Project Plan

Title: Understanding Credit Scores: Investigating the impact of debt management, payment behaviours and demographics on credit scores.

Application Domain and Dataset: I found a suitable dataset on Kaggle. It has 27 attributes and 1 class column of the credit score (Poor (0), Standard (1) and Good (2)). It has data for 12,500 customers over 8 months resulting in 100,000 entries. The attributes are listed below:

No	INPUTS	Description
1	**id**	Unique identifier for each record.
2	**customer_id**	Unique identifier for each customer.
3	**month**	Month of the transaction or record.
4	**name**	Customer's name.
5	**age**	The customer's age.
6	**ssn**	Customer's social security number.
7	**occupation**	The customer's occupation.
8	**annual_income**	The customer's annual income.
9	**monthly_inhand_salary**	The customer's monthly take-home salary.
10	**num_bank_accounts**	Total number of bank accounts owned by the customer.
11	**num_credit_card**	Total number of credit cards held by the customer.
12	**interest_rate**	The interest rate applied to loans or credits.
13	**num_of_loan**	Number of loans the customer has taken.
14	**type_of_loan**	Categories of loans obtained by the customer.
15	**delay_from_due_date**	The delay in payment relative to the due date.
16	**num_of_delayed_payment**	Total instances of late payments made by the customer.
17	**changed_credit_limit**	Adjustments made to the customer's credit limit.
18	**num_credit_inquiries**	Number of inquiries made regarding the customer's credit.
19	**credit_mix**	The variety of credit types the customer uses (e.g., loans, credit cards).
20	**outstanding_debt**	Total amount of debt the customer currently owes.
21	**credit_utilization_ratio**	Proportion of credit used compared to the total credit limit.
22	**credit_history_age**	Duration of the customer's credit history.
23	**payment_of_min_amount**	Indicates if the customer pays the minimum required amount each month.
24	**total_emi_per_month**	Total Equated Monthly Installment (EMI) paid by the customer.
25	**amount_invested_monthly**	Monthly investment amount made by the customer.
26	**payment_behaviour**	Customer's payment habits and tendencies.
27	**monthly_balance**	The remaining balance in the customer's account at the end of each month.
28	**credit_score**	The customer's credit score (target variable: "Good," "Poor," "Standard").

Link to original dataset:
https://www.kaggle.com/datasets/parisrohan/credit-score-classification/data?select=test.csv
Link to cleaned dataset:
https://www.kaggle.com/datasets/iremnurtokuroglu/credit-score-classification-cleaned-dataset/data

The original dataset has been worked on by a contributor on Kaggle and they have cleaned the dataset for their project (link provided above). Is it okay if I use the cleaned dataset for my own analysis?

The dataset has a good variety of attributes related to the customer which I believe will give me flexibility in my project however, I do not plan on using all of the attributes in my project. Is that okay?

Most of the attributes in the dataset are numerical. Would this be beneficial for my analysis?

Well-motivated analytical questions: While coming up with the research questions I first divided the attributes into relevant groups to get a better idea of what is available and what is relevant. I came up with three main groups:

Demographic Factors:

1. customer_id
2. age
3. occupation

Financial Factors:

1. annual_income
2. monthly_inhand_salary
3. num_bank_accounts
4. num_credit_card
5. num_of_loan
6. outstanding_debt
7. credit_utlization_ratio
8. credit_history_age
9. monthly_balance
10. amount_invested_monthly
11. num_credit_inquiries

Behavioural Factors:

1. payment_of_min_amount
2. payment_behaviour
3. delay_from_due_date
4. num_of_delayed_payment
5. changed_credit_limit
6. total_emi_per_month
7. credit_mix

Question 1: How does customer debt management (credit utilisation ratio and outstanding debt) impact the likelihood of achieving a Good vs a Poor credit score?

Understanding how debt management influences credit scores can help customers manage their debts more effectively for a higher credit score. This understanding will also enable lenders such as banks to develop financial products that encourage responsible debt management.

Question 2: What is the relationship between customers' payment behaviour (number of delayed payments and payment of minimum amount) and their credit scores?

Identifying how payment habits affect credit scores can help financial institutions understand risk profiles while determining whether to grant loans. It will also help customers understand the extent of the effect of their payment behaviour on their credit score.

Question 3: What is the impact of demographic factors (age and occupation) on credit scores among customers?

Analysing demographic influences on credit scores can help banks and financial institutions tailor their services and outreach efforts to specific customer segments. For example starter credit cards for younger customers and more flexible plans for middle-aged customers.

Question 4: How do factors such as the number of bank accounts and credit cards owned by customers influence the distribution of credit scores among customers?

Understanding this impact will allow banks to offer personalised financial plans e.g. if a greater number of bank accounts result in a higher customers credit score then banks can come up with plans advising customers to open a variety of accounts with them which will be beneficial to both parties.

Question 5: To what extent does annual income and monthly balance impact credit scores?

Exploring this relationship will help lenders such as banks in their risk assessment of customers while evaluating loan applications and credit decisions.

I believe the first three questions lay a good foundation and can be solid analytical questions after some tweaking. I am unsure about questions 4 and 5. I might merge them into one question or come up with a completely new question to act as a fourth research question. Looking forward to your feedback on this.

My (Tentative) Plan

1. Data preparation:
    - Handle any missing values and keep a record of those values
    - Normalise any relevant variables if needed
    - Feature engineering: Create binary indicators and categories for variables such as payment_of_min_amount, occupation and age

2. Exploratory data analysis:
    - Summarise key descriptive statistics for variables related to debt management, payment behaviour, demographics, banking activity and income
    - Plot distributions such as scatter plots and box plots of credit_utilisation_ratio, outstanding_debt, num_delayed_payments and payment_of_min_amount across credit score categories to observe patterns and outliers
    - Visualise the distribution of credit scores across age groups and occupations
    - Use scatter plots and correlation matrices to show relationships between number of banks/credit cards, annual income and monthly balance with credit scores

3. Methodologies:
    - Use Decision Trees to assess how credit_utlisation_ratio and outstanding_debt influence credit scores categories
    - Conduct correlation analysis and multiple regression to examine the relationship of num_of_delayed_payments and payment_of_min_amount with credit scores
    - Calculate Cohen's d to measure effect size between age brackets on credit score categories and use ANOVA to detect significant difference in credit scores across occupations
    - Use KNNs to group together similar customer profiles based on banking activity (number of bank account/credit cards) and credit scores to identify patterns
    - Logistic regression to study to the impact of annual income and monthly balance on credit scores

4. Evaluation and Insights:
    - Assess model performances and reiterate
    - Summarise key insights in order to answer analytical questions

Specific Questions:

1. The source of the dataset is not mentioned on Kaggle. Is it necessary for the dataset to have a documented source?
2. The dataset is slightly unbalanced with regards to credit score category (53% Standard, 29% Poor and 18% Good). Is this okay or should I try to balance the dataset?
3. The dataset has data on 12,500 customers over 8 months which amounts to 100,000 entries. Will this take time to compute, if yes then should I try to reduce the dataset while maintaining data integrity?
4. What are the expectations with regards to the analytical methods used? Is it okay to repeat methods over different analytical questions as long as it is relevant to the question and gives desired results?