Credit Score Analysis

# INTRODUCTION

According to Experian, credit is the capacity to borrow money with the understanding that it will be paid back later [1]. Credit score of an individual can be defined as a numerical value which indicates how likely they are to have the ability to pay back their debts [2]. The credit score calculated by Fair Isaac Corporation is known as the FICO Score and is used by 90 by of the top 100 financial institutions in the United States [3]. The score itself is ranged between less than 500 to greater than 800, with scores closer to 500 being regarded as poor, and closer to 800 as being good [3]. Different models have their own criteria to calculate credit scores. The goal of this study is to use data analysis techniques on a relevant dataset along with domain knowledge to check which factors contribute significantly towards an individual’s credit score. After identifying the relevant factors, the study will analyze how variations in them affect the overall credit score. The results will be validated against existing frameworks that are used for credit score calculations to ensure that the findings are generalizable. Having information about what influences credit scores is beneficial for companies that calculate their own credit scores, to make their scoring models accurate. Moreover, accurate credit scores are beneficial for both borrowers and lenders. Firms that lend money to customers such as banks will make better informed decisions regarding granting credit to new customers or increase credit limits for existing customers [4]. Borrowers will try to improve the factors which influence credit scores to attain a higher score and increase their chances of getting loans, lower interest rates on their mortgage and access to better credit card rewards [5]. These act as sources of motivation for this study.

# ANALYTICAL QUESTIONS AND DATA

As mentioned earlier, the goal of this study is to answer the question “What factors affect an individual’s credit score?” using data analysis techniques. In order to so, this study will use domain knowledge to identify relevant features and come up with analytical research questions which will help achieve the aim of this study. According to the Federal Bank of Cleveland, factors which affect an individual’s FICO credit score are Payment history (35%), Amount owed (30%), Length of credit history (15%), How much new credit (10%), Type of credit (10%) [6]. Moreover, the Federal Bank of Kansas City states that demographic factors also influence credit scores [7]. Using this information, the analytical questions this study will try to answer are:

1. How do specific debt management features impact the likelihood of an individual achieving a good credit score?
2. What is the relationship between an individual’s payment behavior and their credit score?
3. To what extent do demographic factors such as age and income influence an individual’s credit score?

The answers to these research questions will give a fair understanding of the main factors that affect credit scores and help achieve this study’s aim. For analysis, this study will use a dataset from Kaggle [8] which has been used previously for credit score classification hence is highly suitable. The analysis plan is to identify relevant features from this dataset regarding customers’ demographic factors, their debt management and payment behavior, and use them to implement customer segmentation followed by comparing the distribution of credit scores in each segment to study the effect of these features. A limitation of this dataset is that since it a synthetic dataset, the results of this study may not be as generalizable, but validating the findings with existing work will help in this regard.

# ANALYSIS

The dataset contains information across 8 months for 12,500 customers. It was checked for null values and duplicates; none were found. Customers under the age of 18 and having a monthly balance greater than 0 but no bank accounts were dropped. ‘Credit\_history\_age’ was dropped since many customers had a credit history greater than their age. Data derivation was performed to convert relevant categorical columns to numerical. ‘Type\_of\_loan” was converted from string to list making it usable. Class column (credit\_score) consists of 3 classes: 0 (Poor), 1 (Standard) and 2 (Good).

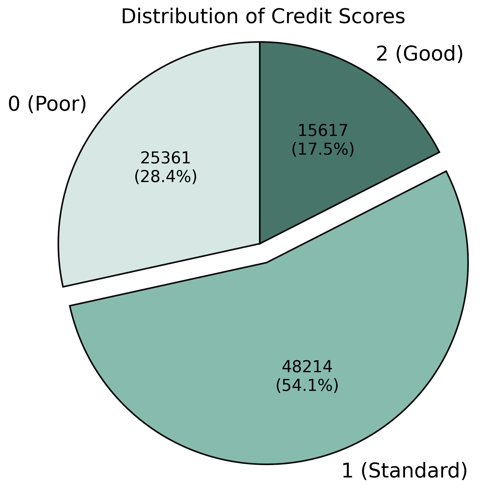


Fig 1

EDA involving univariate analysis of features and bivariate analysis with credit score was performed. Followed by ANOVA test to check which numerical features have a significant relation with credit score, and eta-squared to validate the effect size. For categorical columns, Chi-square test was done to check if they are associated with credit score followed by Cramer’s V to validate level of association. Outstanding debt, number of loans, number of credit cards are selected as debt management features and payment of just minimum amount, number of delayed payments, delay from due date are selected as payment behaviour features since they have a large effect size. Customer segmentation will be done using K Means on selected features so that the distribution of credit scores across the segments can be compared to study the effect of these features on the credit score. Data across the 8 months is aggregated using the average (for consistency) so that each datapoint represents an individual customer and segmentation can be performed.

1. *Customer Segmentation on Debt Management*

EDA and statistical tests mentioned earlier are performed again on the aggregated features prior to clustering for validation. Fig 2 shows the mean of each debt management feature decreasing as the score increases. Fig 3 shows their eta-squared values in relation to credit score.

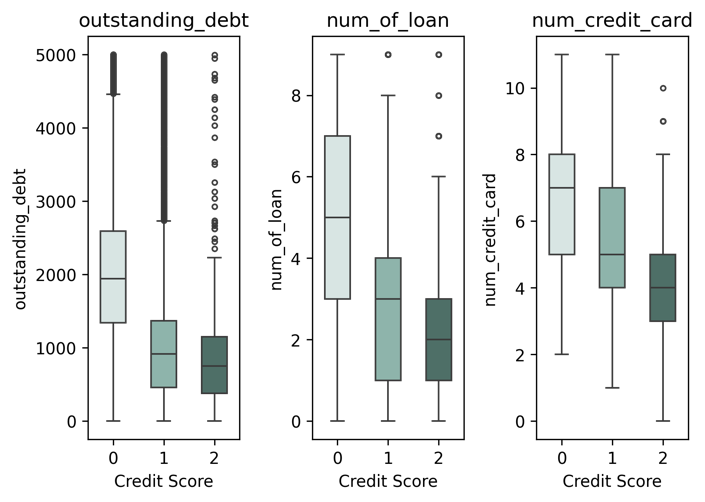


Fig 2

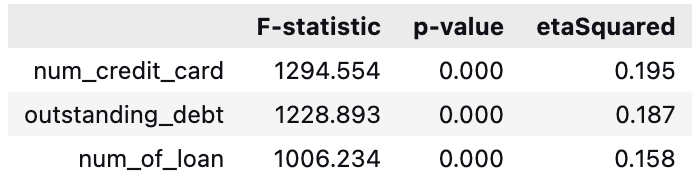


Fig 3

All three variables have a significant effect size hence clustering is performed using them as seen in Fig 4.

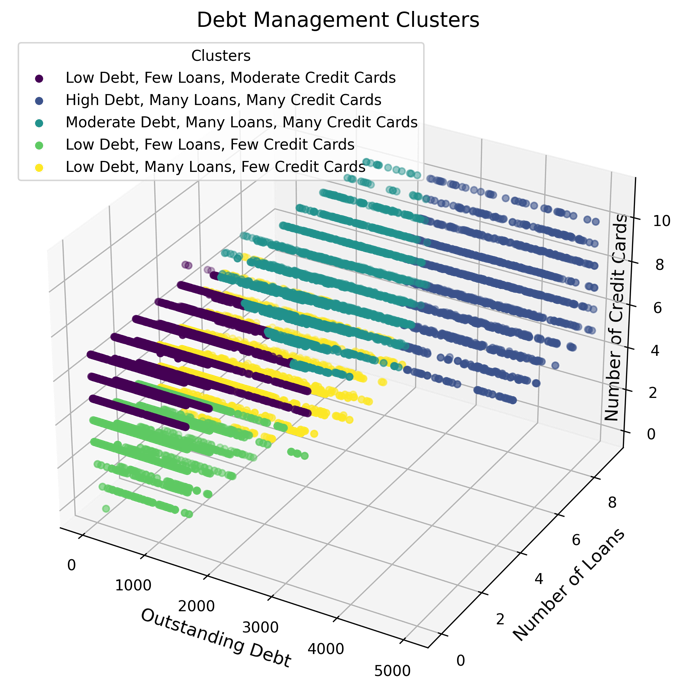


Fig 4

The ideal number of clusters is validated using silhouette plots and scores. Fig 4 shows clear separation of clusters, where each cluster has its own description for interpretation. This groups the customers into different segments and the distribution of credit scores in each segment are interpreted and compared which helps to analyze how the debt management features impact credit score.

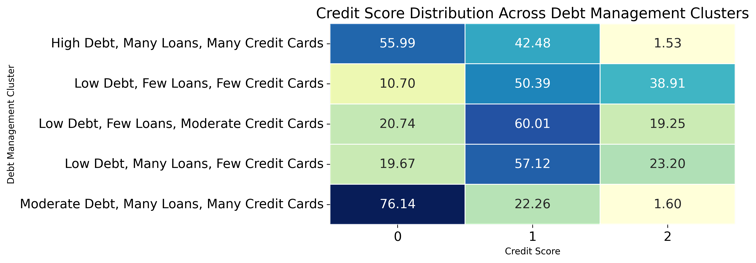


Fig 5

Fig 5 shows that low debt, few loans and credit cards cluster has the highest percentage of customers with a good credit score whereas moderate debt, many loans and credit cards has the highest percentage of poor credit score customers.

1. *Customer Segmentation on Payment Behaviour*

The means of payment behaviour features decrease as the credit score increases as seen in Fig 6.

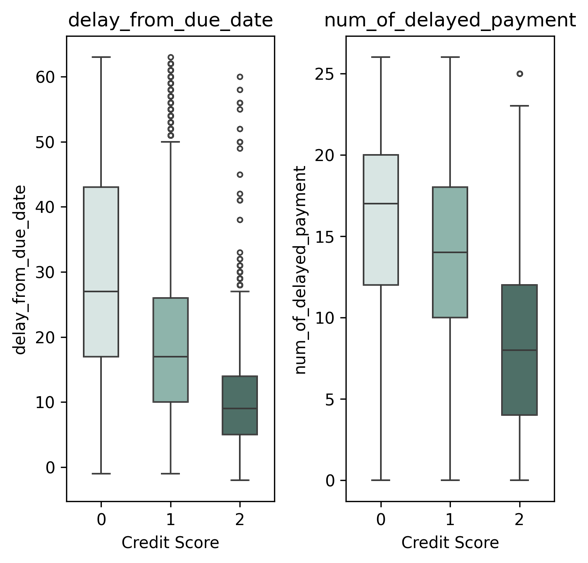


Fig 6

Fig 7 displays that customers who do not pay just the minimum amount (0) and pay more have a significantly greater percentage of good (2) credit scores.

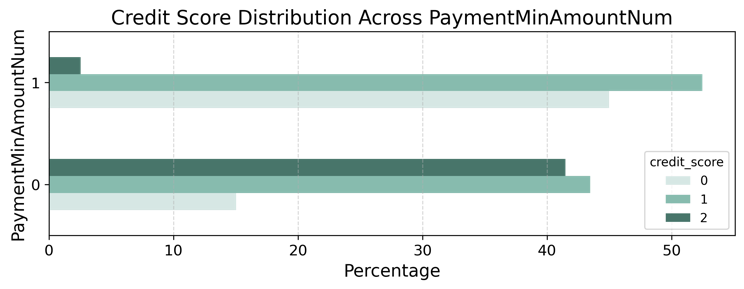


Fig 7

The effect sizes and level of association in relation to credit score shown in Fig 8 validates that all three features have a meaningful impact on credit score.

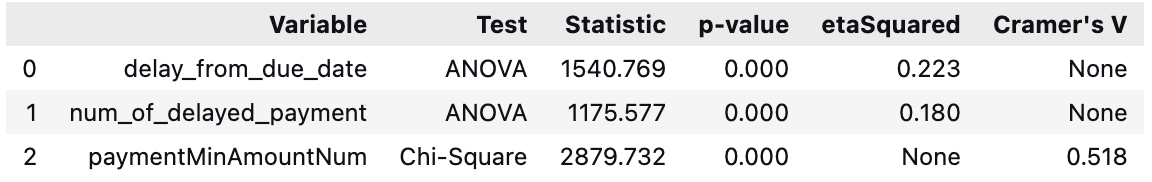


Fig 8

Performing clustering using these features segments the customers as shown in Fig 9.

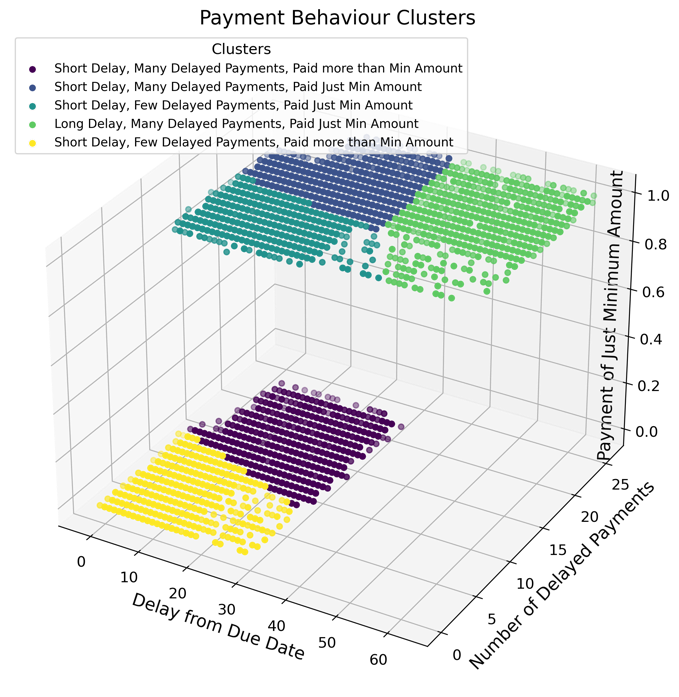


Fig 9

Silhouette plots and scores are used to validate the ideal number of clusters which ensures that the clusters in the plot are distinct and can be interpreted using their description. Comparing the distribution of credit scores in each segment will help understand the relationship between payment behaviour and credit scores.

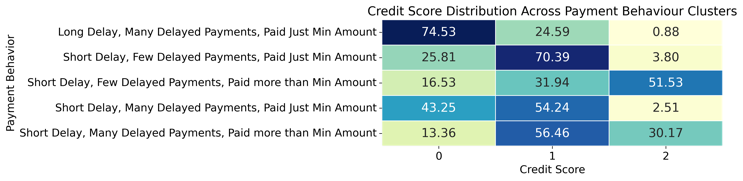


Fig 10

Fig 10 illustrates that clusters in which customers pay more than just the minimum amount have a greater percentage of good credit scores whereas the greatest percentage of poor credit scores is in the cluster where there are long delays in payment, many delayed payments and payment of just the minimum amount.

1. *Customer Segmentation on Demographic Factors*

The means of age and annual income increase as the credit score increases however, this increase is small as shown in Fig 11.

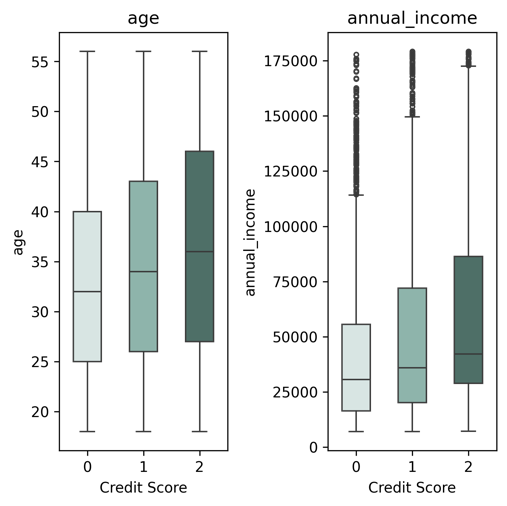


Fig 11

Fig 12 shows that their eta-squared values in relation to credit score is very small.

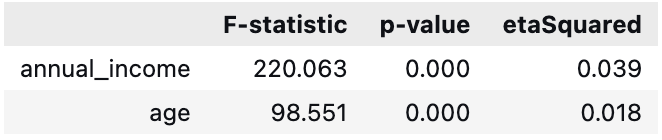


Fig 12

Although the effect sizes are small, clustering is done to explore whatever relationship there might be between demographic factors and credit score.

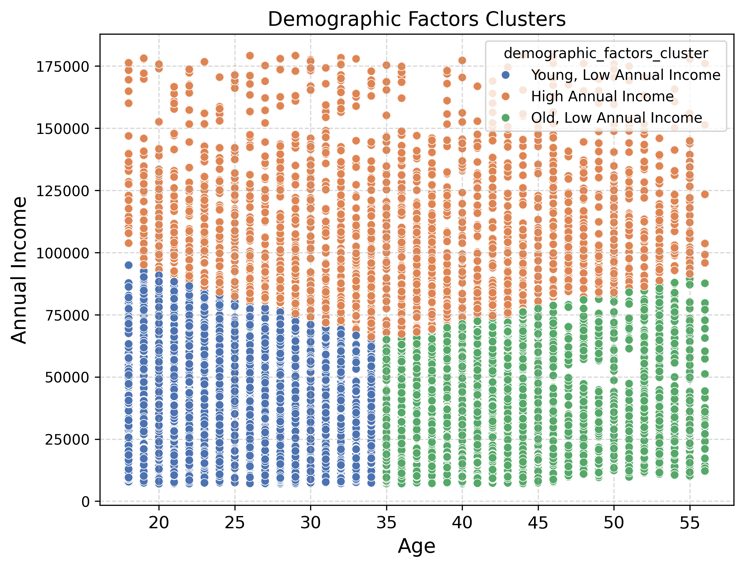


Fig 13

Fig 13 demonstrates that clustering based on demographic factors does not segment the customers very well even though 3 clusters had the highest silhouette score of 0.42. The distribution of credit scores in each segment is interpreted and compared to study the influence of demographic factors on credit scores.



Fig 14

Fig 14 shows that the cluster with high annual income has the highest percentage of good credit scores followed by old age, low income and then lastly young age, low income.

1. *Customer Segmentation on Combined Features*

The study has analyzed the influence of debt management, payment behaviour and demographic factors in isolation but in the real world credit scores are calculated while considering these factors together. Hence PCA is applied to feature engineer demographic features into one component, payment behaviour features into another component and demographic factors into another. The explained variance ratio of debt management and payment behaviour components are 0.67 and 0.69 respectively however, for demographic factors it is 0.54 thus it is replaced with the scaled weighted sum of age and income. Clustering is performed using these three components.

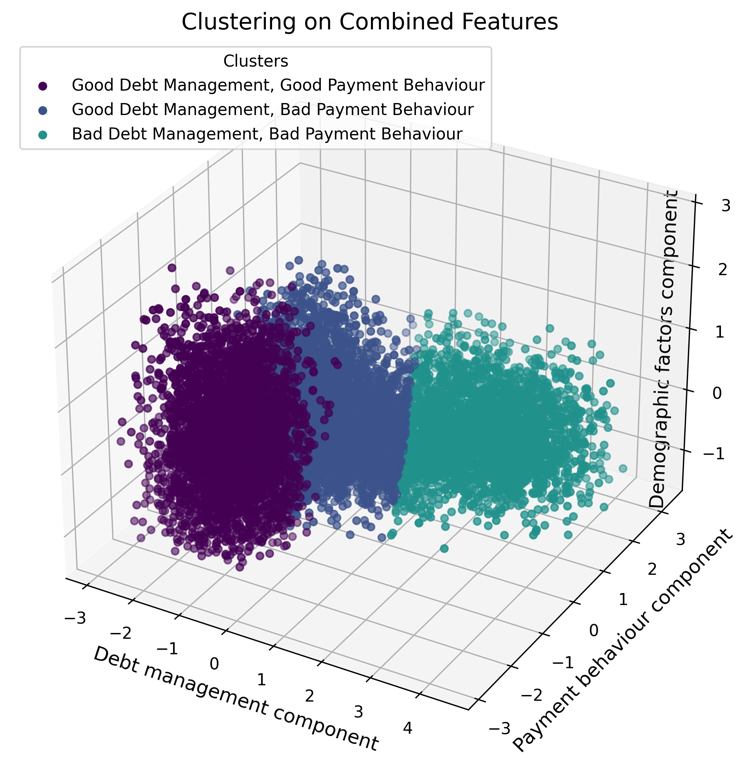


Fig 15

The number of clusters are validated by silhouette scores. Fig 15 shows that the clusters do not change along the demographic factors component thus it has little influence. PCA represents a linear combination of the features hence lower values of the debt management component indicate low values of its features implying good debt management. Same is the case for payment behaviour. The relationship between the components and credit scores is interpreted by comparing the credit score distribution in each cluster. In order to quantify the effect of each component on the credit score, logistic regression is performed to classify credit scores using the three components. The model’s accuracy is used to validate it whereas coefficients of each component are used to interpret the quantified effect of that component. The results are presented in the next section.

# FINDINGS, REFLECTIONS AND FURTHER WORK

The number of clusters are validated by silhouette scores. Fig 15 shows that the clusters do not change along the demographic factors component thus it has little influence. PCA represents a linear combination of the features hence lower values of the debt management component indicate low values of its features implying good debt management. Same is the case for payment behaviour. The relationship between the components and credit scores is interpreted by comparing the credit score distribution in each cluster. In order to quantify the effect of each component on the credit score, logistic regression is performed to classify credit scores using the three components. The model’s accuracy is management component indicate low values of its features implying good debt management. Same is the case for payment behaviour. The relationship between the components and credit scores is interpreted by comparing the credit score distribution in each cluster. In order to quantify the effect of each component on the credit score, logistic regression is performed to classify payment behaviour. The relationship between the components and credit scores is interpreted by comparing the credit score distribution in each cluster. In order to quantify the effect of each component on the credit score, logistic regression is performed to classify credit scores using the three components. The model’s accuracy is management component indicate low values of its features implying good debt management. Same is the case for payment behaviour. The relationship between the components and credit scores is interpreted by comparing the credit score distribution relationship between the components and credit scores is interpreted by comparing the credit score distribution in each cluster. In order to quantify the effect of each component on the credit score, logistic regression is performed to classify payment behaviour. The relationship between the components and credit scores is interpreted by comparing the credit score distribution in each cluster. In order to quantify the effect of each component on the credit score, logistic regression is performed to classify credit scores using the three components. The model’s accuracy The relationship between the components and credit scores is interpreted by comparing the credit score distribution in each cluster. In order to quantify the effect of each component on the credit score, logistic regression is performed to classify credit scores using the three components. The model’s accuracy is management component indicate low values of its features implying good debt management. Same is the case for payment behaviour. The relationship between the components and credit scores is interpreted by comparing the credit score distribution relationship between the components and credit scores is interpreted by comparing the credit score distribution in each cluster. In order to quantify the effect of each component on the credit score, logistic regression is performed to classify payment behaviour. The relationship between the components and credit scores is interpreted by comparing

##### REFERENCES

1. https://www.experian.com/blogs/ask-experian/credit-education/faqs/what-is-credit/
2. https://www.sciencedirect.com/science/article/pii/S0167268111001259?casa\_token=oCX4zwDOL90AAAAA:3yW3f9bMf2HNCDj23872YdcPsyqnnvBNz4hBQYd1WqdTczjY1p1eP9luirCqfi1PypItevzdFw
3. https://www.sciencedirect.com/science/article/pii/S1051137711000441?casa\_token=UzqWsYKDhRkAAAAA:oVC2AvmcUZnpuaE0c-dbTPG5C1PhZ8tnj1TgyTBdYrpoqAncnLI1CCfA3VHdAfCI53r2Arek2g.
4. https://www.sciencedirect.com/science/article/pii/S0169207000000340?casa\_token=yw1Z9wdF4xoAAAAA:5QKISrz0kq8r9er1s2kl5kGuSHRaU0ME6Az0hZ-SQ5Zoipf7OvL\_CssGhkfXJMFH72MJOnSvDA
5. https://www.experian.com/blogs/ask-experian/why-would-you-want-a-good-credit-score/
6. https://www.clevelandfed.org/publications/economic-commentary/2010/ec-201016-your-credit-score-is-a-ranking-not-a-score
7. https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=2042711
8. https://www.kaggle.com/datasets/iremnurtokuroglu/credit-score-classification-cleaned-dataset

##### WORD COUNTS

|  |  |
| --- | --- |
| **Section** | **Word Count** |
| Introduction | 298 |
| Analytical Questions and Data | 298 |
| Analysis | 1000 |
| Findings, Reflections and Further Work |  |
| Total |  |