Project Summary

|  |  |
| --- | --- |
| Batch details | PGP-DSE April’23 Gurgaon |
| Team members | Anmol Razdan, Aviral Negi, Anshul Saini, Srishti Salian,  Pratishtha Srivastava |
| Domain of Project | Finance |
| Proposed  project title | Credit Score Classification |
| Group Number | Group-1 |
| Team Leader | Anmol Razdan |
| Mentor Name | Ms. Vibha Santhanam |

Date: 08-12-2023

|  |  |
| --- | --- |
| Vibha Santhanam | Anmol Razdan |
| Signature of Mentor | Signature of Team Leader |

Table of Contents

|  |  |  |
| --- | --- | --- |
| SI NO | TOPIC | PAGE NO |
| 1 | Overview | 3 |
| 2 | Business Problem  Statement | 3 |
| 3 | Topic Survey in Depth | 4 |
| 4 | Critical Assessment of  Topic Survey | 5 |
| 5 | Methodology To Be  Followed | 6 |
| 6 | References | 17 |
|  |  |  |

Project Details

## OVERVIEW

This project aims to develop a predictive model for classifying individuals into credit score categories based on financial data. The dataset will undergo comprehensive preprocessing, followed by exploratory data analysis and feature engineering to enhance model accuracy. Various machine learning algorithms will be considered, with the selected model evaluated and fine-tuned through hyperparameter tuning. The final model will be validated and tested on separate datasets, and the project will conclude with a detailed report outlining the methodology, insights, and recommendations for implementing the model in credit scoring systems, prioritizing ethical considerations throughout the process.

BUSINESS PROBLEM STATEMENT

The problem is to develop a predictive model for classifying individuals into credit score categories based on their financial data and other relevant factors. Accurate credit scoring is crucial for the financial institutions to assess the reliability of applicants and manage risk effectively as it leads to less default.

1. What would you achieve by this project?

This project aims to transform credit assessment for financial institutions by developing a predictive model for classifying individuals into credit score categories. The implementation of an accurate credit scoring system is crucial for assessing applicant reliability and effectively managing risk, ultimately leading to a reduction in defaults. Through this initiative, we seek to enhance decision-making processes, improve operational efficiency, and increase customer satisfaction by providing fair and precise credit evaluations. The project also addresses regulatory compliance, ensuring alignment with industry standards.

1. How would this help the business or clients?

Financial organizations receive thousands of applications for seeking credit from various prospective customers of diverse backgrounds and needs. Lending credit becomes a very essential decision, because if the customer does not return, there will be imbalance of cash inflow and outflow for the organization and thus, it may suffer losses. Checking and determining manually each application is not feasible when customers demand result as soon as they submit their application. Thus, it becomes essential to build a Machine Learning model for using the customer data whom credit was given along with their computed credit score category and use this data to determine the credit score category for the incoming applications. This will help to save manual efforts while screening applications and work can be achieved at higher scale and volume with lesser manual error.

1. What is the further scope of the project?

The project focuses on making the credit score system more effective and applicable in the real world. This includes creating a versatile machine learning algorithm to improve predictions and collaborating with financial institutions for practical use. We're also looking into the possibility of a dynamic credit scoring system that can adjust in real-time to new information. This approach aims to not only enhance accuracy but also to adapt to changing financial scenarios through ongoing improvement and collaboration.

1. Limitation of the project?

Limitation of the project is that it faces the requirement for substantial volumes of high-quality data. The accuracy of machine learning algorithms hinges on having extensive and reliable datasets. In instances where the available data is of poor quality or limited in scope, our algorithm may encounter difficulties in generating accurate predictions. This limitation emphasizes the importance of securing a robust and comprehensive dataset for the success of our credit score classification project.

TOPIC SURVEY IN BRIEF

1. Problem understanding

The project aims to develop a predictive model for credit scoring, but potential challenges may arise in accurately predicting creditworthiness during unforeseen economic changes or personal financial crises, which historical data may not fully capture. Balancing model complexity and interpretability is crucial for practical implementation in financial decision-making. Addressing ethical concerns related to potential biases and ensuring ongoing model adaptability are also critical aspects of the project's challenge landscape.

1. Current solution to the problem

The current solution involves employing machine learning algorithms, such as logistic regression, decision trees, or ensemble methods, to analyze historical financial data and predict credit scores. Regular model updates, continuous monitoring, and efforts to address biases contribute to improving accuracy and reliability. Ethical considerations involve transparency in decision-making and compliance with regulations, while ongoing refinements accommodate evolving financial landscapes and user needs.

1. Proposed solution to the problem

The proposed solution involves implementing an ensemble of machine learning models, combining the strengths of various algorithms to enhance the accuracy of credit score predictions. Additionally, integrating alternative data sources, aims to provide a more comprehensive assessment of creditworthiness. A dynamic model update mechanism, real-time decision support, efficiency, and transparency in addressing the challenges of predicting credit scores in a rapidly changing financial landscape.

CRITICAL ASSESSMENT OF TOPIC SURVEY

Q1. Find the key area, gaps identified in the topic survey where the project can add value to the customers and business.

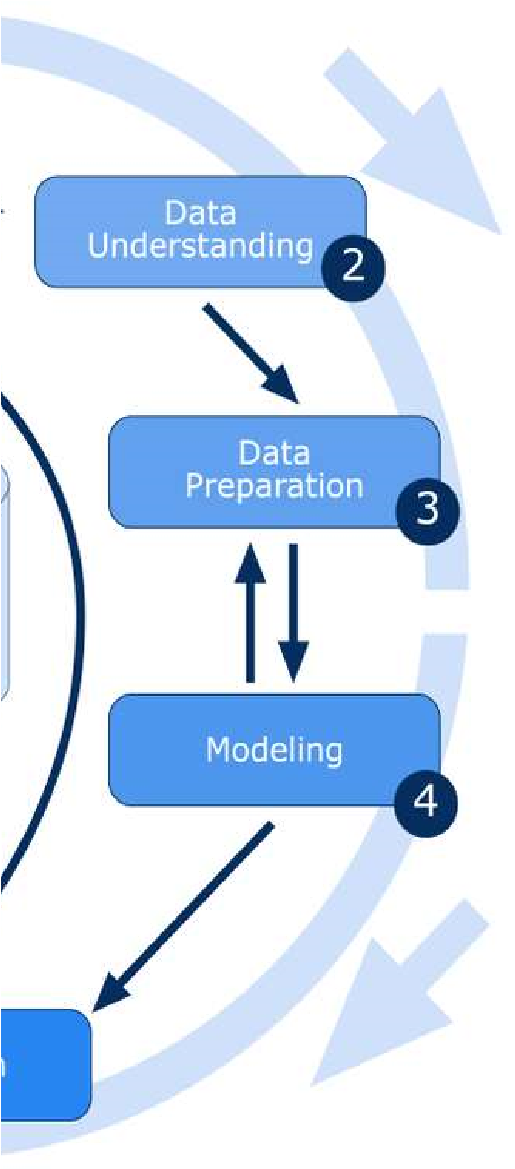
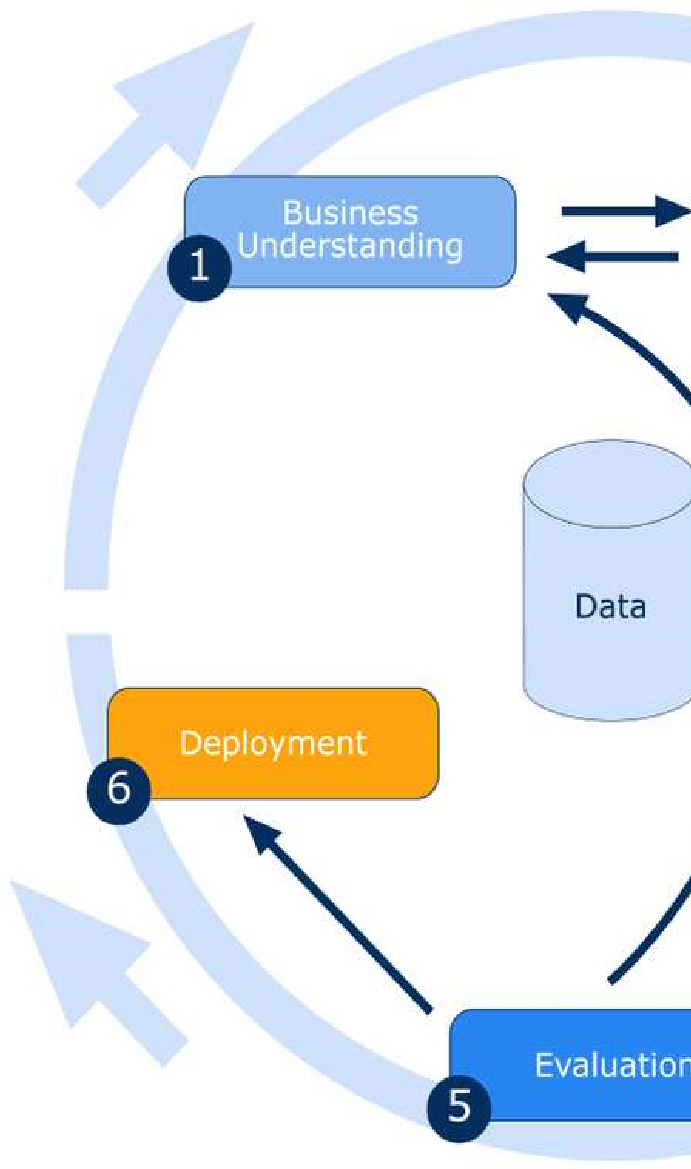
The project adds value by enhancing credit score accuracy, providing real-time decision support, and integrating alternative data to address gaps in traditional models. Additionally, a focus on ethical considerations, bias mitigation, and dynamic model updates ensures a more reliable and adaptable credit scoring system, benefitting both customers and the business.

Q2. What key gaps are you trying to solve?

The project adds value by enhancing credit score accuracy, providing real-time decision support, and integrating alternative data to address gaps in traditional models. Additionally, a focus on ethical considerations, bias mitigation, and dynamic model updates ensures a more reliable and adaptable credit scoring system, benefitting both customers and the business.

METHODOLOGY TO BE FOLLOWED

* Business Understanding



* Data Understanding
* Data Preparation
* Modelling
* Evaluation
* Deployment

# Business Understanding

Business understanding for credit score classification is crucial for effective risk management and informed lending in the financial sector. Credit scores, representing an individual's creditworthiness, influence personalized lending approaches based on credit history. Business decisions, including loan approval, interest rate determination, and customer segmentation, depends on these insights. This nuanced understanding not only reduces risk but also ensures regulatory compliance, fostering responsible lending practices and sustainable financial growth.

PROBLEM STATEMENT:

The problem is to develop a predictive model for classifying individuals into credit score categories based on their financial data and other relevant factors. Accurate credit scoring is crucial for the financial institutions to assess the reliability of applicants and manage risk effectively as it leads to less default.

# Data Understanding

Rows: 100000

Variables: 28

## Variable Description

* **ID:** Unique ID of the record
* **Customer\_ID:** Unique ID of the customer
* **Month:** Month of the year
* **Name:** The name of the person
* **Age:** The age of the person
* **SSN:** Social Security Number of the person
* **Occupation:** The occupation of the person
* **Annual\_Income:** The Annual Income of the person
* **Monthly\_Inhand\_Salary:** Monthly in-hand salary of the person
* **Num\_Bank\_Accounts:** The number of bank accounts of the person
* **Num\_Credit\_Card:** Number of credit cards the person is having
* **Interest\_Rate:** The interest rate on the credit card of the person
* **Num\_of\_Loan:** The number of loans taken by the person from the bank
* **Type\_of\_Loan:** The types of loans taken by the person from the bank
* **Delay\_from\_due\_date:** The average number of days delayed by the person from the date of payment
* **Num\_of\_Delayed\_Payment:** Number of payments delayed by the person
* **Changed\_Credit\_Card:** The percentage change in the credit card limit of the person
* **Num\_Credit\_Inquiries:** The number of credit card inquiries by the person
* **Credit\_Mix:** Classification of Credit Mix of the customer
* **Outstanding\_Debt:** The outstanding balance of the person
* **Credit\_Utilization\_Ratio:** The credit utilization ratio of the credit card of the customer
* **Credit\_History\_Age:** The age of the credit history of the person
* **Payment\_of\_Min\_Amount:** Yes if the person paid the minimum amount to be paid only, otherwise no.
* **Total\_EMI\_per\_month:** The total EMI per month of the person
* **Amount\_invested\_monthly:** The monthly amount invested by the person
* **Payment\_Behaviour:** The payment behaviour of the person
* **Monthly\_Balance:** The monthly balance left in the account of the person
* **Credit\_Score:** The credit score of the person

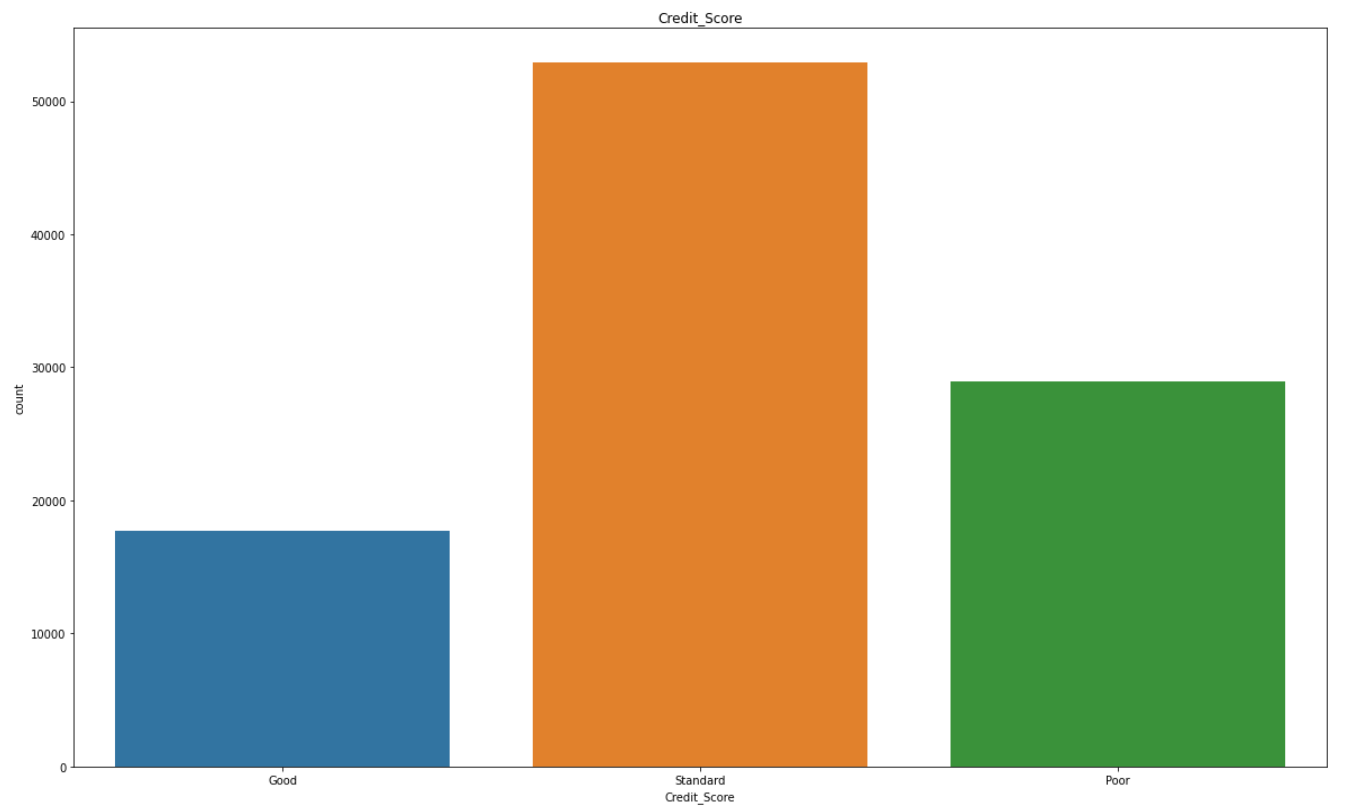
Target Variable

* We have derived our target, i.e. Credit Score
* We have 3 categories in target variable:

Good

Standard

Poor



Data Preparation

Data preparation is a crucial step to ensure the quality and relevance of the data for model training. Here are a few key steps that we have introduced in our project i.e. credit score classification:

* Data Cleaning
* Garbage value/ missing value treatment
* Outlier treatment

Data Cleaning

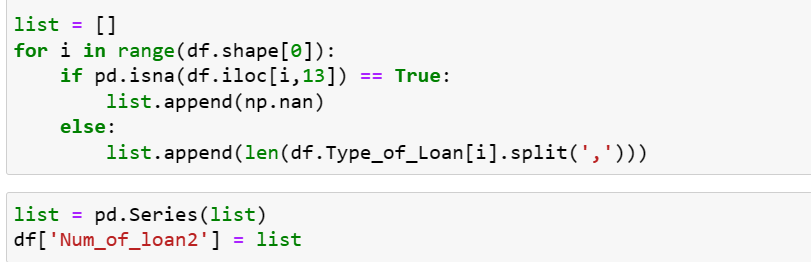
In the systematic process of data cleaning, we have undertaken a comprehensive approach to refine our dataset. Our efforts began by scrutinizing the various features, employing rigorous steps to ensure the highest data quality. By implementing a series of cleansing techniques and using advanced data manipulation methods, we have successfully enhanced the reliability and integrity of our dataset.

* Num\_of\_Loan: Replaced ‘\_’, ‘-’ with empty value.

A white card with black text

Description automatically generated

* Type\_of\_Loan: Generated a new column by deriving values through the segmentation of data in Type\_of\_Loan



* Credit\_history\_age: Cleaned credit\_history\_age and converted it into number from month name.

A screenshot of a computer code

Description automatically generated

* Age: Replaced ‘\_’, ‘-’ with empty value and then changed the data type to integer.

A screenshot of a computer

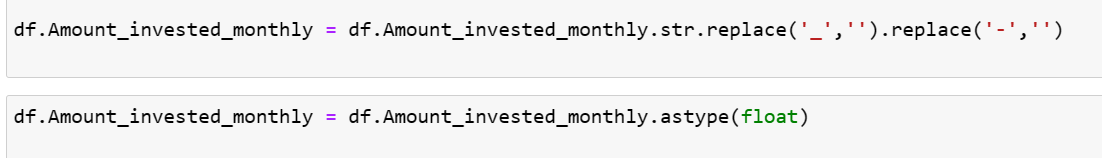
Description automatically generated

* Annual\_income: Replaced ‘\_’, ‘-’ with empty value and then changed the data type to float.

A screenshot of a computer program

Description automatically generated

* Amount\_invested\_monthly: Replaced ‘\_’, ‘-’ with empty value and then changed the data type to float.

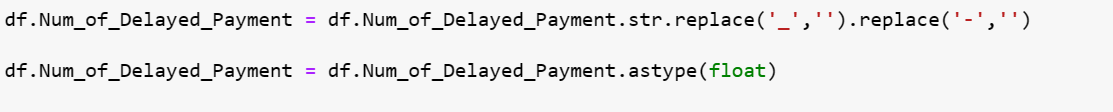


* Monthly\_balance: Replaced ‘\_’, ‘-’ with empty value and then changed the data type to float.

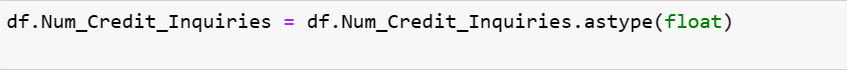
A screenshot of a computer

Description automatically generated

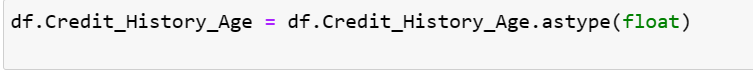
* Num\_of\_delayed\_payments: Replaced ‘\_’, ‘-’ with empty value and then changed the data type to float.



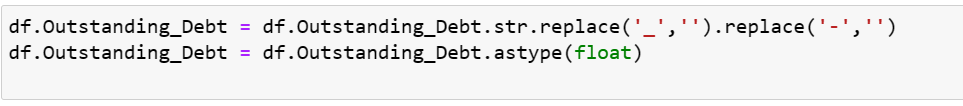
* Num\_Credit\_Inquiries: Changed the data type to float.



* Credit\_History\_Age: Changed the data type to float.



* Outstanding\_Debt: Replaced ‘\_’, ‘-’ with empty value and then changed the data type to float.



* Changed\_Credit\_Limit: Replaced ‘\_’, ‘-’ with empty value and converted empty value to nan values. Lastly changed the data type to float.

A screenshot of a computer program

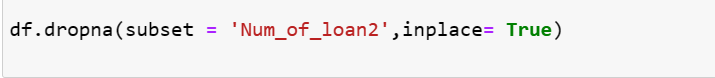
Description automatically generated

Garbage value/ missing value treatment

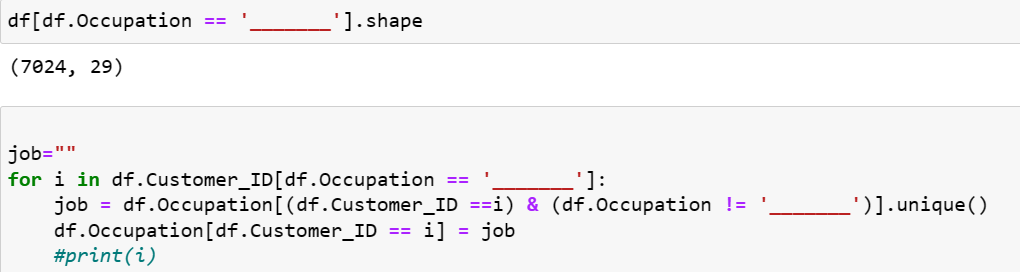
In fixing issues with our data, we took care of two main things: garbage values and missing values. We checked the data closely and got rid of any weird or wrong values to make sure our info is solid. For the missing stuff, we used smart methods to fill in the gaps, so our data is complete and ready for our credit score project. These steps help make sure our model gives accurate results.

* Num\_of\_loan2: Filling null values in the Num\_of\_loan2 column by replacing them with zeros obtained from the old loan column (Num\_of\_Loan). Lastly dropped the null values.



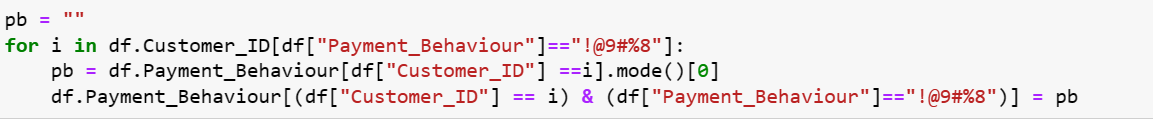


* Occupation: Fixed the ' \_\_\_\_\_\_\_ ' in the occupation column by checking the customer IDs. For customers without ' \_\_\_\_\_\_\_ ', we found their unique occupation and filled ' \_\_\_\_\_\_\_ '.



A screenshot of a computer

Description automatically generated

* Payment\_Behaviour: Fixed the ‘!@9#%8’ in the payment behaviour column by checking the customer IDs. For customers with ‘!@9#%8’ , we have taken the mode value of payment behaviour column of each customer and filled the ‘!@9#%8’ respectively. Remaining ‘!@9#%8’ left value (As the garbage value is just 1.7%) will be filled with mode of all the given values.



* Credit\_Mix: Fixed the ‘\_’ in the credit mix column by checking the customer IDs. For customers with ‘\_’, we have taken the mode value of credit mix column of each customer and filled the ‘\_’ respectively. Remaining ‘\_’ left value (As the garbage value is just 0.7%) will be filled with mode of all the given values.

A close-up of a number

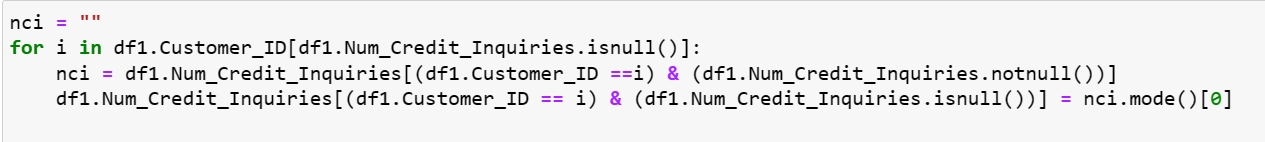
Description automatically generated

A screenshot of a computer

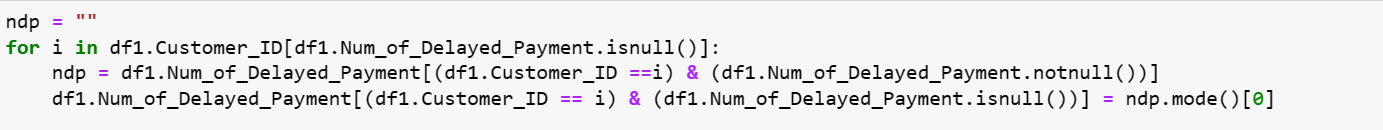
Description automatically generated



* Num\_Credit\_Inquiries: Fixed the null values in the Num credit inquiries column by checking the customer IDs. For customers with null values, we have taken the mode value of not null values of num credit inquiries column of each customer and filled the null values respectively.



* Num\_of\_Delayed\_Payment: Fixed the null values in the Num\_of\_Delayed\_Payment column by checking the customer IDs. For customers with null values, we have taken the mode value of not null values of Num\_of\_Delayed\_Payment column of each customer and filled the null values respectively.



* Columns Discarded
* Name
* Monthly\_Inhand\_Salary
* Num\_of\_Loan
* Type\_of\_Loan
* ID
* SSN
* KNN imputer

In our project, we're using a technique called K-Nearest Neighbors (KNN) imputation to fill in missing values in specific columns like Changed\_Credit\_Limit, Credit\_History\_Age, Amount\_invested\_monthly, and Monthly\_Balance

of our data. KNN looks at nearby data points and estimates what the missing values could be based on similar ones. This helps us keep our data complete and accurate for better analysis, making sure we don't lose important information."

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Outlier Treatment

In our project (credit score classification), we had some unusual data points that didn't fit with the rest. We dealt with these outliers using different research methods and techniques. By doing this, we made sure that these unusual data didn't disrupt our overall analysis. This helped us get more accurate and reliable results in the end. These steps help make sure our model gives accurate results.

* Age: labelled ages surpassing 100 as NaN in our project, ensuring us cover all bases and avoid any potential issues in our data. This decision was aimed at maintaining the integrity of our data analysis. For customers with null values, we have taken the not null values of Age column of each customer and filled the null values respectively.



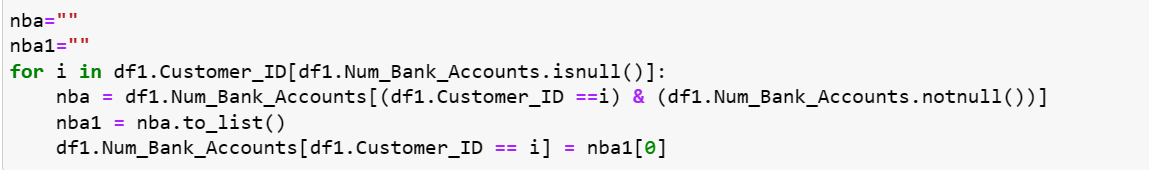
A white background with black text

Description automatically generated

* Num\_Bank\_Accounts: labelled Num\_Bank\_Accounts surpassing 10 and below 0 as NaN since data contains some garbage value, ensuring us cover all bases and avoid any potential issues in our data. This decision was aimed at maintaining the integrity of our data analysis. For customers with null values, we have taken the not null values of Num\_Bank\_Accounts column of each customer and filled the null values respectively.

The average number of accounts can vary, but many individuals might have a combination of savings and checking accounts, along with one or two other specialized accounts.





* Num\_Credit\_Card: labelled Num\_Credit\_Card surpassing 10 as NaN since data contains some garbage value, ensuring us cover all bases and avoid any potential issues in our data. This decision was aimed at maintaining the integrity of our data analysis. For customers with null values, we have taken the not null values of Num\_Credit\_Card column of each customer and filled the null values respectively.



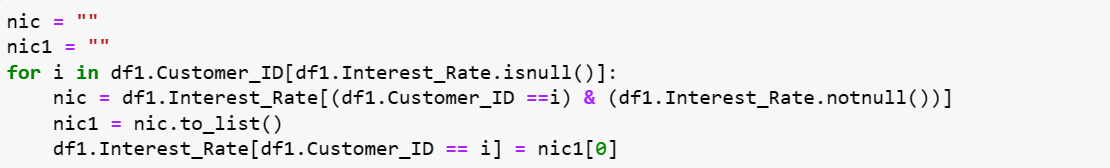
A computer code with numbers and symbols

Description automatically generated

* Interest\_Rate: labelled Interest\_Rate surpassing 34 as NaN since data contains some garbage value, ensuring us cover all bases and avoid any potential issues in our data. This decision was aimed at maintaining the integrity of our data analysis. For customers with null values, we have taken the not null values of Interest\_Rate column of each customer and filled the null values respectively.

After 34% interest rate the number of entries is constant so we proceeded with 34% and above it all is junk.





* Num\_of\_Delayed\_Payment: labelled Num\_of\_Delayed\_Payment surpassing 28 as NaN since data contains some garbage value, ensuring us cover all bases and avoid any potential issues in our data. This decision was aimed at maintaining the integrity of our data analysis. For customers with null values, we have taken the mode value of Num\_of\_Delayed\_Payment column and filled the null values respectively.





* Num\_Credit\_Inquiries: labelled Num\_Credit\_Inquiries surpassing 17 as NaN since data contains some garbage value, ensuring us cover all bases and avoid any potential issues in our data. This decision was aimed at maintaining the integrity of our data analysis. For customers with null values, we have taken the mode value of Num\_Credit\_Inquiries column and filled the null values respectively.

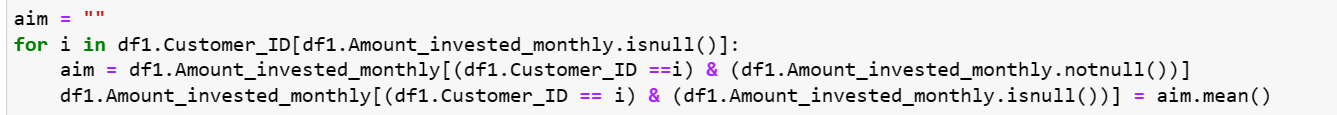
After 17 the change in number of entries is almost constant so we proceeded with it and make other null.



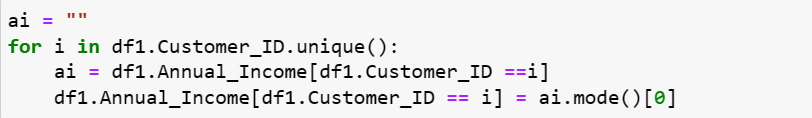


* Amount\_invested\_monthly: labelled Amount\_invested\_monthly having 10000 as NaN, ensuring us cover all bases and avoid any potential issues in our data. This decision was aimed at maintaining the integrity of our data analysis. For customers with null values, we have taken the mean values of Amount\_invested\_monthly column for each customer and filled the null values respectively.





* Annual\_Income: Adopted a systematic approach by computing the mode of the annual income column and subsequently imputing this mode value for each customer across all rows. This method ensures uniformity and reliability in addressing missing or undefined annual income data point.



* Total\_EMI\_per\_month: In analyzing the Total\_EMI\_per\_month, we employed the interquartile range (IQR) with a specific focus on the upper limit coefficient set at 5. Then labelled Total\_EMI\_per\_month having values greater than upper limit as NaN, ensuring us cover all bases and avoid any potential issues in our data. This decision was aimed at maintaining the integrity of our data analysis. For customers with null values, we have taken the mean values of Total\_EMI\_per\_month column for each customer and filled the null values respectively.

A screenshot of a computer

Description automatically generated

Univariate Analysis

In our univariate analysis of the Annual Income column, we observed a median value of around $35,000, reflecting the central tendency of the data. The lower limit was found to be $10,000, indicating the starting point for a significant portion of the income distribution within the dataset. The upper limit was set at $150,000, serving as a threshold for most income values. Notably, incomes beyond this upper limit were found to be till $180,000, signifying a smaller but noticeable segment of higher-income individuals.

A graph showing a blue rectangle with a black line

Description automatically generated

Here in Total\_EMI\_per\_month we can see that majority of people have EMI below $300 per month which is good but there are lots of outliers in this that have high amount of EMI.

A comparison of a graph

Description automatically generated

From the chart of Amount\_invested\_monthly we can see that majority people are investing a very less amount of below $500.

But there are people that are investing heavily also.

A comparison of a graph

Description automatically generated

Here we can see that most people have less than $1000 at the end of the month in their account which is less but there are lots of outliers also that have more than $1000.

A graph of a monthly balance

Description automatically generated

Bivariate Analysis

Numerical With categorical variable

From below we can see that people with good credit score have high annual income than the other two categories and same goes for people with standard credit scores but we can see that there are few outliers which have high annual income but low credit scores which may be due to other factors.

A graph showing a number of income

Description automatically generated with medium confidence

Below we can see that the people with good credit scores have high monthly balance at the end of the month as compared to people with standard and poor credit scores which is understandable. There are outiers present in this case also but they are few.

A screenshot of a graph

Description automatically generated

Categorical with Categorical variable

Credit mix with credit score

A graph of a bar chart

Description automatically generated with medium confidence

* Among our customer base, those with a ‘standard’ credit score surpass 30,000 count in ‘standard’ credit mix, overshadowing both good and bad credit mix categories.
* Notably, individuals with a credit score categorized as 'good' have the highest count at 15,000, predominantly associated with a good credit mix. In contrast, the count for 'bad' credit mix is minimal, standing around 100.
* For customers with poor credit scores, the majority align with a bad credit mix, with the highest count reaching approximately 15,000. Intriguingly, 'good' credit mix exhibits the lowest count among poor credit scores.

Payment Behaviour with credit score

A graph of different colored bars

Description automatically generated

* Among our customer base, those with a ‘standard’ credit score with 15,000 counts in ‘Low\_spent\_Small\_value\_payments’, overshadowing other categories in Payment\_Behaviour.
* Notably, individuals with a credit score categorized as 'good' have the highest count at 4,000, predominantly associated with ‘Low\_spent\_Small\_value\_payments’, ‘High\_spent\_Medium\_value\_payments’ and ‘High\_spent\_Large\_value\_payments’ category in Payment\_Behaviour. In contrast, the count for other categories in Payment\_Behaviour is minimal.
* For customers with poor credit scores, the majority align with a ‘Low\_spent\_Small\_value\_payments’ in Payment\_Behaviour, with the highest count reaching approximately 10,500. Intriguingly, ‘High\_spent\_Small\_value\_payments’, ‘Low\_spent\_Large\_value\_payments’, ‘High\_spent\_Large\_value\_payments’ Payment\_Behaviour exhibits the lowest count among poor credit scores.

Interest Rate with credit score

A graph of different colored lines

Description automatically generated

* Customers with a 'standard' credit score show a high concentration when the interest rate is in the range of 5 to 20. This suggests that a significant portion of customers with a 'standard' credit score tends to choose loan products with interest rates within this specified range.
* Customers with a good credit score exhibit a high concentration when the interest rate is in the range of 1 to 12. This indicates that a substantial number of customers with a good credit score prefer loan products with lower interest rates, possibly reflecting their creditworthiness.
* This implies that a notable percentage of customers with a poor credit score tends to opt for loan products with higher interest rates, possibly reflecting the increased risk associated with lending to individuals with lower creditworthiness.

Number of Loan2 with credit score

A graph of different colored bars

Description automatically generated

* People with a 'standard' credit score tend to have more loans when the number of loans is between 0 and 4. This suggests they are moderately conservative in taking on debt.
* Individuals with a 'good' credit score also prefer having fewer loans, especially in the range of 0 to 4. This indicates responsible borrowing and likely higher creditworthiness.
* Those with a poor credit score are more likely to have a higher number of loans, particularly in the range of 2 to 7. This could imply financial strain or higher risk associated with this group.

Encoding

Applied simple mapping technique to the 'Credit\_Score' column. The original values 'Standard,' 'Poor,' and 'Good' were converted to numerical representations for ease of analysis. Specifically, we assigned the values 1, 0, and 2, respectively, to these categories. This encoding facilitates the use of machine learning algorithms that require numerical input, allowing us to effectively incorporate the credit scores into our models. The transformation enhances the interpretability and compatibility of the data, contributing to a more streamlined and efficient analytical process.

Implemented label encoding as a technique to convert categorical data into numerical format. Specifically, we applied label encoding to variables where the categories have an inherent ordinal relationship. Label encoding assigns numerical labels in a way that reflects the hierarchy or order of these categories. By employing label encoding, we aim to enhance the efficiency of our models in handling categorical features, ultimately contributing to the overall accuracy and effectiveness of our project's analyses and predictions.





A screenshot of a computer program

Description automatically generated

Modelling

In the context of our project, the modelling phase plays a pivotal role in extracting meaningful patterns and making predictions based on the available data. Employing machine learning techniques, our modelling approach aims to uncover relationships and dependencies within the dataset. We utilize a systematic process of feature engineering, where relevant variables are identified and transformed to enhance their predictive power. The selection of an appropriate model is guided by the nature of the problem at hand, considering factors such as data distribution, complexity, and interpretability. Following model selection, rigorous training and validation procedures are implemented to ensure robust performance and mitigate overfitting. The effectiveness of the model is evaluated using various metrics, and fine-tuning may be applied to optimize its predictive accuracy. Throughout this modelling phase, our goal is to develop a reliable and interpretable model that contributes valuable insights to the objectives of our project, ultimately facilitating informed decision-making based on the data at hand.

|  |  |  |
| --- | --- | --- |
| Model Name | Accuracy Score | Cohen-Kappa-Score |
| Decision Tree | 0.71 | 0.52 |
| Random Forest | 0.64 | 0.36 |
| Xg Boost | 0.77 | 0.62 |
| AdaBoost | 0.65 | 0.42 |

We can conclude that XG boost is the best model with highest accuracy score and Cohen-Kappa Score. The top three features affecting our model are Credit Mix, Interest Rate and Outstanding Debt