The casual inference between high credit limit and keeping existing credit card customers

Huyun Li - 1005964618

December 17, 2021

Abstract

In every country, there are so many commercial banks that lead to a large competition among banks. In Canada, the big five banks are Royal Bank, the Bank of Montreal, Canadian Imperial Bank of Commerce, the Bank of Nova Scotia, and Toronto-Dominion Bank. Besides, there are many other small banks. Customers now have more choices on banks to open their credit card account. Banks are also finding ways to attract new customers and retain old customers to prevent them from switching to other banks. This paper is trying to help bank managers to find factors and ways to keep customers. The Propensity Score Matching is used in the paper based on the observational data trying to find the casual inference between credit limit and retaining old customers.

Keywords: Experiment, Machine Learning Algorigthm, Survival Analysis.

Propensity Score Matching, Casual Inference, Observational Study, bank improvement, credit card service

Introduction

A number of studies show that people prefer credit cards with higher credit limit because they can make large purchases efficiently and earn more rewards like cash back, points, or travel miles [9]. People tend to not regard credits in credit cards as their own money [5], so they are more likely to spend more with high credit limit. When they spend more, they will get higher credit score which increases rewards they receive from banks [9]. This report is going to look at the effect of factor credit limit on if customers will stay in credit card service. The research question is that high credit card limit have a casual inference on if they will keep the relationship with banks or not. The analysis helps bank managers to find solutions and prevent their customers from ending their relationship or being attracted by other banks.

The Propensity Score Matching will be used to do the analysis. The dataset is collected through observational study, because it is unethical to assign customers into high credit limit and low credit limit randomly. Propensity Score Matching can do the similar thing as randomized experiments. It can redistribute data into units that receive the treatment and units that do not receive the treatment. It is proper in this case where randomized experiments are unethical.

According to research, income, total transaction and average card utilization ratio are factors that affect amount of credit limits [4]. People with higher income level tend to spend more. With more transaction, credit scores are higher, so the credit limit will increase. Lower average card utilization ratio means lower amount of money people owe, which increases credit scores and credit limit. Thus, those factors will be used to do the logistic regression model to compute the propensity score.

According to research, over 40% Canadians have a $10,000 or higher credit limit [4]. Thus, in this report, credit limit which is higher than and equal to $10,000 is defined as high credit limit, and credit limit lower than $10,000 is defined as low credit limit.

Terminology

Credit limit is the maximum outstanding customers can have on their card at one time [4].

Average card utilization ratio is the amount of money customers owe divided by the credit limit.

Keeping relationship with banks in this report means being existing customers of banks and do not leave credit card services.

Attrited Customers are customers who leave credit card services.

Hypotheses

The hypotheses is that high credit limit (credit limit is larger than $10,000) causes people to be more likely to stay in credit card service.

Introduce the report

In data part, the data "Credit Card Customers" gained from website Kaggle will be imported. Then the data cleaning will be done to keep variables needed for the analysis and NAs are removed. The basis analysis will be done to give readers a brief idea of the dataset. In Methods part, the steps and assumption of Propensity Score Matching will be stated. In results part, Propensity Score Matching and Logistic Regression will be used to do the analysis. First, the propensity score which is the probability of being treated (people with high credit limit) by logistic regression will be computed. Then, matching function from arm package will be used to match the observations with closest propensity scores that are not treated to each one that is treated, and observations that are not matched will be removed. Finally, the logistic regression will be used to find the casual inference between high credit limit and keeping relationship with banks.

Data

Data Collection Process

The data is called "Credit Card Customers", and it is obtained from website Kaggle. This data is collected by Sakshi Goyal. He collected the data "Credit Card Customers" through doing API. He grabbed data from dataset and information provided by the website with the URL as https://leaps.analyttica.com/home. It is a website offering data science and machine learning program for people to find business problems to solve and respective datasets to work upon. The latest update was in 2020. It consists of 10,000 customers with their demographics and credit card information, like age, salary, marital status, credit card limit, and credit card category, etc.

Data Summary

In the data "Credit Card Customers", there are 10127 observations and 23 variables. It consists of 10,000 customers with their demographics and credit card information, like age, salary, marital status, credit card limit, and credit card category, etc. The data is collected for a bank manager who is disturbed by more and more customers leaving their credit card service and want to find a way to provide customers with better service. In this report, I would like to use variables, Income Category, Credit Limit, Total Transaction Amount, Average Utilization Ratio, Attrition Flag to do the analysis.

Cleaning Process

First, I select 7 variables I need to use in my analysis which are Attrition Flag, Income Category, Credit Limit, Total Transaction Amount, Average Utilization Ratio. I rename them as customer, income, credit limit, transaction, utilization ratio correspondingly. This makes it easier to get accessed to when I need to use them later.

Second, I filter the unknowns in variable income in my dataset.

Third, I make variable credit limit be numeric and create a new binary variable treatment. Values in treament are separated into 0 and 1. Credit limit lower than $10,000 is 0, and credit limit higher than $10,000 is 1.

Finally, I create a new binary variable customer binary. Values are separated into 0 and 1. Attrited Customers who leave credit card services are 0, and existing customers who are still in credit services are 1.

Important variables

Variable customer is a binary variable which contains customers who are either attrited customers and existing customers. Attrited Customers are customers who leave credit card services, and existing customers are customers who are still in credit card services.

Variable income is the income level of credit card customers. It is classified into Less than $40K, $40K - $60K, $60K - $80K, $80K - $120K, $120K +.

Variable credit limit is the maximum outstanding customers can have on their card at one time.

Variable transaction is the total amount of transaction of each credit card customer.

Variable utilization ratio is the percent amount of money customers owe banks.

Variable treatment is a binary variable which credit limit lower than $10,000 is 0 (control group), and credit limit higher than $10,000 is 1 (treatment group).

Table 1: Numerical Summaries

	min	max	mean
credit limit	1438	34516	8456
utilization ratio	0.0	0.9990	0.2831

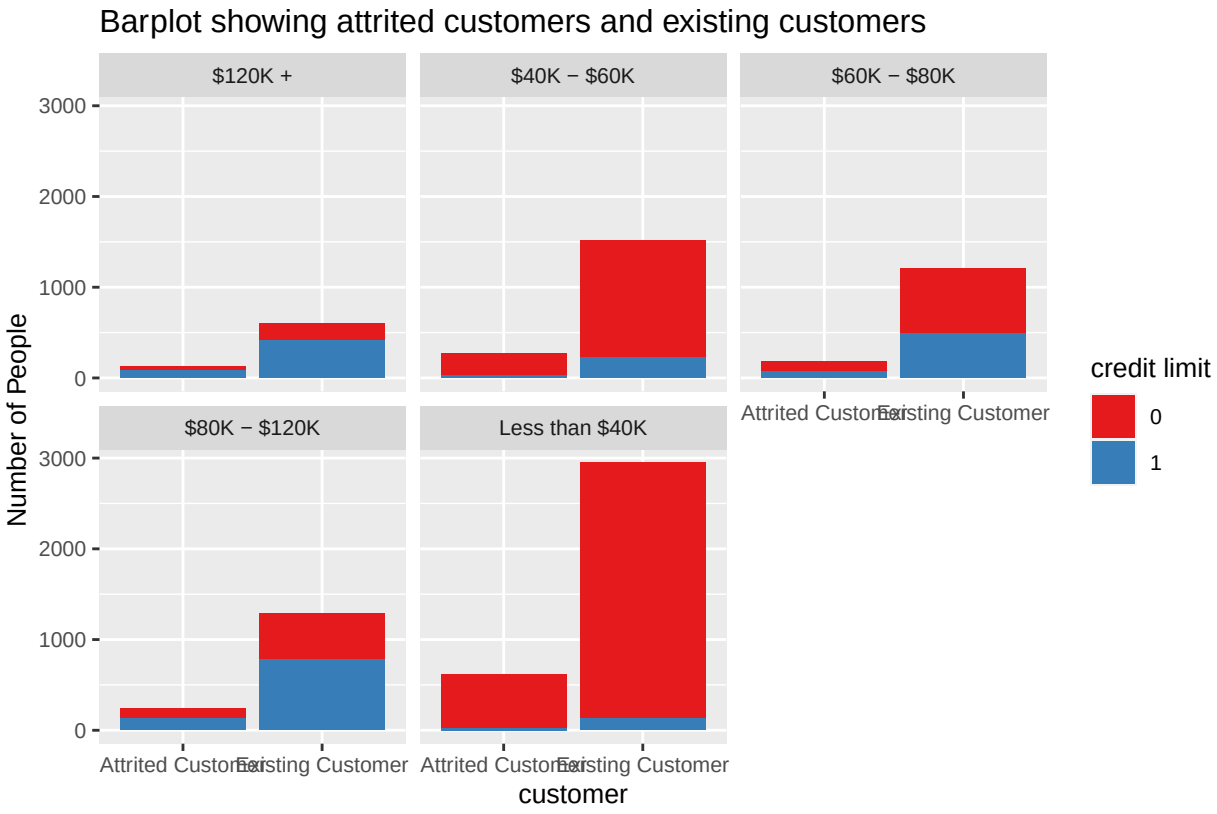
Table 1 shows the numerical summaries of numeric variables. The maximum cost of credit limit is 3.4516×10^4. The minimum cost of credit limit is 1438.3. The mean of cost of credit limit is 8522.8347643. The maximum cost of utilization ratio is 0.999. The minimum cost of utilization ratio is 0. The mean of cost of utilization ratio is 0.2815647.

Table 2: Categorical Summaries

customer	Attrited Customer	Existing Customer			
	1328	7020			
income	$120K +	$40K - $60K	$60K - $80K	$80K - $120K	Less than $40K
	680	1658	1300	1399	3311

Table 2 is the summaries of categorical variables customer and income. It shows the number of each category in each variable.

Plots



Plot 1

Plot 1 is the barplot showing the distribution of attrited customers and existing customers for high credit limit and low credit limit under different income level. This plot shows that there are fewer existing customers or attrited customers in high credit limit than in low credit limit, because total numbers of people with high credit limit is lower than total numbers of people with low credit limit. Thus, Propensity Score Matching is necessary to redistribute data into the treatment group and control group so that there are same numbers of customers in treatment group (people with high credit limit) and control group (people with low credit limit).

All analysis for this report was programmed using `R version 4.0.4`.

Methods

Method introduction

The method used for the report is Propensity Score Matching and Logistic Regression. In this report, I'm going to study the casual inference between high credit limit and if customers will stay in credit card service. However, according to the plot 1, total numbers of customers with high credit limit are different with total numbers of customers with low credit limit. Thus, we need to redistribute the data. Propensity Score Matching is used to do the distribution, because we cannot randomly distribute customers into high credit limit and low credit limit, which is unethical. The data is classified into treatment group (people with high credit limit) and control group (people with low credit limit). Logistic regression Model is used to compute the estimated probability of being treated, which is the estimated probability of people with high credit limit. The treated and controlled observations are matched based on the propensity score and observations not matched are removed from the dataset. Through Propensity Score Matching, the dataset is distributed evenly into high credit limit and low credit limit. Then Logistic Regression is used again to find the casual inference between high credit limit and if customers will stay in the credit card service.

Assumption

The sample of dataset should be large and the data should include enough numbers of both treated and untreated units [6]. There are 8,348 observations in the dataset. There are 1328 customers in control group and 7020 customers in treatment group.

The treatment factor should be relevant with other variables used in computing propensity score [6]. The logistic regression result shows that all variables are significant, so they are all relevant to treatment factor (credit limit).

The R packages used for Propensity Score Matching are broom, arm and huxtable [2].

Results

	(1)
(Intercept)	1.474 ***
	(0.107)
income$40K - $60K	-2.461 ***
	(0.125)
income$60K - $80K	-1.149 ***
	(0.117)
income$80K - $120K	-0.364 **
	(0.116)
incomeLess than $40K	-3.788 ***
	(0.132)
transaction	0.000 ***
	(0.000)
utilization_ratio	-8.634 ***
	(0.303)
N	9015
logLik	-2806.527
AIC	5627.054

*** p < 0.001; ** p < 0.01; * p < 0.05.

The logistic regression model is $log(\frac{P}{1-P}) = 1.474 - 2.461 * income\$40K - \$60K - 1.149 * income\$60K - \$80K - 0.364 * income\$80K - \$120K - 3.788 * incomeLessthan\$40K + 0.0001119 * transaction - 8.634 * utilizationratio + \epsilon$

First, I run a logistic regression model to see if all variables are relevant with treatment factor and use for propensity score computation. The logistic regression model result shows that p values of all variables

income, transaction, and utilization ratio are very small, so these variables are all significant. This follows the assumption of Propensity Score Matching that all variables should be relevant to treatment factor. I will use this model to estimate the propensity score.

Then I match treatment group with control group based on similar propensity score and remove observations in the dataset that are not matched. The glimpse of the dataset after matching is in the Appendix, A2: Materials.

After redistributing data into treatment group and control group, the percentage of high credit limit customers and percentage of low credit limit customers are almost the same. This can be seen from the plot 2 in Appendix, A3: Supplementary Plots.

	(1)
(Intercept)	1.130 ***
	(0.048)
treatment	0.632 ***
	(0.075)
N	4776
logLik	-2322.350
AIC	4648.700

*** p < 0.001; ** p < 0.01; * p < 0.05.

The final model is $log(\frac{P}{1-P}) = 1.130 + 0.632 * treatment + \epsilon$

This shows the casual inference between if people stay in credit card service and high credit limit. The treatment factor credit limit is very significant to lead to casual inference because the p value is smaller than 0.001. If people have high credit limit (credit limit is higher than $10,000), the log odd of staying in credit card service increases by 0.632. The probability of staying in credit card service increases by 0.653.

Conclusions

The goal of the report is to find if high credit limit has a casual inference on people's choice to stay or leave credit card service. The hypotheses is that high credit limit causes people to stay in credit card service. Through the Propensity Score Matching and Logistic Regression, the result is when people have high credit limit (credit limit is higher than $10,000), they have larger probability to stay in credit card service. Thus, people are more likely to keep using credit cards rather than switching to other banks if they have higher credit limit. This is reasonable because people are more likely to spend more than needed with high credit limit. If they switch to another banks and open a new account, the credit limit may not be as high as before and they cannot spend as much as before.

Weaknesses

The Propensity Score Matching is a method trying to be like a completely randomized experiment, but it is not as efficient as fully blocked randomized experiment [7]. Thus, it does not do well in controlling the change of other variables.

Also, the high credit limit is defined as credit limit that are higher than $10,000 in this report. It may cause some bias, because it is only defined in this report. In reality, it is hard to define the dividing line between high credit limit and low credit limt.

Next Steps

In this report, I only work on one factor credit limit and figure out the casual inference between credit limit and if people stay in credit card service. According to research, there are many other factors affect customers' decisions on leaving or staying. For example, some banks offer discounts on flights or offer film tickets. Some banks offer promotions or special offers when opening a new account. However, variables in dataset are limited. In the future, I will look at the casual inference on those factors if I'm able to find or collect dataset with more complete information.

Discussion

In conclusion, the dataset, Credit Card Customers from the website Kaggle is used to analyze if there is a casual inference between credit limit and staying in the credit card service. Through Propensity Score Matching and Logistic Regression, the result shows that high credit limit causes people to be more likely to stay in the credit card service with the bank. Thus, it is recommended that banks can keep in contact with customers more. They can send them emails and give them suggestions on how to increase their credit limit.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: December 15, 2021)

2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: December 15, 2021)

4. Gregorski, M. (2021). *The Average Credit Card Limit in Canada.* The Motley Fool. https://www.fool.ca/personal-finance/credit-cards/the-average-credit-card-limit-in-canada/. (Last Accessed: December 15, 2021)

5. Hardekopf, B. (2018). *Do People Really Spend More With Credit Cards?.* Forbes. https://www.forbes.com/sites/billhardekopf/2018/07/16/do-people-really-spend-more-with-credit-cards/?sh=2a4270141c19. (Last Accessed: December 15, 2021)

6. The World Bank Group. (2021). *Propensity Score Matching.* The World Bank. https://dimewiki.worldbank.org/Propensity_Score_Matching. (Last Accessed: December 15, 2021)

7. Gary, K., and Richard, N. (2019). *Why Propensity Scores Should Not Be Used for Matching.* Political Analysis, 27, 4, Pp. 435-454. https://gking.harvard.edu/publications/why-propensity-scores-should-not-be-used-formatching. (Last Accessed: December 15, 2021)

8. American Bankers Association. (2021). *Survey: Younger Customers More Likely to Switch Banks for Better Digital Features.* ABA Banking Journal. https://bankingjournal.aba.com/2021/02/survey-younger-customers-more-likely-to-switch-banks-for-better-digital-features/. (Last Accessed: December 15, 2021)

9. Fontinelle, A. (2021). *6 Benefits of Increasing Your Credit Limit.* Investopedia. https://www.investopedia.com/financial-edge/0212/6-benefits-to-increasing-your-credit-limit.aspx. (Last Accessed: December 15, 2021)

All analysis for this report was programmed using `R version 4.0.4`.

Appendix

A1: Ethics Statement

This report contains complete steps and processes of data source, data cleaning, and how the method is used to do the analysis. Thus, it is reproductive and readers can repeat the study based on this report.

The data Credit Card Customers is from the website Kaggle, and it is open data and for free. Everyone can get access to.

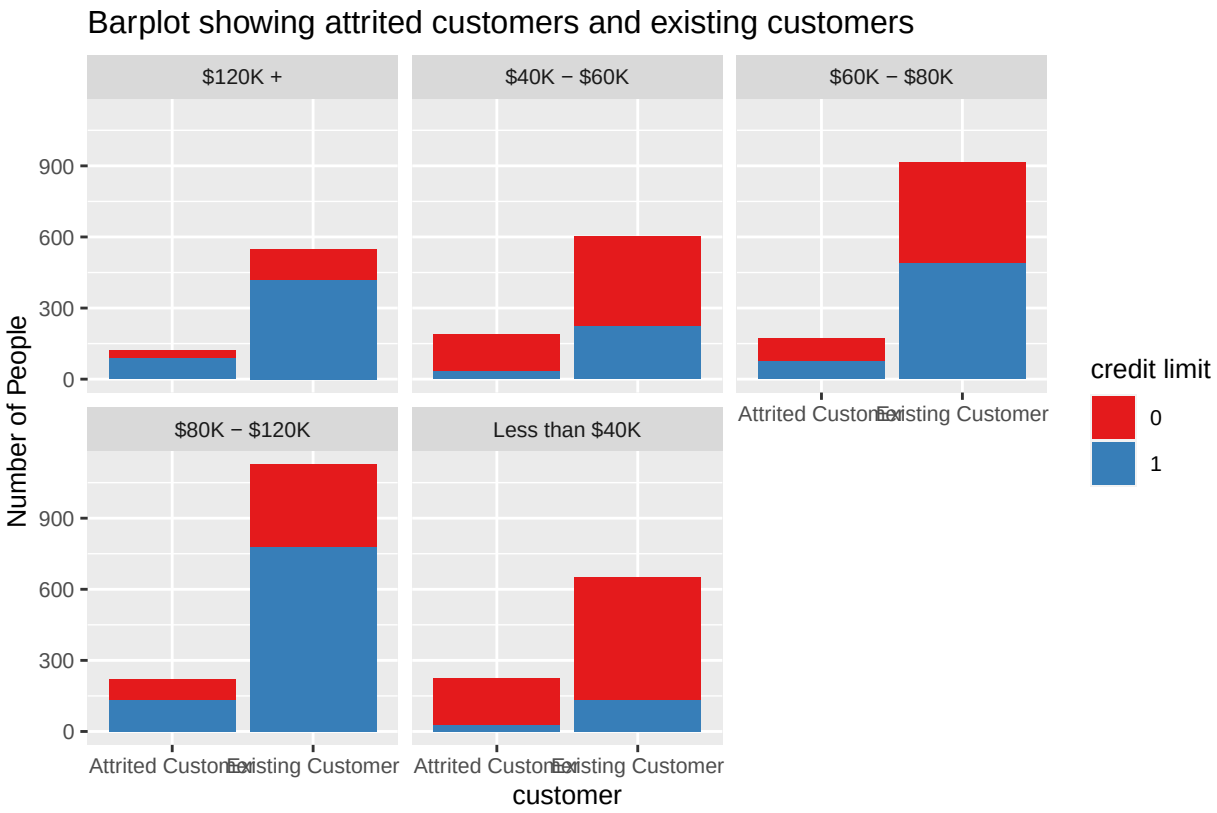
The assumption of method Propensity Score Matching is stated in Method part of this report.

All code, R package, method, source and data are cited in this report.

A2: Materials

```
## Rows: 4,776
## Columns: 11
## $ customer             <chr> "Attrited Customer", "Existing Customer", "Existi~
## $ income               <chr> "Less than $40K", "Less than $40K", "Less than $4~
## $ credit_limit         <dbl> 10195.0, 4920.0, 11261.0, 2326.0, 3968.0, 11176.0~
## $ transaction          <dbl> 1563, 1420, 1158, 5413, 4128, 2682, 2121, 13090, ~
## $ utilization_ratio    <dbl> 0.234, 0.232, 0.221, 0.276, 0.236, 0.217, 0.200, ~
## $ treatment            <dbl> 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1~
## $ customer_binary      <dbl> 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0~
## $ predicted_propensity <dbl> 0.01537309, 0.01539237, 0.01641997, 0.01643738, 0~
## $ match.ind            <dbl> 2982, 2981, 3025, 3024, 3146, 3145, 3206, 3205, 3~
## $ cnts                 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ pairs                <dbl> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9~
```

A3: Supplementary Plots



Plot 2