BankChurners Report

Wuji Shan

12/10/2021

Abstract

Credit card has taken up a significant part of people's lives in the current society. The "Credit Card Customers" dataset is data of the consumer credit card portfolio of a bank, including 10,000 customers with their age, gender, income, marital status, education level, and etc. The manager wants to know the reason behind customer attrition.

This report aims to deal with this customer attrition problem. I made exploratory data analysis and built a multilevel model, which uses utilizing Education_Level as the varying intercept and Gender as the varying slope. This report are consisted of 5 main parts: Introduction, Method, Result, Discussion, and Limitation. Other explorations besides what in Method part are all put in Appendix.

Introduction

Customer churning is a serious situation faced by many corporations and organizations. It has become a significant point that how to deal with this issue and keep customers. To achieve that goal, one of the main focuses of corporations may be detecting the reasons why customers have made the churning decision via analyzing past data, so that they could take effective actions to prevent the customer leaving situation better.

The dataset I choose for this project is published on Kaggle: Credit Card customers – BankChurners Dataset, which includes 10127 observations and 23 variables. After data cleaning and processing, I will make exploratory data analysis from various angles, and use a multilevel model to see what and how attributes may influence the choice of customers to churn from the bank credit card service.

Method

Data Cleaning and Processing

After looking into the data, I have finished the following processing steps to prepare for the next step of EDA and model fitting:

1. Subset down to 8 variables and factor all categorical variables;

2. Removed customer observations with "unknown" answer of categorical variables;

3. Split the data into 2 groups based on customer type: Existing and Attrited Customer.

After my data cleaning and processing, the data set utilized later has 7081 observations and 10 variables besides customer type.

Exploratory Data Analysis

Correlation Analysis

To find which and how features affect customer attrition, first, I'm interested in detecting the overall relationship between variables and the attrition flag of existing & attrited customer group after subseting down the data set. Here, I choose to plot a correlation funnel graph to show whether they are correlated with attrition. Three of variables chosen are numerical, and five of them are categorical.

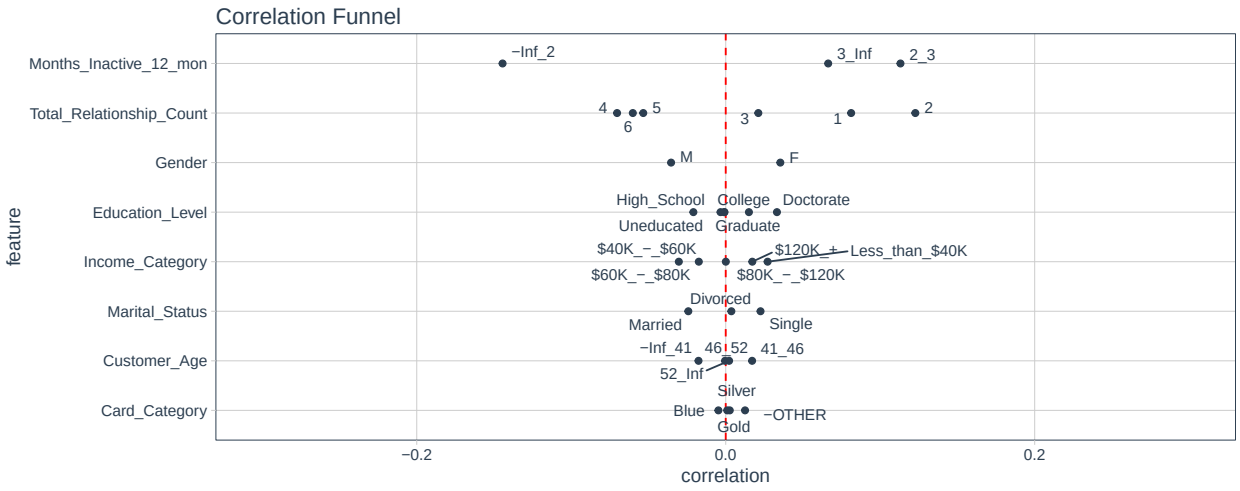


Figure 1: correlation analysis

Via investigating these 8 variables as a correlation funnel graph, we can observe the order of their importance to churning options is from the top to bottom. x-axis going larger than 0 represents that the variable type is more likely to make the choice of attrition; on the contrary, x-axis smaller than 0 represents that they are more likely to exist.

Level of Total Relationship Count & Inactivity

Inactive months and total relationship count are the top 2 variables in the correlation funnel, so I want to investigate them to prepare for the comparison of visualization and model fitting.

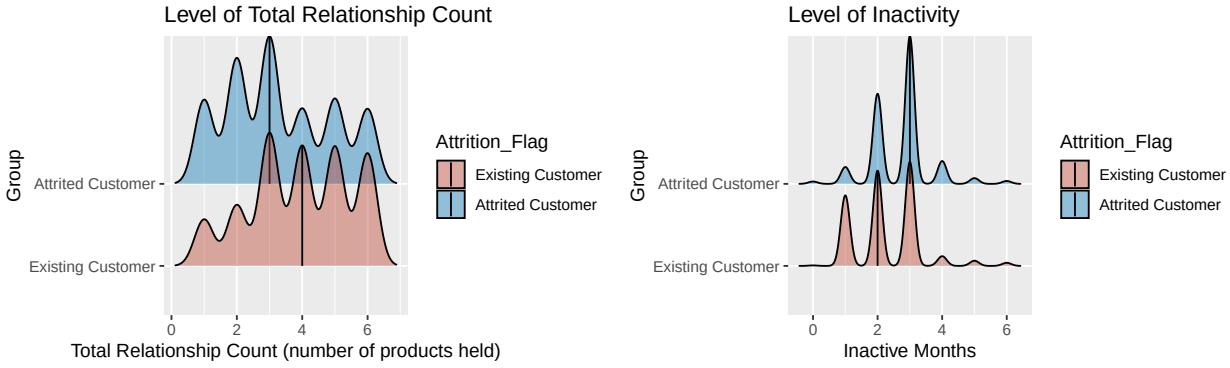


Figure 2: Level of Total Relationship Count & Inactivity

These 2 plots in figure 3 shows that Attrition Customers have lower relationship count (products held) and higher levels of Inactivity.

Education Level & Income category

Education level and income category are the top 2 categorical variables besides gender in the correlation funnel, so I want to investigate them to prepare for the comparison of visualization and model fitting. Also, I will explore relationship between combination of various variables including these two and the attrition choice.

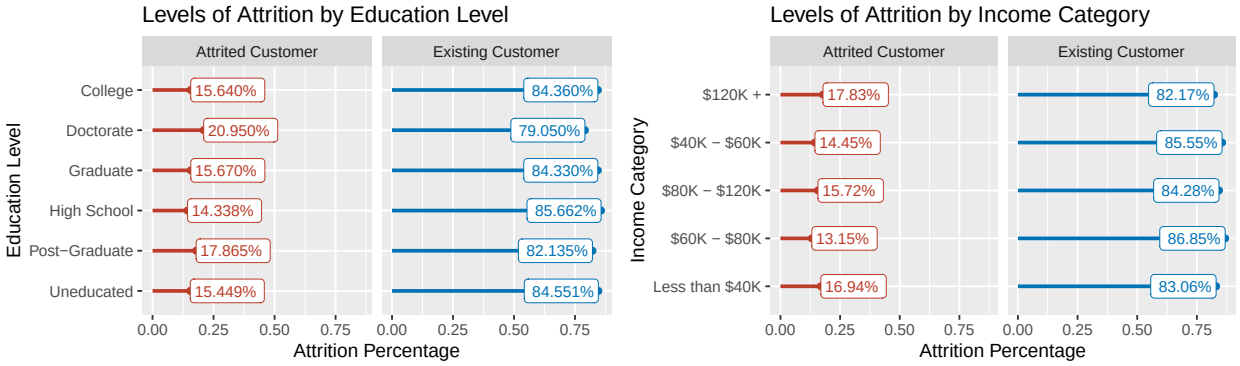
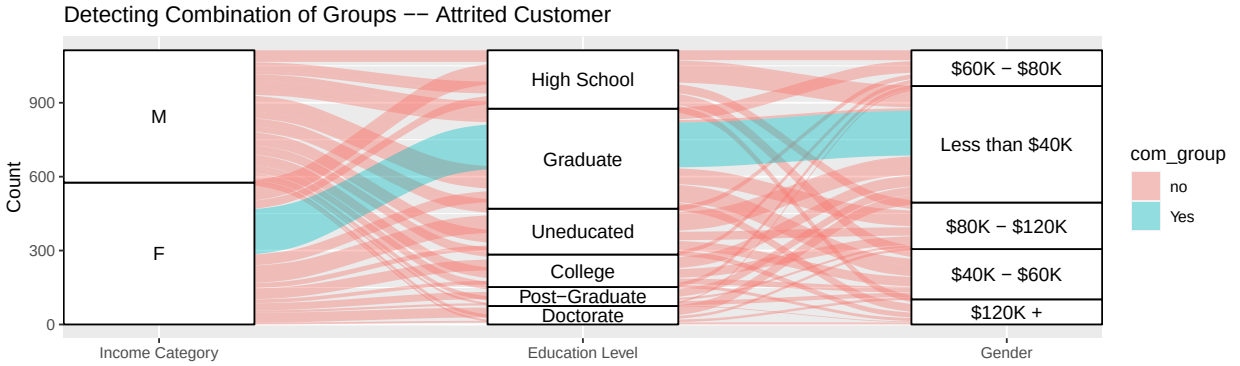


Figure 3: Education & Income Proportion Analysis

We can observe that the highest Education Level Doctorate has a highest proportion among attrition customers. The lowest proportion group is high school level, which takes up lower proportion than uneducated level. Also, we can observe that the highest income category 120K+ dollars has a highest proportion among attrition customer group. followed by Less than 40K dollars, which is the lowest income category.

Customer Attrition Combination

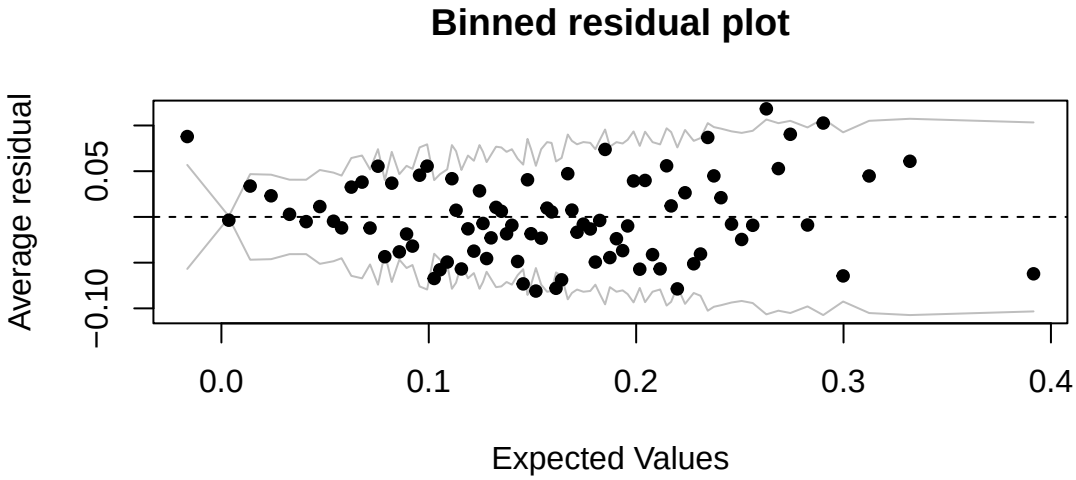
After investigating individual variable, I'm wondering whether I can find a combination of variables which are more likely to make churning options. From Figure 1, we can see that Gender, Education_Levl, and Income_Category are three categorical variables having largest correlation, so we choose to first combine them.



We can observe that Female Graduates in the low income category take up the most proportion of Attrited customer group.

Model Fitting

I used a multilevel model to fit the data, utilizing Education_Level as the varying intercept and Gender as the varying slope. I have attempted several multilevel models to get better fitting, including 'lmer' and 'glmer' (which is in the appendix). Comparing their results, I finally chose 'stan_glmer' because the response of this data set is binary, which needs to use logistic methods. The binned residual plot shows that 'stan_glmer' is a better model fit here as well, because most points fall inside the confidence bands, and there is not a distinctive pattern to the residuals. Additionally, I removed some variables which owns coefficients too small to identify and compare.



Result

Model Coefficients

Because many variables are fit in the model, I only choose to take one example formula for card type category:

$AttritionFlag = 0.0007 - 0.0253 * (GoldCard) - 0.0059 * (SilverCard) + 0.0483 * (PlatinumCard)$

The coeffcient of Card_CategoryGold and Card_CategorySilver are both negative, but that of Card_CategoryPlatinum is positive, which represents that in average, customers owning gold or silver cards are more likely to stay but customers with platinum cards are more liekly to choose to churn from the bank's credit card service.

Fixed Effects

Variable	Estimate	s.d.	t value
Customer_Age	-0.0002	0.0005	-0.434
Marital_StatusSingle	0.0136	0.0088	1.570
Marital_StatusDivorced	0.0126	0.0159	0.812
Income_CategoryLess than $40K	0.0173	0.0187	0.656
Income_Category$80K - $120K	0.0266	0.0148	1.784
Income_Category$40K - $60K	0.0032	0.0168	-0.053

Variable	Estimate	s.d.	t value
Income_Category$120K +	0.0452	0.0180	2.462
Card_CategoryGold	-0.0253	0.0399	-0.641
Card_CategorySilver	-0.0059	0.0184	-0.304
Card_CategoryPlatinum	0.0483	0.1099	0.501
Total_Relationship_Count	-0.0344	0.0027	-12.516
Months_Inactive_12_mon	0.0558	0.0043	13.138

Random Effects

Type	Intercept	GenderM
High School	-0.007	0.008
Graduate	-0.009	0.032
Uneducated	-0.002	0.008
College	0.001	-0.002
Post-Graduate	0.008	0.006
Doctorate	0.006	0.047

Discussion

From the fitting results of the multilevel model, we can see some variables types are statistically significant, including Income_Category$120K +, Total_Relationship_Count, and Months_Inactive_12_mon – effects not that much but there exist. The model results are similar to what I've mentioned before in the first two parts of EDA - correlation analysis and level of Total Relationship Count & Inactivity, which are really top 2 important variables affecting response most both in the EDA and model fitting results. Most variables state positive impacts on attrition choice.

When it only comes to findings of this report, the bank could take some actions to prevent their customer attrition better. One is to put more efforts on lower income group. Although their average spending seems not much significant, their population amount takes up nearly 40% of all customers in this data set. The other is focusing on customers' activity level. We can observe that the longest inactive months of attrited customers is 6 months. During these 6 months, the bank can detect the customer's inactivity and take actions like calling to do market research, introduce new products, or invite them to bank's promotions to arouse consumers' interest.

One interesting finding is both EDA and model results show that customers with platinum cards are more likely to leave the service. Platinum cards are generally related to customers with more products held and active transactions, which are both more correlated to existing customers, but card category of plstinum shows the contrary correlation, meaning that the results of customer churning situation may have been also affected by some other factors not included but worth being explored in the future analysis.

Limitation

Limitation of the model process is that the use of variables is not enough based on the whole number of 23 variables, because I only chose 8 of them to investigate. Also, there exist two variables Trans_Count_Amt and Trans_Count_Ct, which are really two most correlated variables when doing funnel analysis. However, each time I tried to add them into model fitting via 'glmer' and 'stan_glmer', the binned residual plot all shows a distinctly strange pattern and most of points fall outside the confidence bands. In the future analysis, I will add more features to do research and find better methods to make well-fitting model.

Reference

1. Goyao, Sakshi, Credit Card customers, Kaggle,
   https://www.kaggle.com/sakshigoyal7/credit-card-customers

2. Introducing Correlation Funnel - Customer Churn Example,
   https://cran.r-project.org/web/packages/correlationfunnel/vignettes/introducing_correlation_funnel.html

3. ggpubr: Publication Ready Plots, STHDA,
   http://www.sthda.com/english/articles/24-ggpubr-publication-ready-plots/81-ggplot2-easy-way-to-mix-multiple-graphs-on-the-same-page/

4. Interpreting Residual Plots to Improve Your Regression, qualtrics
   https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

5. Brunson, Jason Cory, Alluvial Plots in ggplot2, 2020.12.04,
   https://cran.r-project.org/web/packages/ggalluvial/vignettes/ggalluvial.html

Appendix

8 Variables Utilized:

Variable	Explanation
Customer_Age	Customer's Age in Years
Gender	M = Male, F = Female
Education_Level	Educational Qualification of the account holder
Marital_Status	Married, Single, Divorced, Unknown
Income_Category	Annual Income Category of the account holder
Card_Category	Type of Card (Blue, Silver, Gold, Platinum)
Total_Relationship_Count	Total no. of products held by the customer
Months_Inactive_12_mon	Number of months inactive in the last 12 months

Distributions of all numeric variables

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Figure 4: Distributions of all numeric variables

Pie Charts of Attrition_Flag

Model Fitting

```
fit_BankChurners_glmer <- glmer(Attrition_Flag ~ Customer_Age + Marital_Status + Income_Category +
                          Card_Category + Total_Relationship_Count + Months_Inactive_12_mon +
                          (1 + Gender|Education_Level),
                        data = BankChurners)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Attrition_Flag ~ Customer_Age + Marital_Status + Income_Category +
##     Card_Category + Total_Relationship_Count + Months_Inactive_12_mon +
##     (1 + Gender | Education_Level)
##    Data: BankChurners
##
```

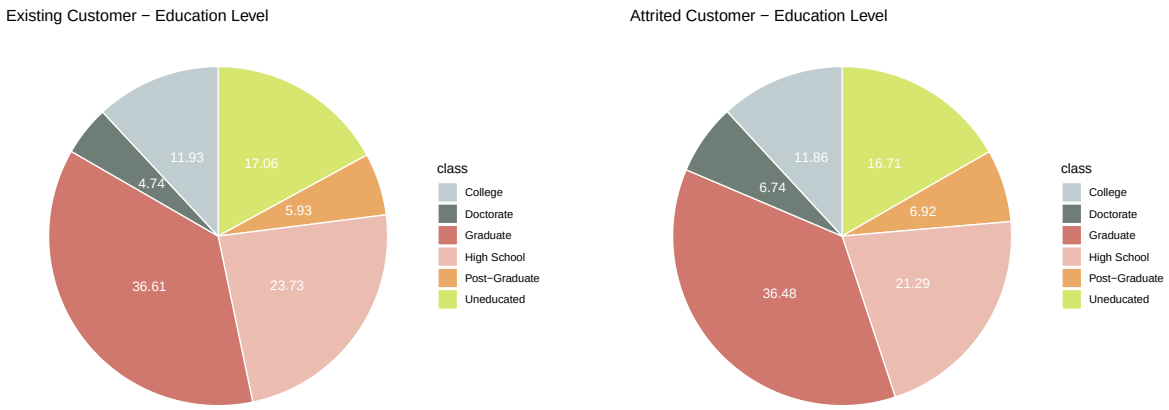


Figure 5: Education Level Proportion Comparison

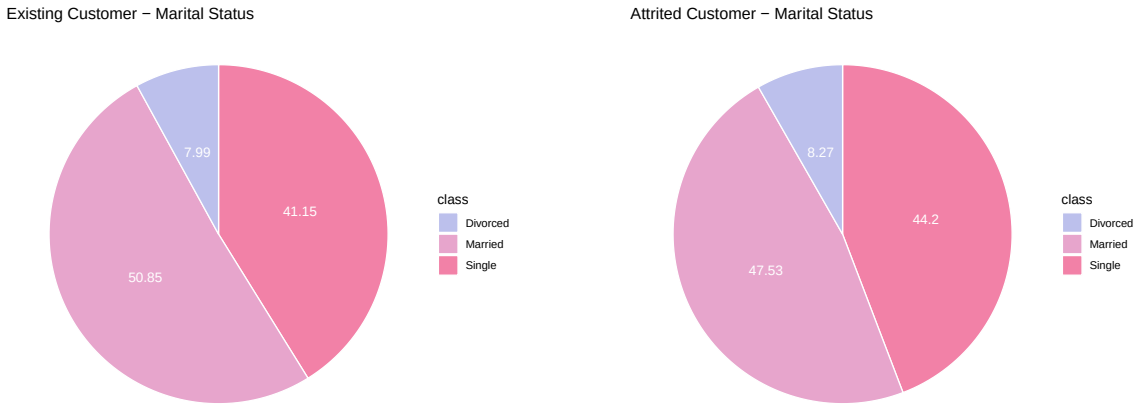


Figure 6: Marital Status Proportion Comparison

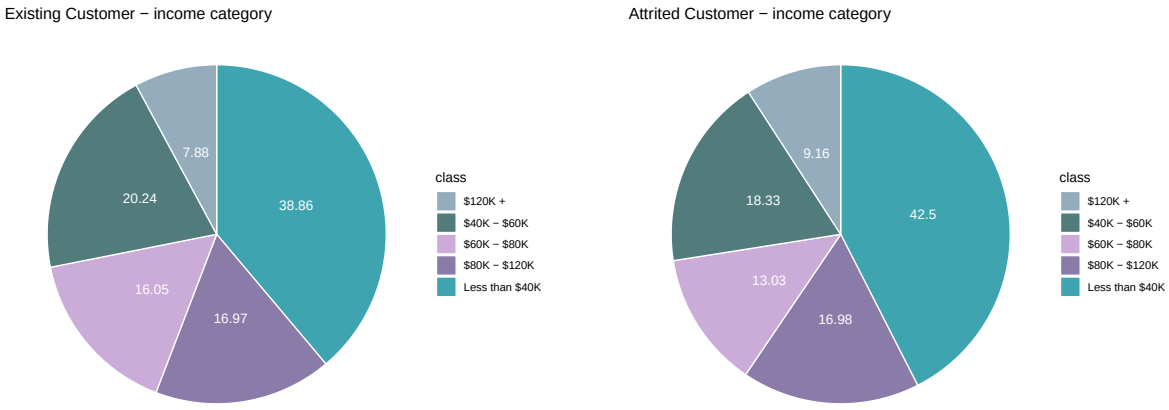


Figure 7: Income Category Proportion Comparison

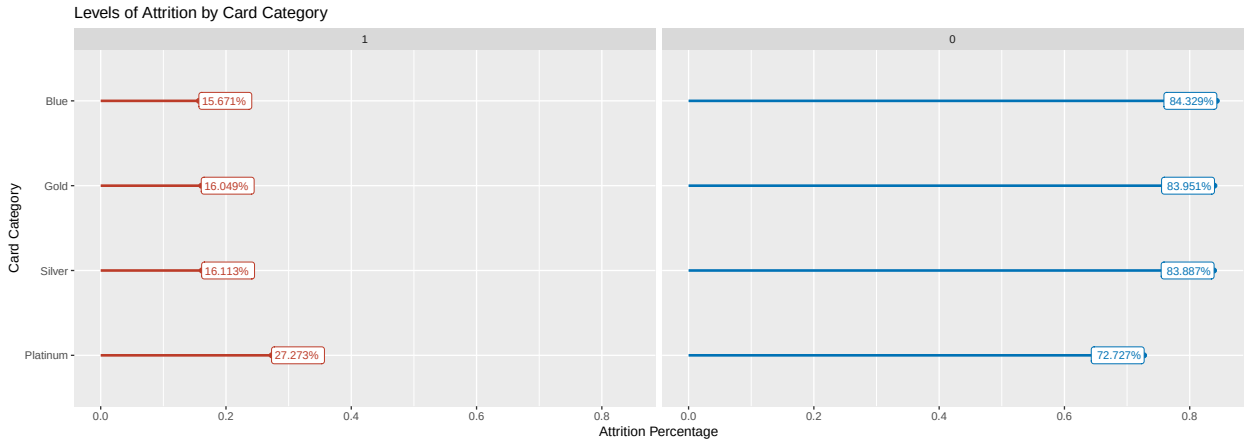


Figure 8: Income Category Proportion Analysis

```
## REML criterion at convergence: 5533.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.4905 -0.5323 -0.3605 -0.1546  2.9657
##
## Random effects:
##  Groups          Name        Variance  Std.Dev. Corr
##  Education_Level (Intercept) 0.0001333 0.01155
##                  GenderF     0.0019602 0.04427  -0.71
##  Residual                    0.1262070 0.35526
## Number of obs: 7081, groups:  Education_Level, 6
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                    0.1441811  0.0308318   4.676
## Customer_Age                  -0.0002291  0.0005283  -0.434
## Marital_StatusSingle           0.0139273  0.0088697   1.570
## Marital_StatusDivorced         0.0130482  0.0160771   0.812
## Income_CategoryLess than $40K  0.0109470  0.0166904   0.656
## Income_Category$80K - $120K    0.0264508  0.0148308   1.784
## Income_Category$40K - $60K    -0.0008346  0.0157814  -0.053
## Income_Category$120K +         0.0451765  0.0183462   2.462
## Card_CategoryGold             -0.0255873  0.0398868  -0.641
## Card_CategorySilver           -0.0056539  0.0186066  -0.304
## Card_CategoryPlatinum          0.0538268  0.1074256   0.501
## Total_Relationship_Count      -0.0343826  0.0027471 -12.516
## Months_Inactive_12_mon         0.0559023  0.0042551  13.138

##
## Correlation matrix not shown by default, as p = 13 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it

## (Intercept)
```

```
##       0.144
```

```
##                (Intercept) GenderF
## High School         -0.007   0.014
## Graduate            -0.013   0.043
## Uneducated          -0.004   0.015
## College              0.002  -0.003
## Post-Graduate        0.002   0.016
## Doctorate           -0.010   0.080
```