

FINAL PROJECT REPORT

Kevin Appiah

Northeastern University

ALY 6015 Intermediate Analytics

May 14, 2023

**Contents**

1. Introduction
2. Exploratory Data Analysis

* Descriptive statistics
* Data Visualization
* Correlation Matrix
* Scatter plots

1. Statistical Method

* Multiple Linear Regression
* Regression Results
* Hypothesis Testing

1. Conclusion
2. References

**Introduction**

This analysis aims to investigate the relationship between customer age, credit limit, revolving balance, and credit card utilization. The goal is to understand how these variables collectively influence credit card usage patterns among customers. To achieve this, a multiple linear regression model was used to estimate the effect of customer age, credit limit, and revolving balance on the average utilization ratio of credit cards. This approach offers a more comprehensive analysis compared to a simple linear regression, as it considers multiple influencing factors simultaneously, which should provide a better fit for the dataset. Initially, I tried using a simple linear regression to investigate a relationship between customer age and credit card utilization. The linear regression analysis revealed no significant relationship between customer age and credit card utilization. The model did not fit the data well, with an R-squared value of 0.000.

The dataset used for this analysis was retrieved from [www.kaggle.com/datasets/whenamancodes/credit-card-customers-prediction](http://www.kaggle.com/datasets/whenamancodes/credit-card-customers-prediction). It provides various features for credit card customers that will allow a more deatiled understanding of the factors that influence credit card utilization.

**Exploratory Data Analysis**

Before diving into modeling, an exploratory data analysis (EDA) was conducted to understand the data and its underlying structure.

1. The dataset was loaded into a Jupyter Notebook using the pandas library's read\_csv function. After successfully loading the data, the head function was used to preview the first five rows. This step provided a quick overview of the data structure, including the number and names of the columns, and the types of data stored (categorical or numerical). The dataset consists of 10127 observations and 23 columns.

A screenshot of a computer

Description automatically generated with medium confidence

**Descriptive Statistics**

The statistical description of the data was retrieved from the data set by using the describe function to gain more insight into the data such as the mean, standard deviation, minimum and maximum.

A screenshot of a computer

Description automatically generated with medium confidence

**Data Visualization**

Histograms for each variable in the dataset to gain a fair knowledge of their distributions. Special attention was paid to the independent variables ('Customer Age', 'Credit Limit', and 'Total Revolving Balance') and the dependent variable ('Average Utilization Ratio').

A picture containing diagram, text, plan, line

Description automatically generated

This is a pie chart that demonstrates the percentage of customer count by income category, the largest proportion of customers, approximately 35.2%, have an income of less than 40K. This group forms most of the customers. Making up 17.7% of customers fall in the income range of 40 to 60K. Notably, the smallest fraction of customers, just 7.2%, earn more than 120K.

A picture containing text, screenshot, diagram, font

Description automatically generated

The bar chart presents the customer count across different card categories. The most common card type, as indicated by a significantly higher bar, is the 'Blue' category, accounting for 9,346 customers and the ‘Platinum’ category has the least number of people with just 20 customers.

A picture containing text, diagram, plot, line

Description automatically generated

This histogram visualizes the distribution of customer ages in the dataset. The shape of the histogram suggests a near-normal distribution, characterized by a single peak, and similar tails on either side of the peak. This indicates that most customers fall within the middle age range, with fewer customers at the younger and older ends.

A picture containing diagram, plot, screenshot, line

Description automatically generated

Notably, this distribution is right-skewed, which means that fewer consumers have high utilization ratios and that most customers have lower utilization rates. Due to their conservative spending patterns or lower credit demands, this skewness shows that most customers do not typically use all their credit card limits.

A picture containing diagram, plot, screenshot, line

Description automatically generated

**Correlation Matrix**

The correlation matrix helped us understand which column types will serve our purpose and which do not. The column types that had many false positives usually ended up being the less reliable and unnecessary values. For the calculation of the correlation matrix, only columns with numerical values are chosen and plotted with the function corrplot ()

A screenshot of a graph

Description automatically generated with low confidence

**Scatter Plots**

To establish a correlation between variables when one of them is related to the other a scatterplot is the best place to start. Specifically, scatterplots were created for 'Customer Age', 'Credit Limit', and 'Total Revolving Balance' against 'Average Utilization Ratio'. Each scatterplot can illuminate how changes in these independent variables might be associated with changes in the utilization ratio, offering valuable insights into the dynamics of credit card usage among customers

The scatterplot between 'Total Revolving Balance' and 'Average Utilization Ratio' displays a clear upward trend, which means there is a strong positive correlation. As the revolving balance increases, points are generally higher on the y-axis, indicating a higher utilization ratio. This plot suggests that customers who maintain higher balances on their cards tend to utilize a higher percentage of their available credit.

A picture containing line, screenshot, diagram, plot

Description automatically generated

The scatterplot between 'Customer Age' and 'Average Utilization Ratio' reveals a random distribution of points, which means there is a weak correlation. There isn't any clear pattern in the plot, indicating that age does not significantly or consistently influence the utilization ratio.

A picture containing screenshot, line, diagram, text

Description automatically generated

The scatterplot between 'Credit Limit' and 'Average Utilization Ratio' shows a downward trend, reflecting a negative correlation. As the credit limit increases along the x-axis, the points tend to be lower on the y-axis, indicating customers with higher credit limits tend to have a lower utilization ratio.

A picture containing text, line, plot, screenshot

Description automatically generated

NB: Due to many data, I reduced the sample size using the sample method in pandas.

These scatter plots were plotted using a sample size of 500

**Method**

To determine the impact of customer age, credit limit, and revolving debt on the average credit card utilization ratio, a multiple linear regression model was used. In comparison to a simple linear regression, this method provides a more thorough analysis because it considers multiple variables at once, which should provide a better fit for the dataset.

A screenshot of a computer

Description automatically generated

**Regression results**

**The R-squared** value is 0.649, which means our model explains 64.9% of the variance in the average credit card utilization ratio.

**The coefficient of Customer\_Age** is -4.009e-05. This means that for each one-year increase in age, we would expect the Average Utilization Ratio to decrease by 0.00004009 units.

**The coefficient of Credit\_Limit** is -1.548e-05. This implies that for each one-unit increase in the Credit Limit, we would expect the Average Utilization Ratio to decrease by 0.00001548 units.

**The coefficient of Total\_Revolving\_Bal** is 0.0002. This suggests that for each one-unit increase in the Total Revolving Balance, we would expect the Average Utilization Ratio to increase by 0.0002 units, assuming all other variables are held constant.

**P-values** in this regression context help us to identify which variables are significant predictors of the outcome variable, which can be useful in feature selection and model interpretation. A small p-value (≤ 0.05) indicates strong evidence that the coefficient is different from zero. On the other hand, a large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis, suggesting the coefficient is not statistically different from zero.

**Hypothesis Testing**

A screenshot of a computer

Description automatically generated with medium confidence

**Conclusion**

In conclusion, the multiple linear regression analysis of the dataset revealed several interesting insights about credit card utilization patterns among customers. Our model, which included 'Customer Age', 'Credit Limit', and 'Total Revolving Balance' as independent variables, was able to explain approximately 64.9% of the variance in the 'Average Utilization Ratio', as indicated by the R-squared value.

The results suggested that both 'Credit Limit' and 'Total Revolving Balance' have a statistically significant relationship with the 'Average Utilization Ratio'.

Specifically, 'Credit Limit' was found to be negatively associated with 'Average Utilization Ratio', implying that as a customer's credit limit increases, their average utilization ratio tends to decrease. This finding may reflect the fact that customers with higher credit limits might not need to utilize a large proportion of their available credit.

On the other hand, 'Total Revolving Balance' showed a positive relationship with 'Average Utilization Ratio', indicating that customers with higher revolving balances tend to have higher average utilization ratios.

In contrast, 'Customer Age' was found to not have a statistically significant relationship with 'Average Utilization Ratio', as its p-value exceeded the significance level of 0.05. This suggests that a customer's age does not necessarily influence their credit card utilization patterns in a statistically meaningful way.

Considering these findings, it is evident that credit limit and revolving balance are important factors that could influence a customer's credit card utilization. These insights could prove valuable for financial institutions in developing strategies to manage credit risk and improve customer service.

**References**

Chauhan, A. (2022a, October 30). *Credit card customers prediction*. Kaggle. https://www.kaggle.com/datasets/whenamancodes/credit-card-customers-prediction

*Data to fish*. Data to Fish. (n.d.-a). https://datatofish.com/multiple-linear-regression-python/

rsundery. (2022, February 14). *Interpreting the results of linear regression using OLS summary*. GeeksforGeeks. <https://www.geeksforgeeks.org/interpreting-the-results-of-linear-regression-using-ols-summary/>

Frost, J. (2023, May 4). *Using histograms to understand your data*. Statistics By Jim. https://statisticsbyjim.com/basics/histograms/