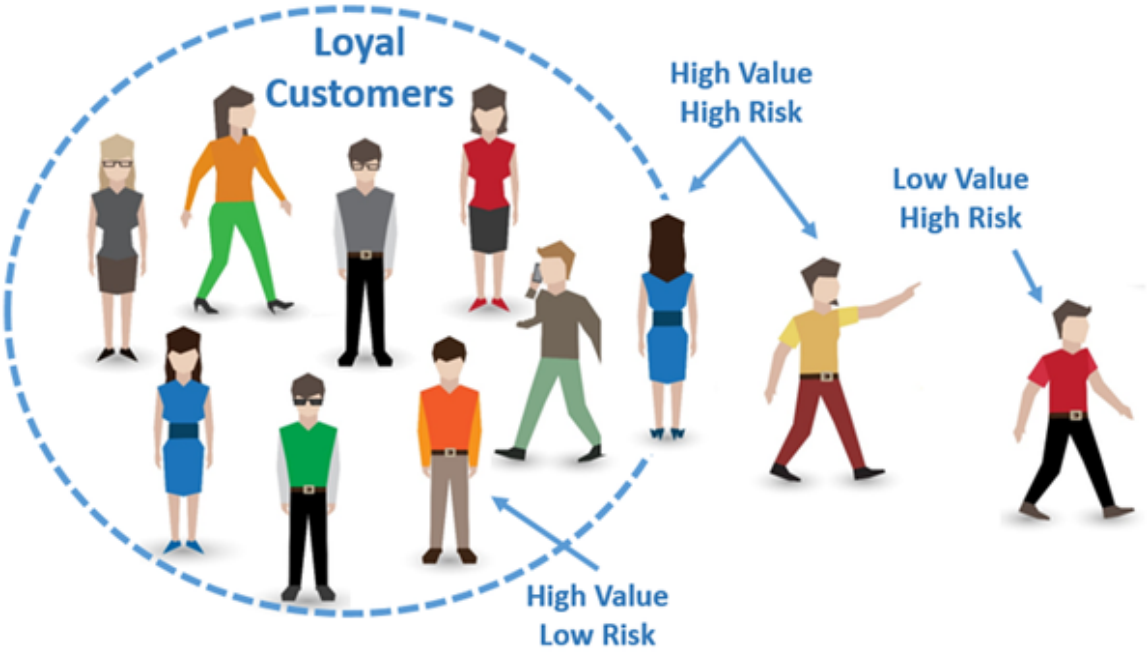
Bank Customer Attrition Prediction
Group 17 Team Project



Team Members:

Maria Rodriguez

Harkirat Johal

Balaji Gopal

Panteha Golkar

Lily Liu

Mohit Verma

TABLE OF CONTENTS

1. INTRODUCTION

1.1. INTRODUCTION TO THE DOMAIN

One of the most profitable businesses for the Student Bank of Waterloo is its Credit Card service. The major contribution for the income from the Credit Card service stems from two sources: Customer Interest and Merchant Fee (or Interchange Fee). Customers of the bank can borrow money from their credit account with an obligation to repay the entire balance in a given period. Interest gets charged on the account for any remaining balance, which is one of the major streams of income for the bank.

Interchange contributes to the second largest source of income. When a merchant (or retailer) accepts a customer's payment via a credit card, the merchant owes the bank an agreed percentage of the sale amount.
With over 20,000 cardholders, the Credit Card Service contributes to about 35% of the Bank's total profit.

1.2. PROBLEM STATEMENT

Over the last 3 years, Student Bank of Waterloo has seen a steady loss of its credit card clientele. Data analysis from 2018, 2019 and 2020 indicate 13%, 16% and 17% of customers have discontinued using the credit card service, each year respectively. The process is showing a sustained effect to be categorized as a Customer Attrition problem.

Keeping aside the 'competitor's effect' resulting in losing the Bank's customers to other service providers, the problem at hand is to understand the driving forces behind the Attrition with a Customer focus. The problem may be analyzed in two main areas: First is to understand credit card customers from different points of view (including demographic profile, card type and utilization) and, Second, to analyze the factors contributing to Customer Attrition.

1.3. BOARD'S DECISION

Overall, the effect is showing a significant impact on the Bank's financial bottom-line and position in the share market. The Board of Directors is taking this matter seriously and would like to take actions and measures to remediate the Customer Attrition problem.

They have realized that the right direction to solve the problems identified under 'problem statements' is to perform a detailed statistical analysis on their Credit Card Customer's profiles with the data in hand and derive useful conclusions and recommendations to help improve the situation. Moreover, the board is also interested in being able to predict whether a credit card customer is trending towards

discontinuation and eventually customer attrition, so that they can take remedial measures.

The Board has decided to invest in hiring a team of professional Data Scientists having diverse backgrounds to find the solution to their problems.

1.4. DATASET INFORMATION

Data Source: https://leaps.analyttica.com/sample_cases/11

This dataset contains data of over 10,000 credit card accounts with 21 variables of different types in CSV format. Data can be categorized into a customer profile, their credit card history and attrition indicator. Table 2.1 shows the detailed data dictionary.

Table 2.1: Data Dictionary

Variable	Type	Description
Clientnum	Num	Client number. Unique identifier for the customer holding the account
Attrition_Flag	Char	Internal event (customer activity) variable - if the account is closed then 1 else 0
Customer_Age	Num	Demographic variable - Customer's Age in Years
Gender	Char	Demographic variable - M=Male, F=Female
Dependent_count	Num	Demographic variable - Number of dependents
Education_Level	Char	Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.)
Marital_Status	Char	Demographic variable - Married, Single, Unknown
Income_Category	Char	Demographic variable - Annual Income Category of the account holder (< $40K, $40K - 60K, $60K - $80K, $80K-$120K, > $120K, Unknown)
Card_Category	Char	Product Variable - Type of Card (Blue, Silver, Gold, Platinum)
Months_on_book	Num	Months on book (Time of Relationship)
Total_Relationship_Count	Num	Total no. of products held by the customer
Months_Inactive_12_mon	Num	No. of months inactive in the last 12 months
Contacts_Count_12_mon	Num	No. of Contacts in the last 12 months
Credit_Limit	Num	Credit Limit on the Credit Card
Total_Revolving_Bal	Num	Total Revolving Balance on the Credit Card
Avg_Open_To_Buy	Num	Open to Buy Credit Line (Average of last 12 months)
Total_Amt_Chng_Q4_Q1	Num	Change in Transaction Amount (Q4 over Q1)
Total_Trans_Amt	Num	Total Transaction Amount (Last 12 months)
Total_Trans_Ct	Num	Total Transaction Count (Last 12 months)
Total_Ct_Chng_Q4_Q1	Num	Change in Transaction Count (Q4 over Q1)
Avg_Utilization_Ratio	Num	Average Card Utilization Ratio

Note: Columns were renamed for simplicity and readability purposes.

1.5. OBSERVATIONS AND DATA PREPARATION

- Just 16% of customers have attrited, so the data won't be balanced to effectively predict churning customers.
- Although there are no missing or null values, some columns (Education_Level, Marital_Status and Income_Category) have "Unknown" values.
- Not all columns are distributed normally.
- Columns were renamed for simplicity and readability purposes.
- Education_Level: replaced "Unknown" with "Graduate" (being its mode^) and merged "Doctorate" with "Post-Graduate" to reduce unnecessary granularity.
- Marital_Status: replaced "Unknown" with "Married" (being its mode^).
- Income_Category: replaced "Unknown" with $40K (being its mode^).
- To prepare data for regression modeling pandas get_dummies is used to convert categorical data to binary columns.
- One column is dropped from each group as a reference.

Note: Mode^ - category type having the most number of occurrences for any given column.

2. OBJECTIVES

To start, the team would like to employ Hypothesis Testing, which is conducting tests to validate an assumption regarding a population parameter.

A. **Customer Profile:** Validating differences in profiles of credit card users,
   Null Hypotheses:
   a. There is no difference in distribution between male and female customers.
   b. There is no relationship between age and income among the customers.
   c. There is no relationship between marital status and income.
   d. There is no association between education and income.

B. **Credit Card Profile:** Analysing the most popular credit card types,
   Null Hypotheses:
   a. There is no difference in the distribution between various credit card types.
   b. Card type is independent of the customer's age, gender, education, marital status and income.

C. **Credit Utilization Profile:** Factors associated with high use of credit funds among users,
   Null Hypotheses:
   a. There is no difference in the amount or number of transactions between genders.
   b. There is no difference in the number of transactions between different card types.
   c. There is no relationship between the credit limit and utilization ratio.

D. **Attrition Prediction:** Predict which users are likely to stay as customers instead of churning,
   Null Hypotheses:
   a. The outcome of attrition is not affected by any of the base attributes in the dataset, and thus cannot be predicted.

The working team would like to proceed with developing a Predictive Model to predict the trend toward Customer Attrition.

3. ANALYSIS

3.1. EXPLORATORY ANALYSES

Exploration of numerical attributes, showed that attributes Age, Amount Chg, Contacts, Dependents were close to a normal distribution, while the rest of the distributions are skewed. Among categorical attributes, Gender was nearly uniformly distributed. In addition, we identified the relationship between few attributes which are displayed below:

* Credit Limit = Revolving Bal + Avg Open

* Utilization Ratio = Revolving Bal / Credit Limit

3.2. CUSTOMER PROFILE

To examine possible associations between Age, Gender, Marital Status and Education with Income, several null hypotheses as outlined in section 2A were tested.

The following points describe the outcome of the selective hypothesis tests conducted and the results obtained.

1) Hypothesis Test 2.A.b: There is no relationship between age and income among the customers.

   Pre-condition check and Execution:

   While exploring the relationship between Age and Income groups, we were interested to find out whether Age is significantly different across multiple income groups. Before applying the ANOVA test, we confirmed the data conditions. Firstly, the observations were mostly independent within and across groups. Secondly, the Age in each income group indicated nearly normal distribution in the histogram and lastly, boxplot inspection showed us that Age across these groups are comparable.

   Note: Please refer to the table displayed under Appendix titled Figure A.1.1.

   To validate the variability of mean and standard deviation among all groups to be significant, we attempted F-test (using f-oneway function) to evaluate the hypothesis, found a fairly large F-value of 5.80529 and a small P-value of 0.00011 which provided us strong evidence against the null hypothesis. Moreover, in order to further analyze which group was significant and mitigate the risk of finding false differences between groups, we attempted the Bonferroni correction method.

Note: Please refer to the table displayed under Appendix titled Figure A.1.2.

Results:

Based on the ANOVA test's low p-value, we conclude that at least one income group has a different age distribution from the rest. And, based on the outcome of the MultiComparison test, we conclude that the age distribution of $120+ Income Group is the different one. And, Those earning the $120+ are older with a mean age of 47.6 +/- 6.7 (CI 43-53).

2) Other Hypotheses Tests 2.A.a/c/d: There is no difference in distribution between male and female customers. There is no relationship between marital status and income. There is no association between education and income.

Pre-condition check and Execution:

Besides Age, the rest of the variables are categorical, therefore we chose the Chi-Square testing approach. Started with examining whether our dataset met the below conditions: firstly, data observed as frequencies instead of percentages, secondly the levels of variables were mutually exclusive, thirdly the observations under study are mostly independent and lastly each level contained at least 5 observations in a contingency test. We then set 0.05 as the default significance level for all tests. Validated the Gender variable with expected equal frequencies, found a P-value was close to zero, thus data provided strong evidence to reject the claimed hypothesis.

Note: Please refer to Figure A.2 under the Appendix for the visual of this distribution.

We experimented with a pairwise contingency tables test using Chi-Square distribution for Income with Marital Status, Income with Education to explore an association, received P-value of 0.2273 for 1st pair and 0.5979 for 2nd pair. There was a failure to reject the null hypothesis.

Results:

There is a significantly higher proportion of females among the customers. Income level is independent of marital status and education level.

3.3. CREDIT CARD PROFILE

To examine possible associations between a few variables: Age, Gender, Education, Marital Status, Income with Card Types, several null hypotheses as outlined in section 2B are tested.

The following points describe the outcome of the selective hypothesis tests conducted and their results obtained.

1) Hypothesis Test 2.B.a/b: There is no difference in the distribution between various credit card types. Card type is independent of the customer's age, gender, education, marital status and income. There is no association between education and income.

   Pre-condition check and Execution:

   Once again, we confirmed that data pre-conditions were met for conducting the Chi-Square test. We performed a test for the Card Types variable expecting equal frequencies, but found the P-value was close to zero. Thus, data provided strong evidence to reject the claimed null hypothesis.

   We converted Age into categorical variables by grouping them into Age groups, performed pairwise contingency tests for Card Types by Age groups, Gender, Education, Marital Status and Income respectively to explore the relationships. Got the results P-value of 0.0722; 0.0000; 0.2333; 0.0024; 0.0000 respectively. To further explore those pairs with pretty small P-value, which are Gender, Marital Status and Income variables, we re-calculated P-value by dissecting Card to each type instead of considering it as a single variable, data revealed more details of the association.

   Note: Please refer to Figure A.3 under the Appendix section of this report.

   Result:

   The proportions of the card types are significantly different, with Blue predominating, followed by Silver, Gold, and Platinum the least common. The card type held by the customer is independent of his/her age group. Further dissection of the data revealed significant differences in the distribution of genders among the different card types. Card type is independent of educational attainment. And, there are relationships between card types and marital status.

3.4. CREDIT UTILIZATION PROFILE

Based on the credit utilization profile, we wanted to check possible relationships in the variables below: Gender, Transactions Amount, Transactions Count, Card Types, Credit Limit and Credit Utilization Ratio.

The following points describe the outcome of the selective hypothesis tests conducted (as outlined in section 2.C.a/b/c) and their results obtained.

Hypotheses Tests: There is no difference in the amount or number of transactions between genders. There is no difference in the number of transactions between different card types. There is no relationship between the credit limit and utilization ratio.

Pre-condition check and Execution:

1) For examining the relationship between Gender and Transaction Amounts, we chose to perform a T-distribution test after confirming the data met the pre-conditions: each proportion of Genders followed a nearly normal distribution and samples were independent of each other. The result of the test showed a P-value of 0.0123. Following that we ran another examination between Gender and Transactions Count, we applied the same method and obtained a P-value of 0.0000. Therefore, we could safely reject the null hypothesis. Also, the scatter plot merged the difference in amount and volume of the transaction between cardholders' genders.

   Note: Please refer to figure A.4 in the Appendix.

2) To inspect whether Transactions Count are significantly different across all Card Types, we chose ANOVA F-test (using f-oneway function). Results showed a fairly small P-value of 0.0000. Moreover, we performed the Bonferroni correction method to verify our findings.

   Note: Please refer to A.5 in the Appendix.

3) We applied a linear regression model by using the Ordinary least squares(OLS) method when analyzing the relationship between Credit Limit and Utilization Ratio since both variables are continuous. We applied boxcox transformation on the Credit Limit. We got an estimated intercept to be approximately 6.269964 and slope to be approximately -1.888. Hence, revealed that Credit Limit is negatively correlated with Utilization Ratio, as indicated during exploratory analysis. Moreover, we evaluated the linear model by inspecting regression residuals, observed that the residuals were nearly normally distributed and centered close to zero, hence it qualified as a validation for our model.

Note: Please refer to figure A.6 under Appendix.

Results:

Female customers make more frequent transactions. Male customers make more expensive transactions. At least one of the card types has a different distribution for the transaction count. Blue cardholders make less frequent transactions compared to the other card types. Credits up to 4 thousand dollars are well utilized. Beyond 4 thousand, there is a sharp decline in credit utilization.

3.5. ATTRITION PREDICTION

The goal of any organization such as a bank is to retain as many customers as possible or to take the necessary proactive steps to reduce the risk of customer attrition. In this scenario, a business manager wants to understand the primary reasons for attrition and determine the customers who are likely to close their accounts with the bank in near future. Analyzing data would enable understanding of the causes of attrition and likely future occurrence. This information is very valuable to the bank as they can concentrate their efforts towards high-risk groups and preemptively avoid customer attrition.

The customer attrition problem is binary with two possible outcomes (the customer churns, customer does not churn) a logistic regression model is an appropriate choice for modeling. The next step was to fit the model based on available predictors and use a backward elimination process to remove insignificant or non-contributory variables. The null hypothesis is that the outcome of attrition is not affected by any of the attributes, hence it cannot be predicted. This process was iterated until the final model for attrition (response variable) is attained. The model would then be used to determine the odds (p) for a customer to churn based on the predictor values.

To simplify the initial model and to validate regression results, single attribute hypothesis tests were performed to determine and eliminate insignificant predictors. Table A.1 in the Appendix summarizes the results of various tests and their dependence on customer attrition.

To assume a logistic regression model, a few underlying assumptions need to hold true. Firstly, the observations must be independent of each other. In this case, every record of data corresponds to a different customer. The next assumption for logistic regression is there should be little or no multicollinearity between independent variables (predictors). To check for the multicollinearity, a correlation heatmap was produced. An initial heatmap, with all the predictors helped to identify the relationships between these variables. (Please refer to Figure A.7 in the Appendix)

The following heatmap (Figure 1) is a reduced version displaying all numerical variables that have a minimal correlation.**"Duration", "Avg_Open", "Transaction_Amt", "Amount_chg"** and **"Util_ratio**" were dropped.

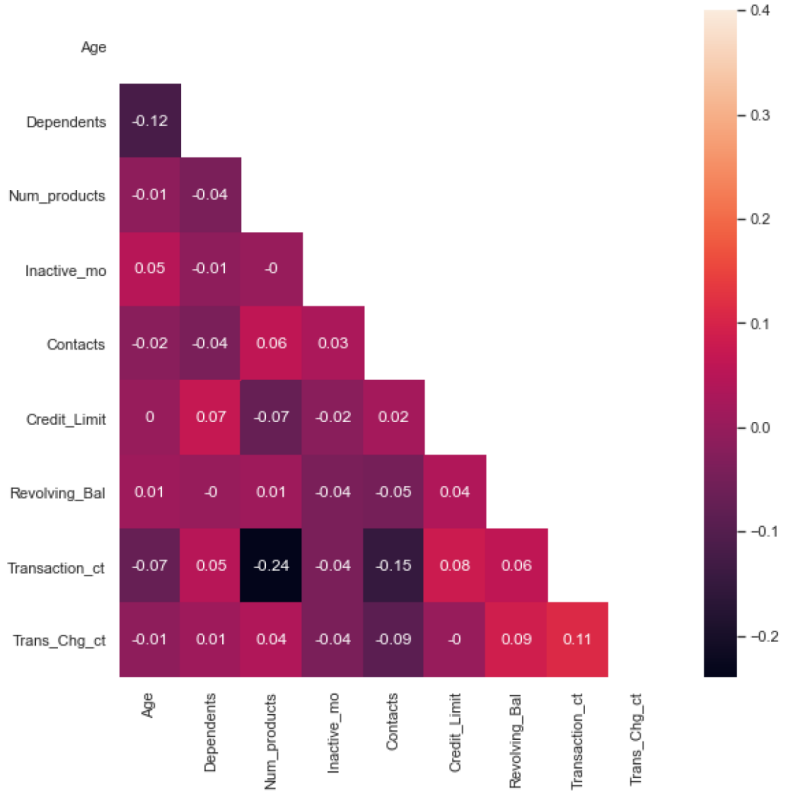


Figure 1: Reduced Correlation Heatmap

Going through the process of fitting the model, eliminating non-significant **(p < 0.05)** or non-contributory (low coefficient, < |0.1|) variables and refitting the model, the final logistic regression model was obtained. A train set of the data (70% of total) was utilized and a test set (30 % remaining) was kept for comparing the predictions (attrition or no attrition) of the model to the actual outcomes. Table 1 below summarizes the significant predictors for attrition (response) and their corresponding regression coefficients.

Table 1: Final Regression Model Predictors, Coefficients and Odds-Ratios

Predictor	Coefficient	Odds-Ratio
Male {0: Female, 1: Male}	-0.2956	0.7357
Dependents	0.1177	1.1014
Num_products	-0.3196	0.7376
Trans_Chg_ct	-5.0904	0.007138
Inactive_mo	0.4450	1.5176
Contacts	0.5728	1.7594

These regression coefficients of the predictors provide the following information about customer attrition:

Liable to attrition

- Customers with more dependents
- Customers with a higher duration of inactivity
- Customers with a higher frequency of contacts with the bank

Less Liable to Attrition

- Male customers less likely to churn than female customers
- Customers that use more bank products
- Customers with a higher change in the number of transactions from Q4 to Q1

After establishing a logistic regression model, the next step is to evaluate the model based on how it predicts the response variable (attrition). Figure A.8 in the Appendix displays the relationship between the predicted values of *Logit(p)* and the odds (*p*), along with the actual values of the response (attrition).

The predictions are compared to the test dataset. Table 2 below summarizes the prediction metrics for the obtained model:

Table 2: Prediction Metrics for Logistic Regression Model

Metric	Value
True Negative, TN	5843
False Positive, FP	105
False Negative, FN	812
True Positive, TP	328
Accuracy, ACC	0.87
Precision	0.76
Recall (Sensitivity)	0.29
Area under ROC Curve, AUROC	0.8026

Figure A.9 in the appendix shows the Receiver Operating Curve (ROC), with the corresponding balanced sensitivity and specificity point, the optimal threshold (Youden's J) point and the default threshold ($d_{th} = 0.5$) point. Evaluating the predicted results using the test dataset, Figure 2 shows the confusion matrix heatmap comparing the results to the observed response (at default threshold=0.5).

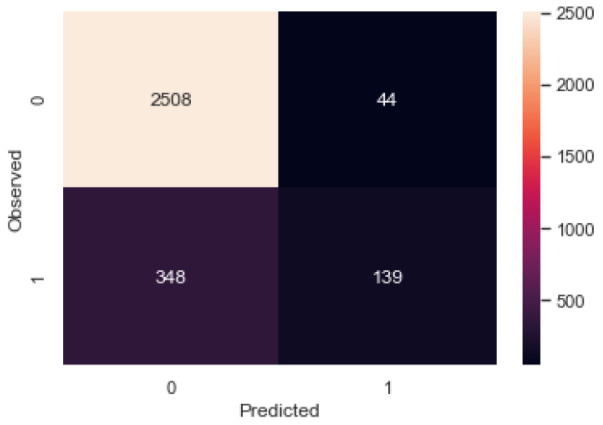


Figure 2: Confusion Matrix Heatmap (Test Dataset)

Table 3 shows a summary of the prediction metrics comparing the model's predictions and the observed results from the test dataset:

Table 3: Prediction Metrics (based test dataset)

Metric	Value
True Negative, TN	2488
False Positive, FP	46
False Negative, FN	359
True Positive, TP	146
Accuracy, ACC	0.87
Precision	0.76
Recall (Sensitivity)	0.29

At a default discrimination threshold of 0.5, the logistics regression model presents a good accuracy of 87% for accurately predicting whether a customer will churn or not. It also has a high area under the Receiver Operating Curve (AUROC) value of 0.81 which indicates that the model performs better than a random estimator for attrition. It also tells us that the model has a high overall capacity to discriminate between positive and negative responses (attrition or no attrition) at different thresholds. The recall(sensitivity) of detecting attrition is low at 29%, however, considering the non-critical status of credit-card attrition, this could be maintained, instead of sacrificing the high precision and accuracy of the prediction model.

4. SUMMARY

Customer profile

Findings:

- ❏ There is a significantly higher proportion of females among the customers.
- ❏ High-income card owners are older, with a mean age of 47.6 +/- 6.7 (CI 43-53).
- ❏ Income level is independent of marital status and education.

Recommendations:

Provide the bank with background on either increasing the satisfaction by offering female-oriented or older-age oriented rewards OR by an increased effort to enhance recruitment of males and low-mid-income groups.

Credit Card Profile

Findings:

- ❏ Blue cards are the most common, followed by Silver, Gold and Platinum.
- ❏ The card type is independent of age and education.
- ❏ Blue cards are popular with females, married customers, and the low-income group.
- ❏ Silver and Gold cards are popular with males, single customers, and those with annual earnings of 70 thousand dollars and above.

Recommendations:

1. The Bank can enhance the recruitment of credit card owners by offering activities and rewards targeted to females, married people, and offering savings-oriented plans for the low-income group.
2. The Bank can offer exclusive rewards targeted towards the male population, single people and higher-income earners.

Credit Utilization Profile

Findings:

- ❏ Females make more frequent transactions.
- ❏ Males make more expensive transactions.
- ❏ Blue cardholders make less frequent transactions.
- ❏ Credit limits up to around 4 thousand dollars are well utilized, with a sharp decline in utilization beyond 4 thousand.

Recommendations:

1. The Bank can tailor the benefits of credit card charges (% per use and % of the amount payable) to a target population, similar to Air Mile rewards.
2. The Bank can gauge the liquidity of their assets by assigning on average 4 thousand credit units per cardholder.

3. Banks can offer lower fees for credit use above 4 thousand to increase utilization.

Attrition Prediction

Summary:

1) Attrition, when analyzed one-to-one with variables, did not show a relationship with age, marital status, education level, number of dependents, card category, duration of the relationship with the bank.
2) Attrition showed an association with:
    a) Gender: females are more liable to attrition than males.
    b) Income: low income (<40 thousand dollars) is more liable to attrition. An income of 70 thousand dollars is less liable to attrition.
    c) Number of bank products used: There is less attrition with higher product use.
    d) Inactivity: Higher attrition with longer inactivity.
    e) Contacts: Higher attrition with increased frequency of contacts with the bank.
    f) Credit limit: Higher attrition with a lower mean credit limit.
    g) Revolving balance: Higher attrition with lower revolving balance.
    h) Transaction counts: Higher attrition with less frequent transaction counts.
    i) Transaction amount: Higher attrition with lower transaction amounts.
3) Logistic regression analysis showed that Attrition may be predicted by gender, number of dependents, number of products used, transaction count changes and account inactivity. At the 0.5 threshold, the accuracy of predicting attrition is 87%, with an AUC of 0.81.
4) Banks can do a qualification check that includes the above parameters to ascertain good customer retention and minimize Customer Attrition.

APPENDIX

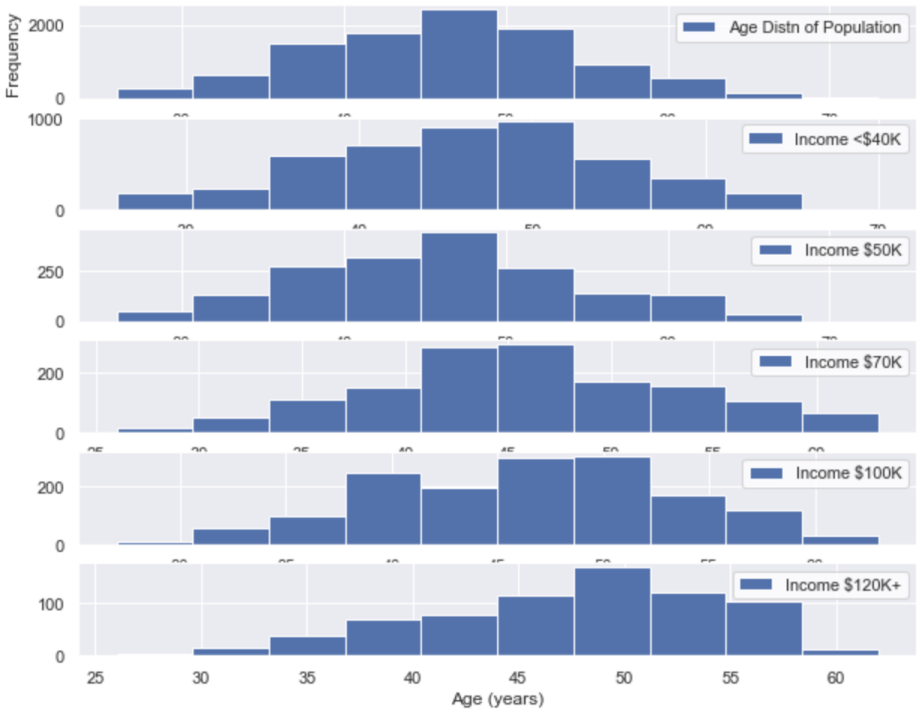


Figure A.1.1: Histogram of Age across Income groups

Test Multiple Comparison ttest_ind FWER=0.05
method=bonf alphacSidak=0.01, alphacBonf=0.005

group1	group2	stat	pval	pval_corr	reject
$100K	$120K	-3.8241	0.0001	0.0013	True
$100K	$40K	0.5678	0.5702	1.0	False
$100K	$50K	1.2649	0.206	1.0	False
$100K	$70K	1.7316	0.0835	0.8345	False
$120K	$40K	3.9699	0.0001	0.0007	True
$120K	$50K	4.3164	0.0	0.0002	True
$120K	$70K	4.9196	0.0	0.0	True
$40K	$50K	0.8704	0.3841	1.0	False
$40K	$70K	1.2835	0.1994	1.0	False
$50K	$70K	0.4131	0.6796	1.0	False

Figure A.1.2: Comparison of Age across Income groups

```
p val=  [0.]
```

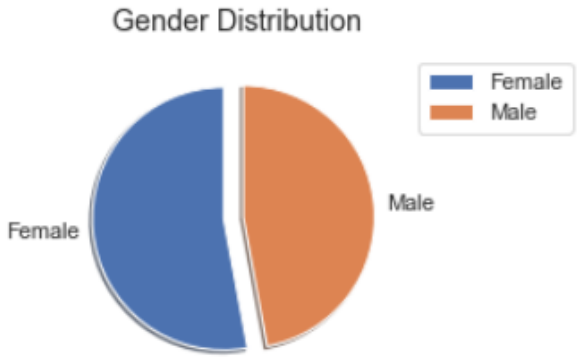


Figure A.2: P-value and distribution for Gender

```
p value =  0.0
```

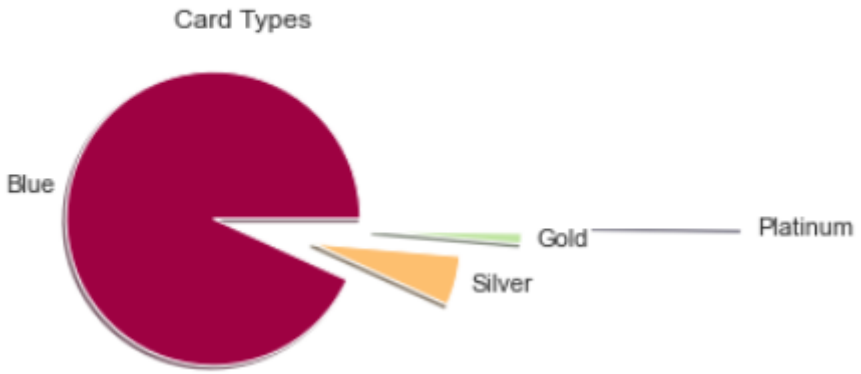


Figure A.3: P-value and distribution for Card Types

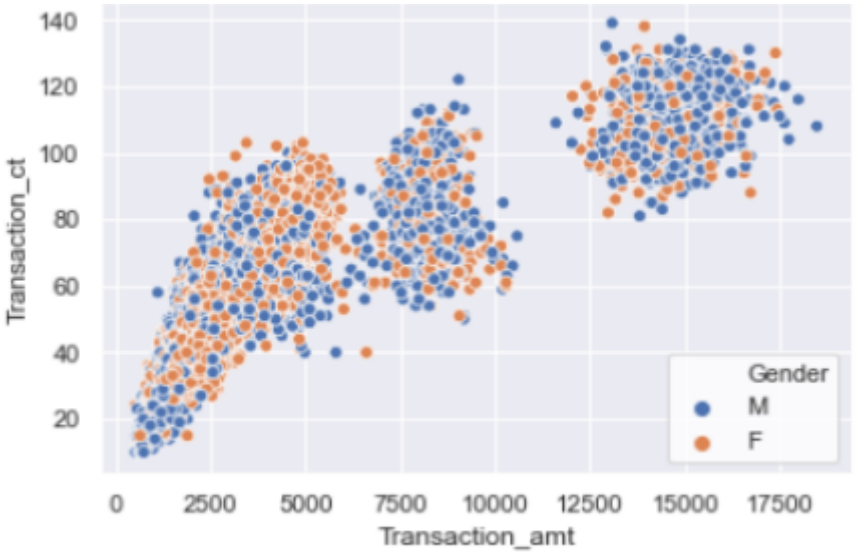


Figure A.4: Correlation Transaction Amount vs Count per Gender

Test Multiple Comparison ttest_ind FWER=0.05 method=bonf
alphacSidak=0.01, alphacBonf=0.008

group1	group2	stat	pval	pval_corr	reject
Blue	Gold	-8.1372	0.0	0.0	True
Blue	Platinum	-4.4577	0.0	0.0001	True
Blue	Silver	-10.5299	0.0	0.0	True
Gold	Platinum	-0.8303	0.4078	1.0	False
Gold	Silver	2.4722	0.0137	0.082	False
Platinum	Silver	1.9839	0.0477	0.2865	False

Figure A.5: Comparison of Transactions Count across Card Types

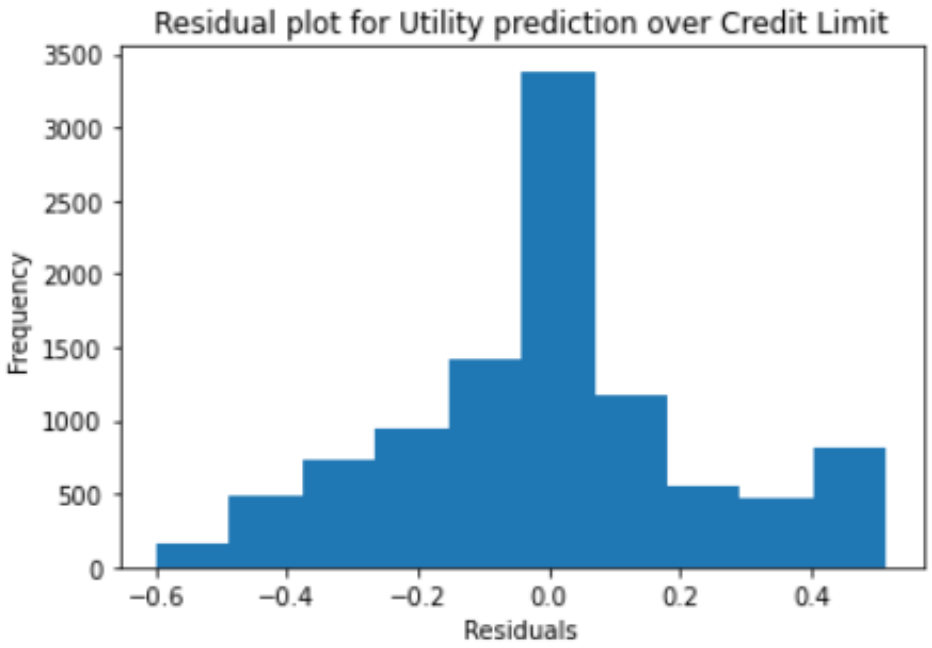


Figure A.6: Residual plot for Utilization prediction per Credit Limit

Table A.1: Summary of single attribute hypothesis tests for attrition prediction

Hypothesis	Test Results	Conclusion
No difference in customer attrition between male and female customers	Chi-Square Statistic = 14.07 p-value = 0.000	Attrition is not independent of gender.
No difference in the mean products used between existing and attrited customers	Mean (attrited) = 3.28 Mean (existing) = 3.91 t-statistic = -15.27 p-value = 0.000	Attrited customers did use less products than existing customers
No difference in the mean number of months inactive between existing and attrited customers	Mean (attrited) = 2.69 Mean (existing) = 2.27 t-statistic = 15.52 p-value = 0.000	Attrited customers have a longer duration of inactivity than existing customers
No difference in the mean number of contacts between existing and attrited customers	Mean (attrited) = 2.97 Mean (existing) = 2.36 t-statistic = 21.02 p-value = 0.000	Attrited customers had a higher number of contacts with the bank than existing customers
No difference in customer attrition based on marital status	Chi-Square Statistic = 3.95 p-value = 0.139	Attrition is independent of marital status

No difference in customer attrition by education level obtained	Chi-Square Statistic = 11.26 p-value = 0.046	Attrition is dependent on the education level
No difference in customer attrition by Age group	Chi-Square Statistic = 6.53 p-value = 0.163	Attrition is independent of age
No difference in customer attrition by credit card type	Chi-Square Statistic = 2.23 p-value = 0.525	Attrition is independent of credit card type
No difference in customer attrition by number of dependents of the customer	Chi-Square Statistic = 9.48 p-value = 0.092	Attrition is independent of number of dependents
No difference in customer attrition by the total number of bank products held by the customer	Chi-Square Statistic = 284.07 p-value = 0.000	Attrition is not independent of the total number of products held.
No difference in customer attrition by the total number of contacts made by the customer in the last 12 months	Chi-Square Statistic = 584.64 p-value = 0.000	Attrition is not independent of the total number of contacts made.
No difference in customer attrition by income category	Chi-Square Statistic = 12.75 p-value = 0.013	Attrition is not independent of the income category.
No difference in the mean revolving balance between existing and attrited customers	Mean (attrited) = 672.82 Mean (existing) = 1256.60 t-statistic = -27.44 p-value = 0.000	Attrited customers had a lower mean revolving balance than existing customers
No difference in the mean number of transactions between existing and attrited customers	Mean (attrited) = 44.93 Mean (existing) = 68.67 t-statistic = -40.25 p-value = 0.000	Attrited customers had a lower mean number of transactions than existing customers
No difference in the mean transaction amount between existing and attrited customers	Mean (attrited) = 3095.03 Mean (existing) = 4654.66 t-statistic = -17.21 p-value = 0.000	Attrited customers had a significantly lower mean transaction amount than existing customers

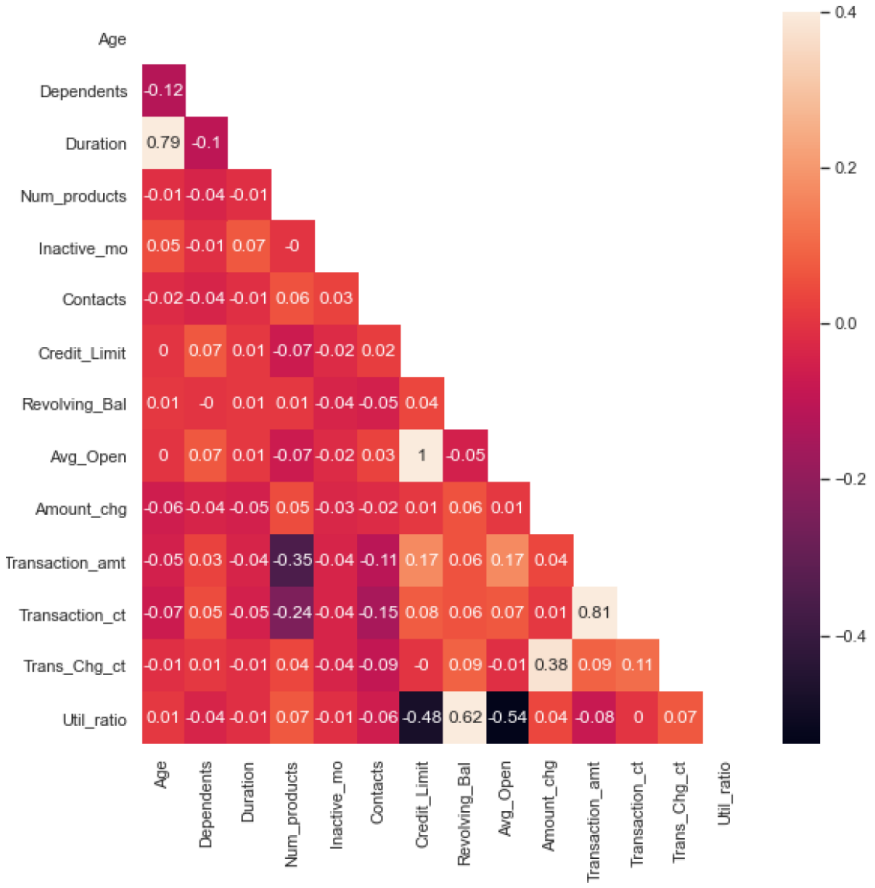


Figure A.7: Initial Correlation Heatmap (all predictors)

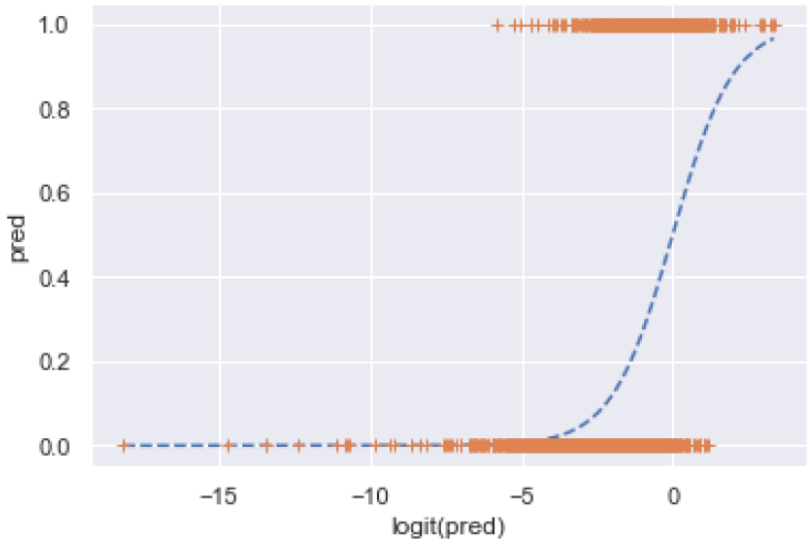


Figure A.8: Logit(p) function with actual response (attrition)

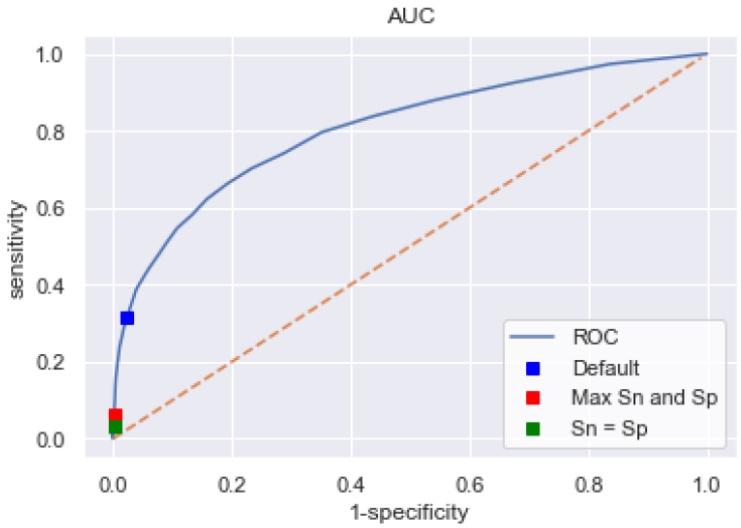


Figure A.9: ROC Curve with 3 critical points (Default, Youden's J and Balance point)