Credit Score Analysis

I. INTRODUCTION

According to Experian, credit is the capacity to borrow money with the understanding that it will be paid back later (Waugh, 2024). Credit score of an individual is defined as a numerical value which indicates how likely they are to have the ability to pay back their debts (Arya et al., 2013). The credit score calculated by Fair Isaac Corporation known as the FICO Score is used by 90 of the top 100 financial institutions in the United States (Smith, 2011). The score 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 (Smith, 2011). 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 an existing framework that is 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. Accurate credit scores are beneficial for both borrowers and lenders. Firms that lend money to customers such as banks will draft improved policies regarding granting credit to new customers or increasing credit limits for existing customers (Thomas, 2000). 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 (Akin, 2023). These act as sources of motivation for this study.

II. ANALYTICAL QUESTIONS AND DATA

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, domain knowledge will be used 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 influencing 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%) (Demyanyk, 2010). Moreover, the Federal Bank of Kansas City states that demographic factors also influence credit scores (Hayashi & Stavins, 2012). 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 vs a poor credit score?
2. What is the relationship between an individual's payment behavior and their credit score?
3. How 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 (Tokuroglu, 2024) 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 generalizable, but validating the findings with existing frameworks will help in this regard.

III. 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).

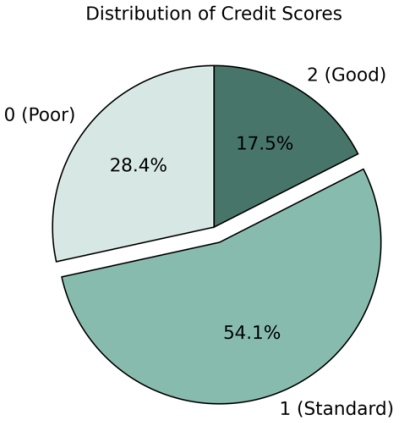

Fig 1

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 KMeans 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 8 months is aggregated using the average (for consistency) so that each datapoint represents an individual customer and segmentation can be performed.

A. 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.

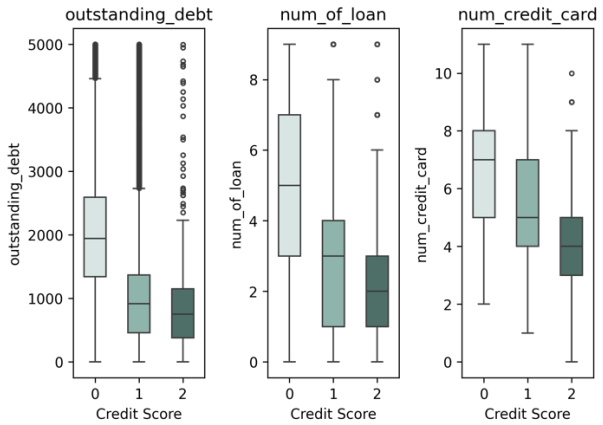

Fig 2

	F-statistic	p-value	etaSquared
num_credit_card	1294.554	0.000	0.195
outstanding_debt	1228.893	0.000	0.187
num_of_loan	1006.234	0.000	0.158

Fig 3

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

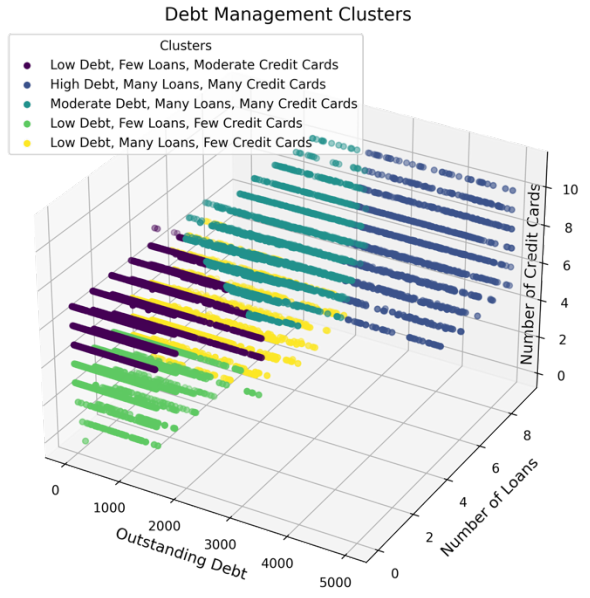

Fig 4

The ideal number of clusters and clustering model is validated using silhouette analysis. 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.

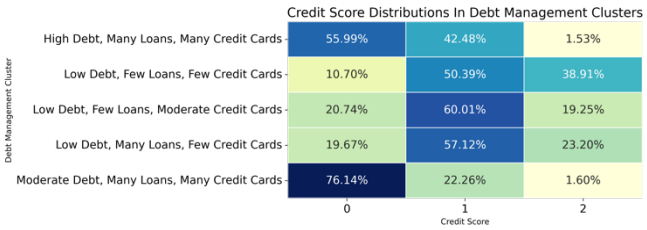

Fig 5

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

B. Customer Segmentation on Payment Behaviour

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

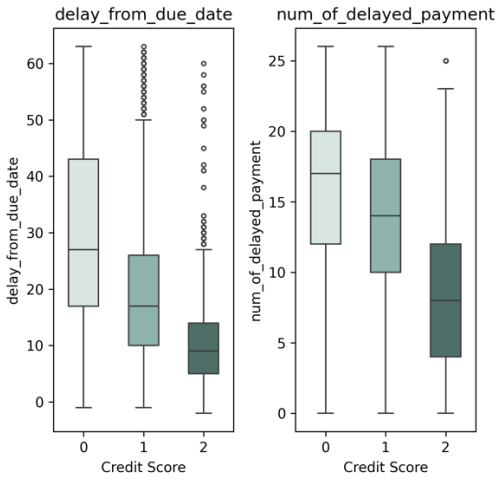

Fig 6

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.

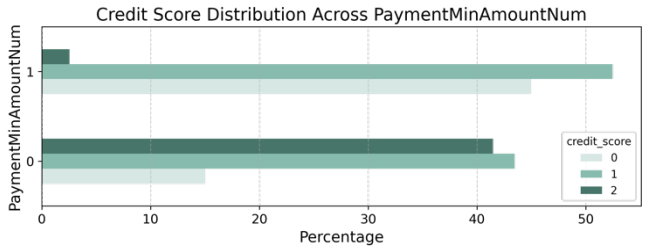

Fig 7

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.

	Variable	Test	Statistic	p-value	etaSquared	Cramer's V
0	delay_from_due_date	ANOVA	1540.769	0.000	0.223	None
1	num_of_delayed_payment	ANOVA	1175.577	0.000	0.180	None
2	paymentMinAmountNum	Chi-Square	2879.732	0.000	None	0.518

Fig 8

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

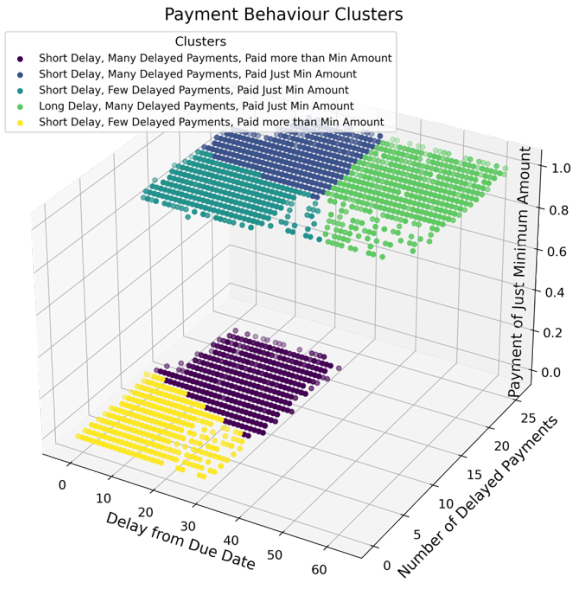

Fig 9

Silhouette analysis is used to validate the clustering model and the ideal number of clusters. The clusters 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.

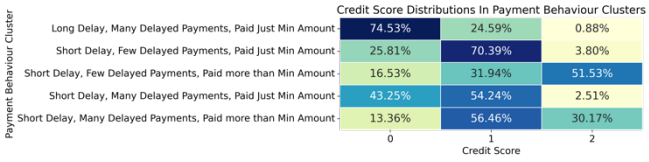

Fig 10

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.

C. 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.

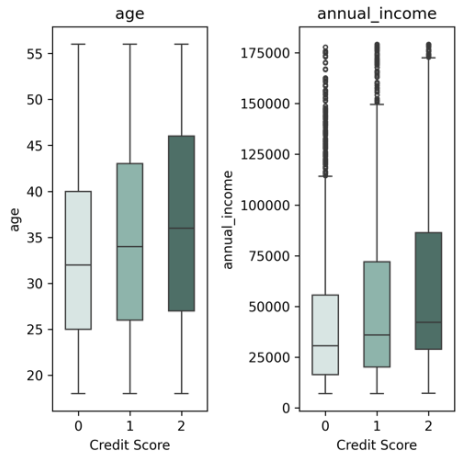

Fig 11

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

	F-statistic	p-value	etaSquared
annual_income	220.063	0.000	0.039
age	98.551	0.000	0.018

Fig 12

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

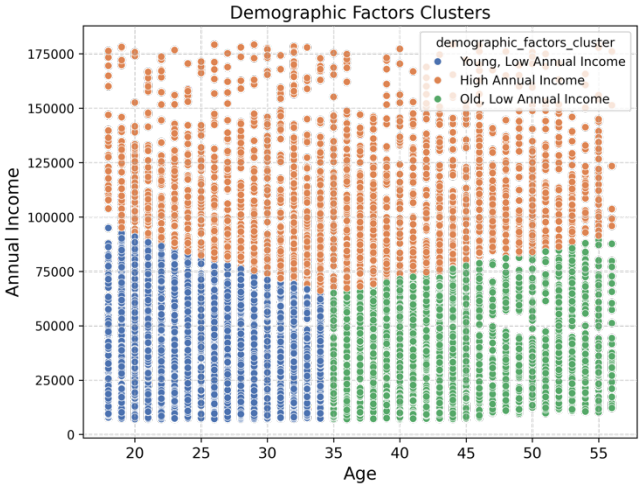

Fig 13

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.

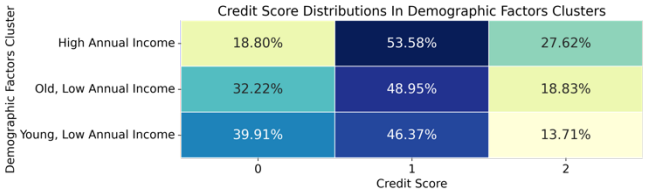

Fig 14

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.

D. Customer Segmentation on Combined Features

This study has analyzed the influence of debt management, payment behaviour and demographic factors in isolation but in reality, credit scores are calculated while considering these factors together. Hence PCA is applied to feature engineer debt management 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.

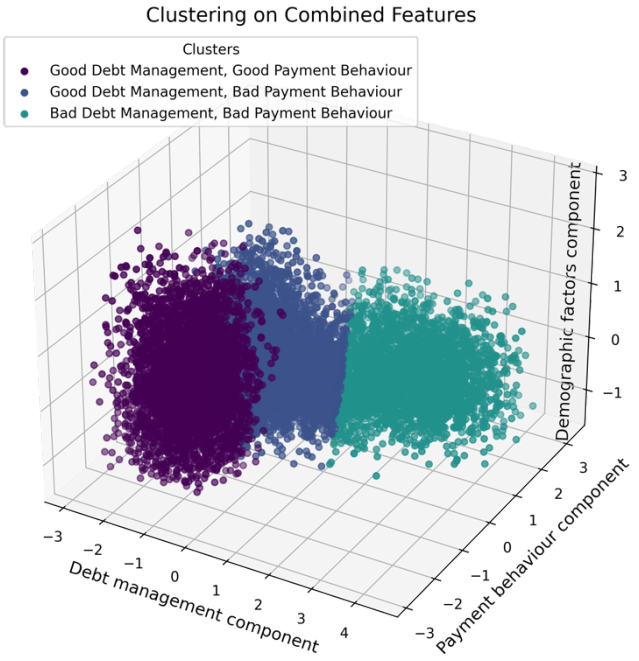

Fig 15

Number of clusters and clustering quality are validated by silhouette analysis where 3 clusters have similar plot thickness and a silhouette score of 0.39. Fig 15 shows that clusters do not change along the demographic factors component thus it has little influence. PCA represents a linear combination of 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. To quantify the effect of each component on the credit score, logistic regression is performed to classify credit scores using these three components. The model's accuracy validates it whereas coefficients of each component are used to interpret the quantified effect of that component. Results are presented in the next section.

IV. FINDINGS, REFLECTIONS AND FURTHER WORK

The impact of each component on credit score can be interpreted by comparing the credit score distribution in each cluster of Fig 15.

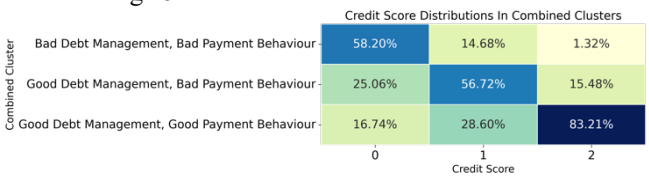

Fig 16

It is observed that the segment of customers who have good debt management and good payment behaviour have the highest percentage (83.21%) of good credit scores regardless of their age and annual income. Whereas, customers who have bad debt management and bad payment behaviour have the highest percentage of poor credit scores (58.20%). To quantify the effect of these components on credit scores, results of the logistic regression model are analyzed. The model uses the same components that were used in clustering, instead of balancing the data and deriving the components again to ensure consistency between the clustering and regression models. It achieves a 66.7% accuracy. The coefficients are shown in Fig 17.

	Credit Score 0	Credit Score 1	Credit Score 2
debtManagementComponent	0.483	-0.084	-0.399
paymentBehaviourComponent	0.410	0.328	-0.738
demographicFactorsComponent	-0.015	0.048	-0.032

Fig 17

Coefficients are converted from log-odds to odd ratios for better interpretability.

	Odds Ratios		
	Credit Score 0	Credit Score 1	Credit Score 2
debtManagementComponent	1.622	0.919	0.671
paymentBehaviourComponent	1.507	1.388	0.478
demographicFactorsComponent	0.985	1.049	0.968

Fig 18

From the previous section, in order to answer the research questions while examining the features in isolation, it can be implied that better debt management by an individual in terms of features such as outstanding debt, number of loans and number of credit cards, is associated with a higher likelihood of a good credit score as opposed to a poor one. Similarly, better payment behaviour of an individual signifies a greater likelihood of them having a good credit score and vice versa. In comparison, the effect of age and income was minimal but it could still be seen that individuals with higher incomes have more chances of achieving a good credit score. To answer the research questions when examining all three components together, Fig 18 shows that a 1 unit increase towards bad debt management increases the odds of a poor credit score by 62.2% and decreases the odds of a good credit score by 32.9%. Similarly, a 1 unit increase towards bad payment behaviour increases the odds of a poor credit score by 50% and decreases the odds of a good credit score by 52.2%. Compared to debt management and payment behaviour, the effect of demographic factors on credit score is very minimal in this case. Overall, for a good credit score, customers should not rely heavily on debt and should have a positive payment behaviour.

The dataset used in this study is synthetic hence the findings are not generalizable which acts a limitation of this work. This was a trade-off since the dataset contains relevant features for this study's analysis. However, the effects of debt management and payment behaviour on credit score are consistent with FICO's scoring framework mentioned earlier (Demyanyk, 2010). Customer data across 8 months was aggregated using averages for customer segmentation which loses variations in the data as it summarizes it to a single value. The average may not be the best representation of the data. Silhouette scores of clustering models in this study are relatively low (0.3-0.4) however, this may be because clustering is done on three dimensions. Hence the thickness of silhouette plots is also considered while validating clustering models as demonstrated in the accompanying Jupyter notebook.

For future work, similar analytical techniques should be applied to a real-world dataset for generalizability. If dealing with customer time-series data, alternate feature engineering methods should be used and evaluated for representing customer data as a single point for segmentation. Segmentation should be done using other techniques such as decision trees and the results can be compared with KMeans. Different datasets should be merged for access to more features which may influence credit score in real life; this study uses a limited set of features from a single dataset.

V. REFERENCES

[1] Waugh, E. (2024) *What is credit?*, *Experian*. Experian. Available at: https://www.experian.com/blogs/ask-experian/credit-education/faqs/what-is-credit/ (Accessed: 20 December 2024).

[2] Arya, S., Eckel, C. and Wichman, C. (2013) 'Anatomy of the credit score', *Journal of Economic Behavior & Organization*, 95, pp. 175–185. doi: 10.1016/j.jebo.2011.05.005.

[3] Smith, B. C. (2011) 'Stability in consumer credit scores: Level and direction of FICO score drift as a precursor to mortgage default and prepayment', *Journal of Housing Economics*, 20(4), pp. 285–298. doi: 10.1016/j.jhe.2011.09.001.

[4] Thomas, L. C. (2000) 'A survey of Credit and behavioural scoring: Forecasting financial risk of lending to consumers', *International Journal of Forecasting*, 16(2), pp. 149–172. doi: 10.1016/s0169-2070(00)00034-0.

[5] Akin, J. (2023) *Why do you want A good credit score?*, *Experian*. Experian. Available at: https://www.experian.com/blogs/ask-experian/why-would-you-want-a-good-credit-score/ (Accessed: 20 December 2024).

[6] Demyanyk, Y. (2010) *Your credit score is a ranking, not a score*, *Economic Commentary*. Federal Reserve Bank of Cleveland. Available at: https://www.clevelandfed.org/en/newsroom-and-events/publications/economic-commentary/economic-commentary-archives/2010-economic-commentaries/ec-201016-your-credit-score-is-a-ranking-not-a-score.aspx (Accessed: 20 December 2024).

[7] Hayashi, F. and Stavins, J. (2012) 'Effects of credit scores on Consumer Payment Choice', *SSRN Electronic Journal*. doi: 10.2139/ssrn.2042711.

[8] Tokuroglu, I. N. (2024) *Credit Score Classification cleaned dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/iremnurtokuroglu/credit-score-classification-cleaned-dataset (Accessed: 20 December 2024).

VI. WORD COUNT

Section	Word Count
Introduction	300
Analytical Questions and Data	300
Analysis	1000
Findings, Reflections and Further Work	600
Total	2200