Factors that Affect the Credit Limits

Xinxuan Lin

March 26, 2021

Abstract

A higher credit limit is always a big attraction for people to apply for credit cards. However, for two people who have similar performance on the covariates, they can be issued with different credit limits. In this report, the dataset *BankChurners* provided by S. Goyal were used to determine the association between credit limits, number of transaction, customer's gender, and their income. Based on the result of the linear regression model and the AIC backward method, there is a positive relationship between credit limits and the number of total transactions and income. Moreover, females are possible to get higher credit limits than males.

1. Introduction

As the widely use of credit cards, credit card companies have to offer credit limits more carefully as a way to prevent financial loss. However, from the customer's perspective, a higher credit limit is always a big motivation for them to use credit cards. Hence, the analysis of the dataset *BankChurners* could help them understand how credit card companies offer credit limits based on their performances, and what they could do to increase their credit limits.

The credit limit is the maximum amount of what the credit card issuer will allow the client to borrow from them (Porter et al., 2020). The initial credit limits are based on the information provided by the customers in the application form. Based on customers' performances, the credit limits would be adjusted. However, people always wondering why others can have higher credit limits when they have similar performances.

This report would determine the important factors that credit card companies would use when offering credit limits. Customers' age, gender income, the frequency of transactions, and education level were first considered to build a mulivariable linear model. These five explanatory variables were first introduced to the model, because it was the basic and common information that companies would get from customers. After getting this initial model, the AIC backward elimination method was applied to find the actual important factors out of these five covariates.

Bankchurners provided by S. Goyal on *Kaggle* was used in this report to determining the association between credit limits and mentioned variables. Information about the dataset was included in the Data section (Section 2.1); the model for prediction of credit limits was introduced in Model section (Section 2.2); Result section (Section 3) explained the relationship between credit limits and important covariates. The limitation of the study was included in Section 4. Conclusion section summarized the association between credit limits and mentioned factors.

2.1 Data

The data *BankChurners* was provided by S. Goyal on Kaggle. It contains personal information of 10,127 customers in the bank. Due to the privacy issue, the names of customers and bank were not provided. Customers were listed out by using 9 digits numbers. Hence, the analysis of this result was based on an observational study.

2.1.1 Method of Collecting Data

The data were collected based on the application form that customers submitted, and also further financial performance after they got their credit cards. The performances were recorded in the bank system.

The response of "age", "gender", "income", and "education level" were collected based on the application form that respondent completed. Further updates of the information are respondents' voluntary action. The non-response answer were recorded as "unknown" in the system.
Frequency of transaction was collected by using the bank monitoring system. Banks have a record of each transaction that customers make by using their bank accounts.

2.1.2 Variables

Education Level:
Education level is a categorical variable. It has 7 categories as following:
1.College
2.Doctorate
3.Graduate
4.High School
5.Post-Graduate
6.Uneducated
7.Unknown

Income_Category:
The variable income in the dataset is also recorded as category. It is separated into 6 different categories:
1.$120K +
2.$40K - $60K
3.$60K - $80K
4.$80K - $120K
5. Less than $40K
6. Unknown

Gender
Gender in *BankChurners* was collected in two different categories: male and female. This dataset contains 5358 females and 4769 males. The number of female respondents was more than 11% of the number of male respondents.

Credit_Limit, Customer_Age, Total_Trans_Ct
Credit limit, customer's age and the frequency of the transaction were collected as numeric variables.

Following are the graphs to show the frequency between the total number of transactions and each income category in different genders. The number of transactions is grouped under 6 intervals.

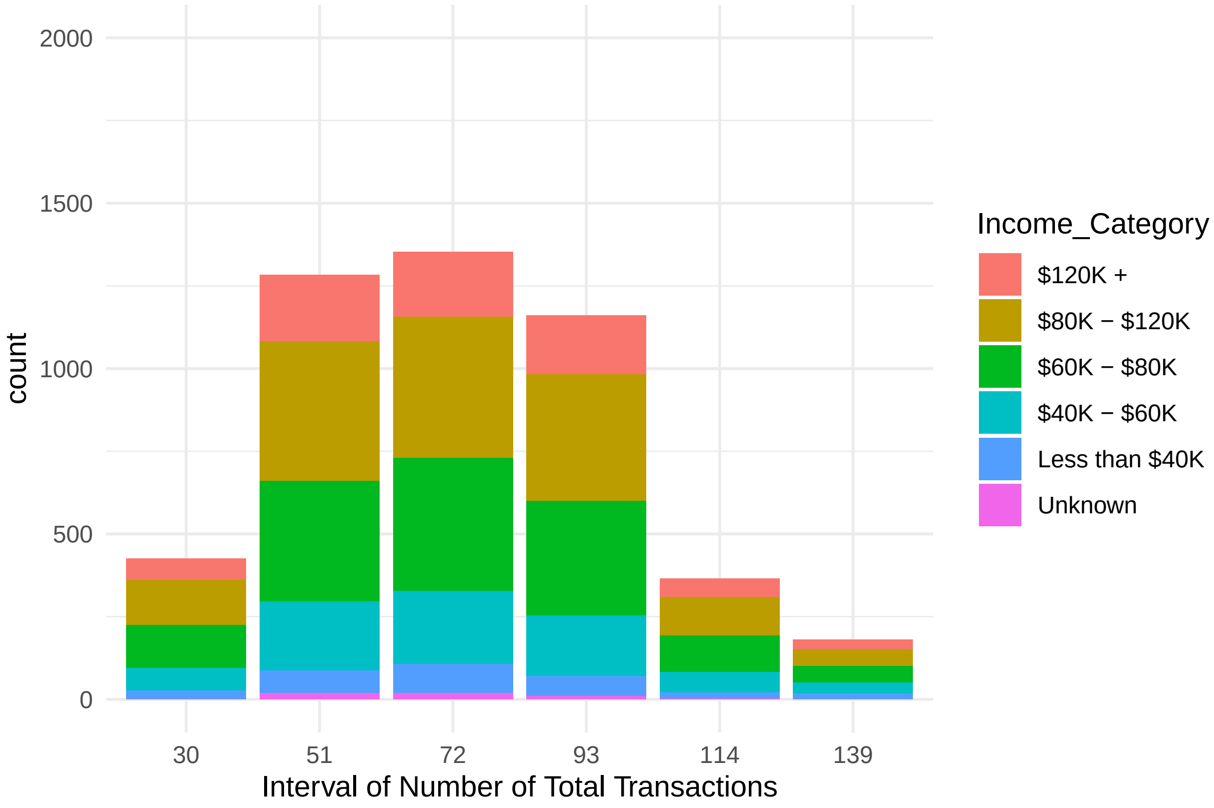

Figure 1a. Total Number of Transaction (Male)

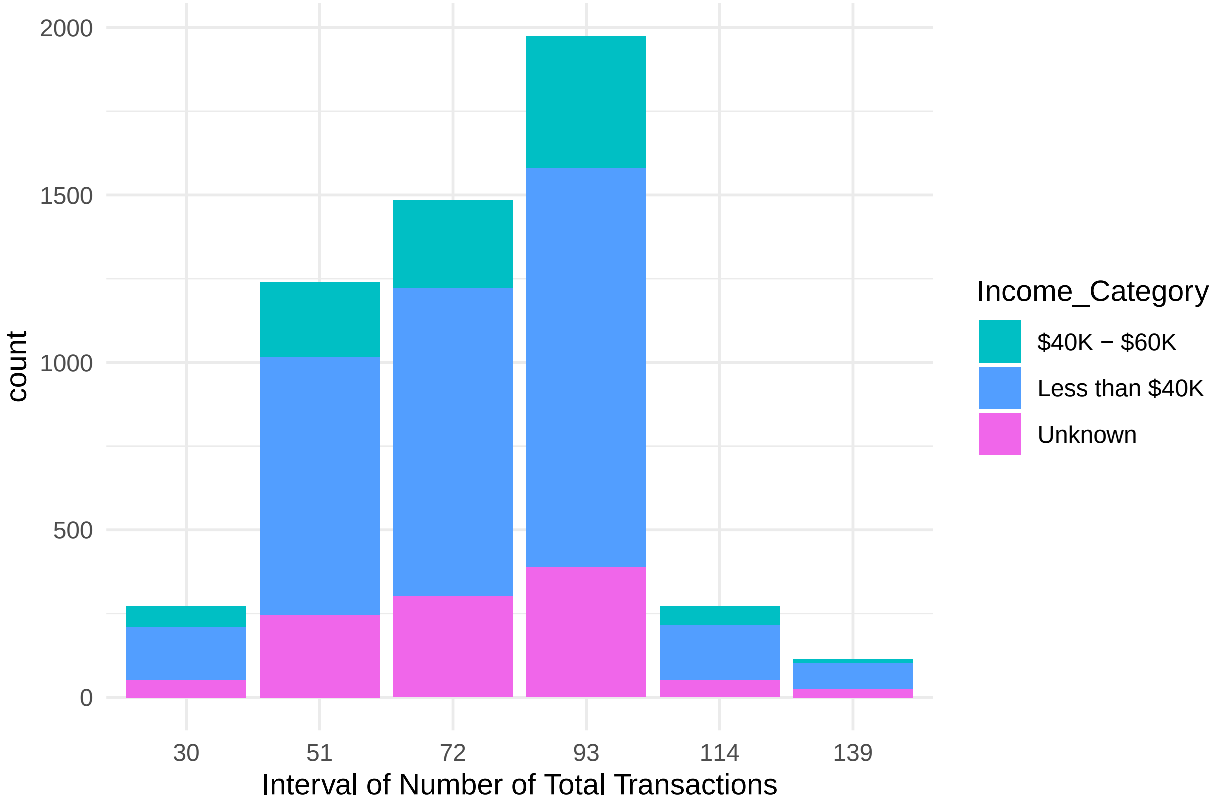

Figure 1b. Total Number of Transaction (Female)

"30": The number of total transaction is between 1 to 30;

"51": The number of total transaction is between 31 to 51;
"72": The number of total transaction is between 52 to 72;
"93": The number of total transaction is between 71 to 93;
"114": The number of total transaction is between 94 to 114;
"139": The number of total transaction is between 115 to 139.

2.2 Model

The amount of credit limit was initially predicted based on age, gender, education level, income, and the count of the transactions (which would be referred to as the initial model). These covariates were first introduced to the model because they were commonly collected from banks' customers. In other words, these covariates were the basic information that credit card companies would use to evaluate applicants.

Since the response variable (credit limit) is a continuous variable, and quantitative and qualitative variables were included as explanatory variables, a multivariables linear model would naturally be introduced to the data.

Based on Figure 1a and 1b, the distribution of count of the transaction is different in genders. Thus the count of transaction-by-gender interaction was considered to be added to the linear model.

The initial model is:

$$CreditLimit = \beta_0 + \beta_1 X_{age} + \beta_2 X_{tc} + \beta_3 X_m + \beta_{4,j} X_{edu,j} + \beta_{5,k} X_{inc,k} + \beta_6 X_{tc} X_m + \epsilon_i$$

where *Credit Limit* represents the credit limit that respondent has.

X_{age} represents the age of the respondent.

X_{tc} represents the count of the transactions.

X_m represents male customer.
$X_m = 1$ if the respondent is male, and 0 otherwise.

$X_{edu,j}$ represents the education level of the respondent. The subscript j means the corresponding *jth* education level from Doctorate, Graduate, High School, Post-Graduate, Uneducated, Unknown, and College.

$X_{inc,k}$ denotes the income category of the respondent. The subscript k implies the actual category that the respondent's income falls into. The income categories includes \$120K+, \$80K - \$120K, \$60K - \$80K, \$40K - \$60K, less than \$40K, and unknown.

ϵ_i is the independent error term.

β_0 shows the expected credit limit of the respondent who is a 0-year-old female, but graduated from college, having income more than \$120K and did not use her credit card, which is unrealistic. Thus β_0 does not have an intuitive interpretation.

β_1 is the coefficient of the variable *age*, it shows the effect of age on the credit limit. A negative β_1 implies that as age increases, the amount of credit limits would decrease by controlling other variables unchanged.

Conversely, a positive β_1 indicates that age is directly proportional to the credit limit.

β_2 is the coefficient of the number of transactions, it represents the relationship between the count of the transaction and the amount of the credit limit. In other words, a positive β_2 implies that there is a positive trend between the number of the transaction and the amount of the credit limit. Conversely. negative β_2 indicates the negative trend between them.

β_3 shows the difference in the amount of credit limit between male and female respondents. Negative β_3 means the male respondent has a lower credit limit. Conversely, a positive β_3 value indicating a higher credit limit that male respondent has than female respondent.

$\beta_{4,j}$ implies the difference in the amount of credit limit between respondent who has the corresponded(jth category in education level) education certificates and the respondent who graduated from college. Similarly, a positive $\beta_{4,j}$ shows that people who have the associated education certificates tend to have higher credit limits than people who graduated from the reference education level of college. Conversely, negative $\beta_{4,j}$ implies that people who graduated from college tend to have higher credit limits.

$\beta_{5,k}$ denoted as the difference in the amount of credit limit between people who has income in the kth interval and the reference income of above $120K. Negative $\beta_{5,k}$ implies that people who have income above $120K would have higher credit limits than those who have income in the associated interval. Conversely, positive $\beta_{5,k}$ shows that people who have income in kth interval tend to have higher credit limits than those who have income above $120K.

β_6 measures the change in size of the effect of the number of total transactions on the amount of credit limit due to the gender of customers. Negative β_6 means a negative effect of the total number of transaction on the credit limit that male has. Positive β_6 indicating males' transaction counts have a positive effect on the amount of credit limit.

Table 1: Initial Model

term	estimate	std.error	statistic	p.value
(Intercept)	17221.924216	706.057702	24.3916668	0.0000000
Customer_Age	-4.947524	9.149394	-0.5407488	0.5886926
Total_Trans_Ct	18.993717	4.603417	4.1260037	0.0000372
GenderM	-1106.001994	487.529011	-2.2685870	0.0233144
Education_LevelDoctorate	-114.723106	416.059059	-0.2757376	0.7827553
Education_LevelGraduate	131.315629	265.513785	0.4945718	0.6209132
Education_LevelHigh School	-87.030789	282.915687	-0.3076209	0.7583771
Education_LevelPost-Graduate	381.336896	397.250599	0.9599404	0.3371082
Education_LevelUneducated	236.129488	299.243403	0.7890884	0.4300789
Education_LevelUnknown	23.849398	297.931714	0.0800499	0.9361992
Income_Category$40K - $60K	-13468.331793	357.865420	-37.6351864	0.0000000
Income_Category$60K - $80K	-8940.907264	335.980863	-26.6113587	0.0000000
Income_Category$80K - $120K	-3858.021165	330.895472	-11.6593350	0.0000000
Income_CategoryLess than $40K	-14685.859459	389.258150	-37.7278149	0.0000000
Income_CategoryUnknown	-8868.888802	434.876776	-20.3940272	0.0000000
Total_Trans_Ct:GenderM	40.110431	6.254369	6.4131861	0.0000000

In order to determine the useful explanatory variables out of 6 mentioned ones, the AIC backward elimination method was applied.

The idea of the AIC backward elimination method is that starting with the model contains all predictors, the predictor that has the highest AIC value would be removed by using *R*. Then refit the model and continue to remove the variable that has the highest AIC value until the model reaches the lowest AIC value. The latest model would be considered as the final model.

Table 2: Reduced Model

term	estimate	std.error	statistic	p.value
(Intercept)	17045.29873	489.199736	34.843230	0.0000000
Total_Trans_Ct	19.02934	4.597258	4.139280	0.0000351
GenderM	-1110.71143	487.193574	-2.279815	0.0226394
Income_Category$40K - $60K	-13447.69110	356.983795	-37.670313	0.0000000
Income_Category$60K - $80K	-8931.42732	335.483718	-26.622536	0.0000000
Income_Category$80K - $120K	-3847.58603	330.505536	-11.641518	0.0000000
Income_CategoryLess than $40K	-14664.42492	388.313733	-37.764374	0.0000000
Income_CategoryUnknown	-8851.13587	433.964104	-20.396009	0.0000000
Total_Trans_Ct:GenderM	40.35028	6.250953	6.455061	0.0000000

Based on the result of AIC elimination, age and education level were removed from the model, and the final model is as following:

$$CreditLimit = \beta_0 + \beta_1 X_{tc} + \beta_2 X_m + \beta_{3,k} X_{inc,k} + \beta_4 X_{tc} X_m + \epsilon_i$$

Credit Limit represents the credit limit that respondent has.

X_{tc} represents the count of total transactions.

X_m represents male customer.
$X_m = 1$ if the respondent is male, and 0, otherwise.

$X_{inc,k}$ denotes the income category of the respondent. The subscript k implies the actual category that the respondent's income falls into. The income categories includes $120K+, $80K - $120K, $60K - $80K, $40K - $60K, less than $40K, and unknown.

ϵ_i is the independent error term.

β_0 shows the credit limit of the female respondent who graduated from college, having income more than $120K and did not use her credit card at all.

β_1 is the coefficient of the number of transactions, it represents the relationship between the count of the transaction and the amount of the credit limit. In other words, a positive β_1 implies that there is a positive trend between the number of the transaction and the amount of the credit limit. Conversely. negative β_1 indicates the negative trend between them.

β_2 represents the difference in the amount of credit limit between male and the reference gender of female. Negative β_2 means the male respondent has a lower credit limit. Conversely, positive β_2 value indicating a higher credit limit that male respondent has than female respondent.

$\beta_{3,k}$ denoted as the difference in the amount of credit limit between people who has income in the kth interval and the reference income of above \$120K.

β_4 measures the change in size of the effect of the number of total transactions on the amount of credit limit due to the gender of customers. Negative β_4 means a negative effect of the total number of transaction on the credit limit that male has. Positive β_4 indicating males' transaction counts has a positive effect on the amount of credit limit.

3. Result

Figure 1a and 1b shows different distributions of total transaction count between males and females, thus it is reasonable that gender has difference effect on count of transaction, which implies that interaction term should be included in the model.

Based on Table 1, p-value for education levels and age were greater than 5% significant level, thus all the education levels and age were not significant associated with the amount of credit limits.

The AIC backward method was applied to determine the useful predictors out of the 6 variables. Based on table 2, we could see that the p-value for gender, income, and count of transaction were all less than 5% significant level. It means that among all the mentioned covariates, only respondents' gender, income, and count of transaction were associated with the amount of credit limits.

According to Table 2, β_1 is 19.029, which means that controlling other variables, as the count of transaction increases, the amount of credit limit would also increase.

β_2 is -1110.711 shows that controlling other variables unchanged, male respondents have credit limit \$1110.71 lower than female respondents.

Negative $\beta_{3,k}$ indicates that fixing other variables, people who have higher income would have higher credit limits as well. But surprisingly, people who did not provide income information would still have higher credit limit than whose income between \$60K - \$80K.

4. Limitations

Due to the privacy issue, some information of the data was not revealed, which would raise some red flags. The observation units were only labeled as 9-digit numbers, we were not able to make sure whether the observation units were from the same household or not. If they were, then the units were not independent. For example, if one family member(other than the household head) were purchasing all the daily supplies, then she/he would have more transactions than other members; the household head could possibly have less transaction but higher income.

Moreover, as mentioned before, the data were lack of background information such as the collecting method, which could lower the accuracy of the analysis.

5. Conclusion

The prediction of credit limits was initially made based on the basic covariates: age, gender, education level, income, frequency of transaction, and transaction count-by-gender interaction. The AIC method suggested that only gender, income, frequency of transactions, and transaction count-by-gender interaction were associated with credit limits. Females tended to have a higher credit limit than males under the same conditions of other covariates. Fixing other variables unchanged, people with higher income were more possible to have higher credit limits. Moreover, under the same conditions of other covariates, as the frequency of transaction increases, credit limits would also increase.

There are still some aspects that require further research on it, such as the background information of the dataset, and the identity of the observation units.

Reference

Goyal, S. (2020, November 19). Credit Card customers. Retrieved April 2, 2021, from https://www.kaggle.com/sakshigoyal7/credit-card-customers

Porter, W., Staff, C., McClanahan, A., DeNicola, L., Hipp, D., & Proctor, C. (2020, November 16). What Is a Credit Limit and How Is It Determined? Retrieved April 2, 2021, from https://www.creditkarma.com/advice/i/what-is-a-credit-limit

Extract default color palette of ggplot2 r package (example): Hex codes. (2020, September 30). Retrieved April 2, 2021, from https://statisticsglobe.com/identify-default-color-palette-names-of-ggplot2-in-r

N/A, N. (n.d.). CRITERION-BASED PROCEDURES. Retrieved April 2, 2021, from http://www.biostat.jhsph.edu/~iruczins/teaching/jf/ch10.pdf