Do Demographic Characteristics Affect Credit Card Churn Rate?

Nguyen Y Nhi Le

12/02/2023

## Abstract

In recent years, credit cards are essential for our modern lifestyle. It brings a convenient and safe way for customers to buy almost anything with just one click or swipe both online and offline. The number of companies that provide credit card products keeps increasing and makes market competitiveness high. They provide credit card products with attractive bonuses and reward policies to attract customers from other companies. Customers who leave a credit card company are called credit card churn. It causes credit companies to lose their customers because of the cancellation or non-renewal of their accounts.

Some studies show that customer churn is a costly, significant loss for credit companies. Harvard Business Review has pointed out that a 5% increase in customers can help the companies gain from a 25% to 85% increase in profit. Because customer retention is critical and issue urgent to any credit card company, maintaining this factor is the key to reducing companies’ operation costs and increasing profit. In order to do that, credit companies have to understand the factors that affect the credit card churn of their customers. If you predict customer churn early and have a prompt solution to prevent churn, it will keep a high customer retention rate and gain more profits.

The customer demographics characteristics are essential information for customers. Understanding these factors and how they relate to customer churn will provide credit card companies insights about their customers and lead to effective strategy marketing campaigns to keep the customers with the company. This research uses a bank churners dataset from Kaggle containing multiple customer demographic variables and their churn status. We will perform multiple analyses, including exploratory data analysis and logistic regression, to find the relationship between customer churn and their demographics. The result of this research will help credit card companies focus on customer groups who will be more likely to churn their credit cards.

## Introduction

Different customer demographics include age, gender, dependent count, education level, marital status, and income can have different affect on the churn decision. For example, married credit card customers might churn more than single credit card customers because credit card companies usually offer attractive promotion for new customers. The analysis will focus to find the relationship between customer churn and customer demographics characteristics. As a result, we can understand which group of customers are more likely to churn their credit cards in the future.

## Literature Review

According to Miao & Wang (2022), if credit card customers use their card more frequently, they are less likely to leave the credit company. The authors conducted research on a dataset with more than 10,000 credit card customers and 21 feature columns to build a prediction model on the loss of credit card companies’ customers. The three important factors that affect the customers’ churn decision are the total transaction amount in the last 12 months, the total revolving balance, and the total transaction count during the previous 12 months.

In addition, Farquad, Ravi & Raju (2009) used a dataset from a bank credit card data that had a growing number of credit card customers and decided to improve their retention rate. Several reasons were mentioned during their research to predict whether their existing customers will be churning out in the future, including technology, customer service quality, geographical location, etc. As a result, the research showed that long-relationship customers are less likely to churn out credit card services in the near future. It also suggested the bank needs to improve their service in purchase journey experience and customer service rather than focusing on customer demographics.

Moreover, de Lima Lemos et al. (2022) conducted research on customer data at a large Brazilian bank. They revealed that bank customers who have a strong relationship with the bank, such as those who also use other bank products and services and engage in bank loans, will be more likely to keep loyalty to the bank.

## Research Question(s)

This research project will address the following question:

* Do customer demographic characteristics affect their decision to churn out their credit card companies?

## Theory

The null hypothesis and alternative hypothesis are:

H0: Neither customer demographic characteristics is related to credit card churn.

H1: At lease one customer demographic characteristics is related to credit card churn.

By analyzing customer demographic characteristics in related to the credit card customer churn, the alternative hypothesis are expected to be true.

## Data

This research project will use a credit card customer dataset from [Kaggle](https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers/data). The dataset has more than 10,000 customer records with 23 various features. Each record contains information about credit card customers, including age, gender, dependent count, education, income, card category, months on book, etc.

## 'data.frame': 10127 obs. of 23 variables:  
## $ CLIENTNUM : int 768805383 818770008 713982108 769911858 709106358 713061558 810347208 818906208 710930508 719661558 ...  
## $ Attrition\_Flag : chr "Existing Customer" "Existing Customer" "Existing Customer" "Existing Customer" ...  
## $ Customer\_Age : int 45 49 51 40 40 44 51 32 37 48 ...  
## $ Gender : chr "M" "F" "M" "F" ...  
## $ Dependent\_count : int 3 5 3 4 3 2 4 0 3 2 ...  
## $ Education\_Level : chr "High School" "Graduate" "Graduate" "High School" ...  
## $ Marital\_Status : chr "Married" "Single" "Married" "Unknown" ...  
## $ Income\_Category : chr "$60K - $80K" "Less than $40K" "$80K - $120K" "Less than $40K" ...  
## $ Card\_Category : chr "Blue" "Blue" "Blue" "Blue" ...  
## $ Months\_on\_book : int 39 44 36 34 21 36 46 27 36 36 ...  
## $ Total\_Relationship\_Count : int 5 6 4 3 5 3 6 2 5 6 ...  
## $ Months\_Inactive\_12\_mon : int 1 1 1 4 1 1 1 2 2 3 ...  
## $ Contacts\_Count\_12\_mon : int 3 2 0 1 0 2 3 2 0 3 ...  
## $ Credit\_Limit : num 12691 8256 3418 3313 4716 ...  
## $ Total\_Revolving\_Bal : int 777 864 0 2517 0 1247 2264 1396 2517 1677 ...  
## $ Avg\_Open\_To\_Buy : num 11914 7392 3418 796 4716 ...  
## $ Total\_Amt\_Chng\_Q4\_Q1 : num 1.33 1.54 2.59 1.41 2.17 ...  
## $ Total\_Trans\_Amt : int 1144 1291 1887 1171 816 1088 1330 1538 1350 1441 ...  
## $ Total\_Trans\_Ct : int 42 33 20 20 28 24 31 36 24 32 ...  
## $ Total\_Ct\_Chng\_Q4\_Q1 : num 1.62 3.71 2.33 2.33 2.5 ...  
## $ Avg\_Utilization\_Ratio : num 0.061 0.105 0 0.76 0 0.311 0.066 0.048 0.113 0.144 ...  
## $ Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_1: num 9.34e-05 5.69e-05 2.11e-05 1.34e-04 2.17e-05 ...  
## $ Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_2: num 1 1 1 1 1 ...

We filter only customer demographics columns: Attrition\_Flag, Customer Age, Gender, Dependent Count, Education Level, Marital Status, and Income Category.

# Select only demographics variables  
df <- churn\_data %>% select(Attrition\_Flag, Customer\_Age, Gender, Dependent\_count, Education\_Level, Marital\_Status, Income\_Category)  
  
head(df)

## Attrition\_Flag Customer\_Age Gender Dependent\_count Education\_Level  
## 1 Existing Customer 45 M 3 High School  
## 2 Existing Customer 49 F 5 Graduate  
## 3 Existing Customer 51 M 3 Graduate  
## 4 Existing Customer 40 F 4 High School  
## 5 Existing Customer 40 M 3 Uneducated  
## 6 Existing Customer 44 M 2 Graduate  
## Marital\_Status Income\_Category  
## 1 Married $60K - $80K  
## 2 Single Less than $40K  
## 3 Married $80K - $120K  
## 4 Unknown Less than $40K  
## 5 Married $60K - $80K  
## 6 Married $40K - $60K

We continue to check if any missing value in the dataset:

## Attrition\_Flag Customer\_Age Gender Dependent\_count Education\_Level   
## 0 0 0 0 0   
## Marital\_Status Income\_Category   
## 0 0

There are no missing values in the dataset.

After filtering out only necessary variables, we recognize that Attrition\_Flag, Gender, Education Level, Marital Status, and Income are categorical variables. Therefore, we convert these variables to factors to make it more easier to analyze data.

# Change categorical columns to factor  
df[-c(2, 4)] <- lapply(df[-c(2, 4)], as.factor)  
  
# We also rename Customer\_Age to Age and Income\_Category to Income  
colnames(df)[colnames(df) == "Customer\_Age"] = "Age"  
colnames(df)[colnames(df) == "Income\_Category"] = "Income"  
  
# Show data structure  
summary(df)

## Attrition\_Flag Age Gender Dependent\_count  
## Attrited Customer:1627 Min. :26.00 F:5358 Min. :0.000   
## Existing Customer:8500 1st Qu.:41.00 M:4769 1st Qu.:1.000   
## Median :46.00 Median :2.000   
## Mean :46.33 Mean :2.346   
## 3rd Qu.:52.00 3rd Qu.:3.000   
## Max. :73.00 Max. :5.000   
##   
## Education\_Level Marital\_Status Income   
## College :1013 Divorced: 748 $120K + : 727   
## Doctorate : 451 Married :4687 $40K - $60K :1790   
## Graduate :3128 Single :3943 $60K - $80K :1402   
## High School :2013 Unknown : 749 $80K - $120K :1535   
## Post-Graduate: 516 Less than $40K:3561   
## Uneducated :1487 Unknown :1112   
## Unknown :1519

Since gender data is F and M, we need to make it more readable by converting it to Female and Male.

Additionally, the Attrition\_Flag variable has only two values: Attrited Customer and Existing Customer. Hence, we transform this to a new column called Churn with Yes if Attrition\_Flag="Attrited Customer" and No if Attrition\_Flag="Existing Customer". After transformation, we drop the Attrition\_Flag column.

# Update gender values  
df <- mutate(df, Gender=as.factor(ifelse(df$Gender == 'F', "Female", "Male")))  
  
# Create new variable Churn to indicate churn status  
df <- mutate(df, Churn=as.factor(ifelse(df$Attrition\_Flag == "Attrited Customer", "Yes", "No")))  
  
# Drop Attrition\_Flag column  
df$Attrition\_Flag <- NULL  
  
summary(df)

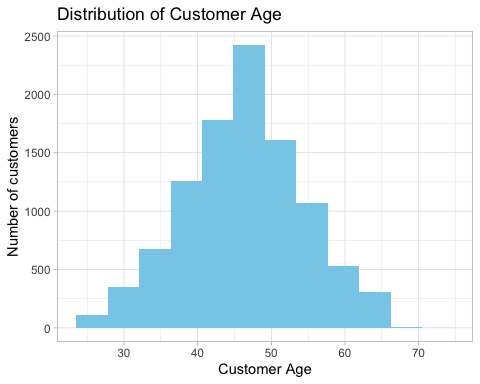
## Age Gender Dependent\_count Education\_Level  
## Min. :26.00 Female:5358 Min. :0.000 College :1013   
## 1st Qu.:41.00 Male :4769 1st Qu.:1.000 Doctorate : 451   
## Median :46.00 Median :2.000 Graduate :3128   
## Mean :46.33 Mean :2.346 High School :2013   
## 3rd Qu.:52.00 3rd Qu.:3.000 Post-Graduate: 516   
## Max. :73.00 Max. :5.000 Uneducated :1487   
## Unknown :1519   
## Marital\_Status Income Churn   
## Divorced: 748 $120K + : 727 No :8500   
## Married :4687 $40K - $60K :1790 Yes:1627   
## Single :3943 $60K - $80K :1402   
## Unknown : 749 $80K - $120K :1535   
## Less than $40K:3561   
## Unknown :1112   
##

Once the data was cleaned and dropped unnecessary columns, we start with descriptive statistics to understand the data. We can see there are 1627 churned customers and 8500 non-churned customers.

For numeric variables Customer Age and Dependent Count, we calculate the summary statistics and use the histogram to display the data spread.

### Customer Age

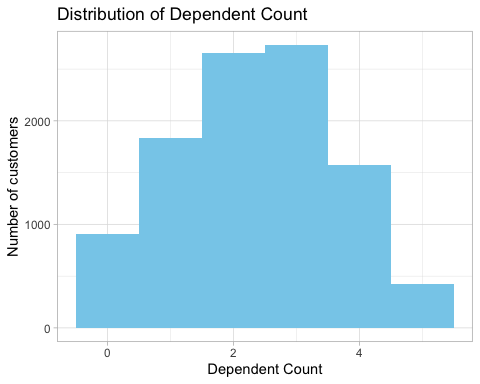
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 26.00 41.00 46.00 46.33 52.00 73.00



The histogram shows the distribution of credit card customers by age. The customer age range is from 26 to 73, with mean = 46.33 and median = 46. The largest bucket of credit card customers are within the age range of 40 to 50.

### Dependent Count

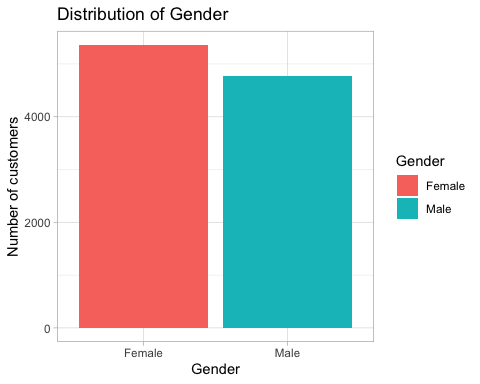
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.000 2.000 2.346 3.000 5.000



The histogram shows the distribution of credit card customers by dependent count. The range of dependent count is from 0 to 5, with mean = 2.346 and median = 2. The largest bucket of credit card customers has 2 or 3 dependents.

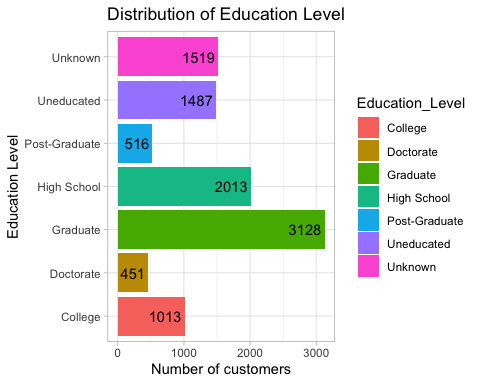
For categorical variables, we count the frequency distribution and use bar charts to understand the distribution of different categories in each variable.

### Gender



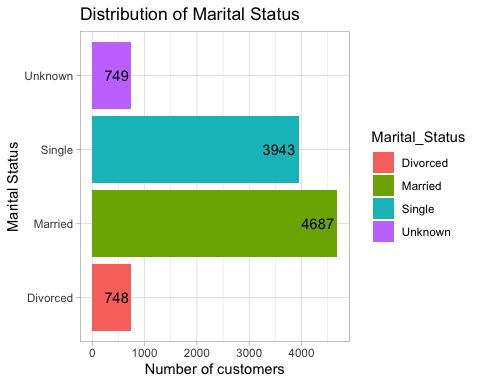
There are more female customers than male customers. The difference between female and male customers is slight, so it won’t significantly impact the data analysis. We can say the gender variable is uniformly distributed.

### Education Level



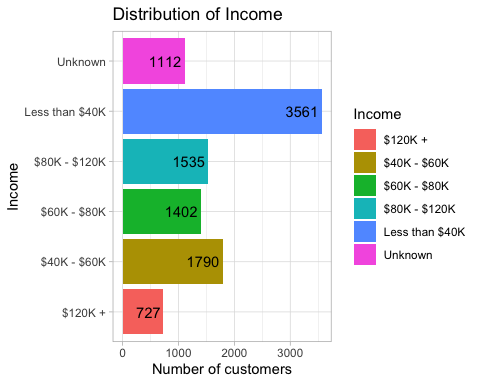
The bar chart shows the distribution of credit card customers by education level. The largest number of credit card customers have graduate degrees. This means graduate people are more likely to have credit cards than people of other education levels.

### Marital Status



The bar chart shows the distribution of credit card customers by marital status. The largest number of credit card customers are married. This means married people are more likely to have credit cards than other marital statuses.

### Income



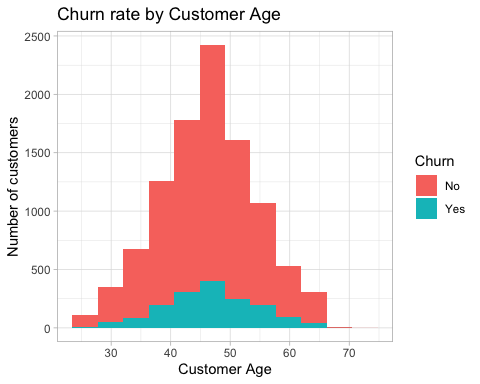
The bar chart shows the distribution of credit card customers by income. The largest number of credit card customers have incomes less than $40K. We can easily understand this because low-income customers might need to use credit cards for their daily expenses or short-term borrowing from credit card companies.

## Methodology

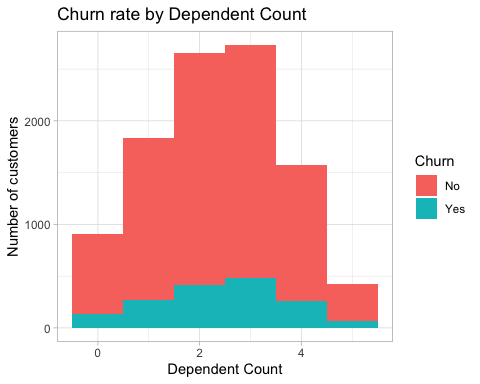
After data was cleaned in the pre-processing steps, it will be used to perform exploratory data analysis to identify the relationship between credit card churn rate and demographics variables.

### Exploratory Data Analysis

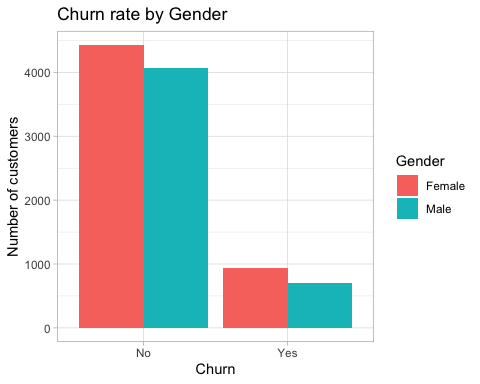
The numeric variables are plotted on histogram plots to show the churn rate.



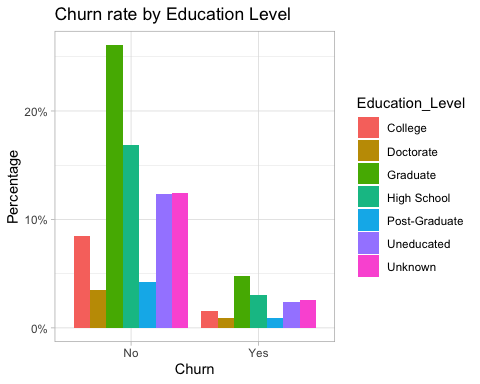
Customers within the age range of 45-55 are more likely to have credit cards and become credit card churners.



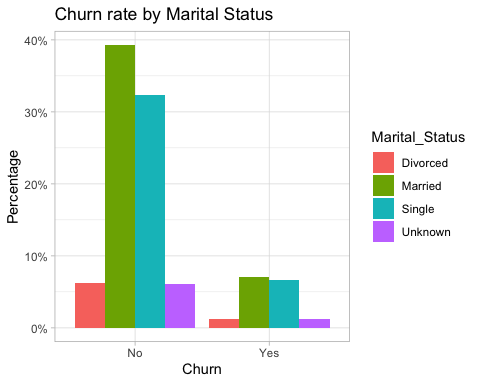
Although the churn rate shows no difference according to dependent count, customers with 2 or 3 dependents contribute to the highest churn rate.

Categorical variables are plotted on bar plots to compare the churn rate and find the largest group of customers more likely to become credit card churners. 

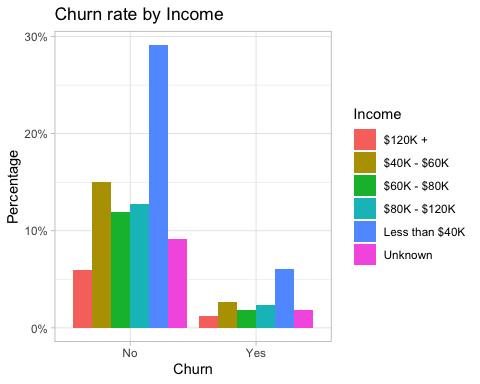
The churn rate for female customers is slightly higher than for male customers. This can be a factor for credit card companies to focus on their advertising campaigns to increase customer retention.



The largest proportion of credit card customers have graduate degrees, and this group of customers also has the highest churn rate.



We can see that most credit card customers are either single or married. The highest churn rate falls among married customers.



Credit card customers with income less than $40K are more likely to have credit cards and then become churners.

### Logistic Regression

Because the dependent variable Attrition is binomial (YES/NO), multiple logistic regression models will be built to test the relationship between customer churn and demographic variables. The first model is built with all independent variables:

churn.glm.all <- glm(Churn~., data = df, family = "binomial")  
summary(churn.glm.all)

##   
## Call:  
## glm(formula = Churn ~ ., family = "binomial", data = df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.702012 0.267776 -6.356 2.07e-10 \*\*\*  
## Age 0.006664 0.003503 1.902 0.05714 .   
## GenderMale -0.332441 0.107179 -3.102 0.00192 \*\*   
## Dependent\_count 0.042133 0.021308 1.977 0.04800 \*   
## Education\_LevelDoctorate 0.384507 0.145293 2.646 0.00813 \*\*   
## Education\_LevelGraduate 0.029270 0.100657 0.291 0.77121   
## Education\_LevelHigh School 0.006753 0.107529 0.063 0.94993   
## Education\_LevelPost-Graduate 0.204230 0.144876 1.410 0.15863   
## Education\_LevelUneducated 0.049814 0.112838 0.441 0.65888   
## Education\_LevelUnknown 0.116304 0.111421 1.044 0.29657   
## Marital\_StatusMarried -0.089166 0.107816 -0.827 0.40822   
## Marital\_StatusSingle 0.051747 0.108403 0.477 0.63311   
## Marital\_StatusUnknown 0.072576 0.139115 0.522 0.60188   
## Income$40K - $60K -0.346161 0.136493 -2.536 0.01121 \*   
## Income$60K - $80K -0.285071 0.125671 -2.268 0.02331 \*   
## Income$80K - $120K -0.102957 0.120731 -0.853 0.39378   
## IncomeLess than $40K -0.305408 0.147982 -2.064 0.03903 \*   
## IncomeUnknown -0.346414 0.164010 -2.112 0.03467 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8927.2 on 10126 degrees of freedom  
## Residual deviance: 8878.0 on 10109 degrees of freedom  
## AIC: 8914  
##   
## Number of Fisher Scoring iterations: 4

We can see the Marital\_Status variable is insignificant (p-value > 0.05). So we build the second model after dropping the Marital\_Status.

churn.glm.two <- glm(Churn~Age  
 + Gender  
 + Dependent\_count  
 + Education\_Level  
 + Income, data = df, family = "binomial")  
summary(churn.glm.two)

##   
## Call:  
## glm(formula = Churn ~ Age + Gender + Dependent\_count + Education\_Level +   
## Income, family = "binomial", data = df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.703653 0.252871 -6.737 1.61e-11 \*\*\*  
## Age 0.006295 0.003500 1.799 0.07208 .   
## GenderMale -0.329456 0.107165 -3.074 0.00211 \*\*   
## Dependent\_count 0.041390 0.021284 1.945 0.05182 .   
## Education\_LevelDoctorate 0.384620 0.145256 2.648 0.00810 \*\*   
## Education\_LevelGraduate 0.028270 0.100614 0.281 0.77873   
## Education\_LevelHigh School 0.006264 0.107475 0.058 0.95352   
## Education\_LevelPost-Graduate 0.203104 0.144813 1.403 0.16076   
## Education\_LevelUneducated 0.052474 0.112794 0.465 0.64177   
## Education\_LevelUnknown 0.118804 0.111357 1.067 0.28603   
## Income$40K - $60K -0.340472 0.136420 -2.496 0.01257 \*   
## Income$60K - $80K -0.283089 0.125626 -2.253 0.02423 \*   
## Income$80K - $120K -0.101232 0.120666 -0.839 0.40150   
## IncomeLess than $40K -0.298881 0.147919 -2.021 0.04332 \*   
## IncomeUnknown -0.338905 0.163982 -2.067 0.03876 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8927.2 on 10126 degrees of freedom  
## Residual deviance: 8884.6 on 10112 degrees of freedom  
## AIC: 8914.6  
##   
## Number of Fisher Scoring iterations: 4

The Age and Dependent\_count variable have p-values are slightly higher than 0.05, we can not say that these variables are insignificant. However, let’s build third model without Marital\_Status, Age and Dependent\_count.

churn.glm.three <- glm(Churn~Gender  
 + Education\_Level  
 + Income, data = df, family = "binomial")  
summary(churn.glm.three)

##   
## Call:  
## glm(formula = Churn ~ Gender + Education\_Level + Income, family = "binomial",   
## data = df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.275764 0.166777 -7.650 2.02e-14 \*\*\*  
## GenderMale -0.356530 0.106483 -3.348 0.000813 \*\*\*  
## Education\_LevelDoctorate 0.390482 0.145172 2.690 0.007150 \*\*   
## Education\_LevelGraduate 0.030372 0.100578 0.302 0.762671   
## Education\_LevelHigh School 0.006915 0.107436 0.064 0.948677   
## Education\_LevelPost-Graduate 0.203065 0.144747 1.403 0.160646   
## Education\_LevelUneducated 0.055060 0.112756 0.488 0.625331   
## Education\_LevelUnknown 0.121795 0.111315 1.094 0.273891   
## Income$40K - $60K -0.371456 0.135874 -2.734 0.006260 \*\*   
## Income$60K - $80K -0.293763 0.125480 -2.341 0.019226 \*   
## Income$80K - $120K -0.107123 0.120578 -0.888 0.374320   
## IncomeLess than $40K -0.339626 0.146930 -2.311 0.020806 \*   
## IncomeUnknown -0.379827 0.163047 -2.330 0.019830 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8927.2 on 10126 degrees of freedom  
## Residual deviance: 8890.7 on 10114 degrees of freedom  
## AIC: 8916.7  
##   
## Number of Fisher Scoring iterations: 4

## Result

Three logistic regression models were built and they show the strong, positive relationship between customer demographics and churn rate. The 3rd model is best model shows that Gender, Education Level, and Income are statistically significant impact in customer churn with p-value < 0.05. Male credit card customers are negative associated with credit card churn and this is significant with p-value < 0.01. Customers in any education level have positive associate with customer churn. Only doctorate degree is significant with p-value=0.007150 < 0.01, while other education levels are not significant because p-value > 0.05. Customers within any income category have negative associated with customer churn. This is not significant for customers within income range $80K - $120K (p-value > 0.05), while other income ranges are significant.

Therefore, the null hypothesis “Neither customer demographic characteristics is related to credit card churn” is rejected we can confirm the alternative hypothesis “At lease one customer demographic characteristics is related to credit card churn”.

## Implication

This research shows the relationship between credit card churn and various customer demographic characteristics. There are multiple other factors related to customer demographic characteristics that might affect credit card churn. These factors include card category, credit limit, total revolving balance, etc.

Further researchers should expand their interest to these factors. They can group customer demographic characteristics into multiple groups and analyze these groups and other factors. Their interest findouts can be helpful for credit card companies to have good marketing campaigns to increase customer retention rates.

## Conclusion

The research aims to answer the question: “Do customer demographic characteristics affect their decision to churn out their credit card companies?”. This research uses a dataset of credit card churners. The exploratory data analysis was conducted to show the relationship between each customer demographics variable and customer churn. Multiple logistic regression models were built to identify the relationship between independent variables and binary dependent variable Churn. The best logistic regression model showed that gender, education level, and income are the most significant credit card churn decision factors.

## References

Heskett, J. L. (2016). “Putting the Service-Profit chain to work”. Retrieved from: <https://hbr.org/2008/07/putting-the-service-profit-chain-to-work>

Miao, X., & Wang, H. (2022). “Customer Churn Prediction on Credit Card Services using Random Forest Method”. Advances in Economics, Business and Management Research.

Farquad, M., Ravi, V., & Bapi, R. S. (2009). “Data Mining Using Rules Extracted from SVM: An Application to Churn Prediction in Bank Credit Cards”. In Lecture Notes in Computer Science.

De Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). “Propension to customer churn in a financial institution: a machine learning approach”. Neural Computing and Applications, 34(14), 11751–11768.