Credit Score Classification

Group 11 - Section 005

Smit Rana (0792056)
Avi Kumar Patel (0790966)
Vaibhav Patel (0772934)
Yash Kumar Patel (0797825)

Instructors:

John Gerassimou
Manjari Maheshwari
St. Clair College of Applied Arts and Technology

Abstract

At the present time, everyone needs to have good credit because it is seen as a sign of trustworthiness. Financial institutions and credit card issuers use credit scores to determine creditworthiness. It enables banks and credit card companies to issue loans to customers with excellent creditworthiness immediately. In today's fast-paced world, doing a manual investigation into each portfolio and generating a credit report both take a significant amount of time. Most of the banks and credit card companies use Machine Learning algorithms to categorize all their customers based on their credit history. So, we want to make classification system for credit scores to get rid of having to do work by hand. The credit score classification project based on machine learning seeks to develop a predictive model that can accurately categorize a customer's creditworthiness based on their credit history, banking data, or financial behavior. The project will involve preprocessing, cleaning, visualization, and machine learning models. The efficient model will be evaluated based on its classification accuracy and other applicable metrics to assess its effectiveness in predicting class of credit score. As an outcome of the effort, we put in on this project, we have developed an application that can be used by financial institutions and other businesses to better categorize the customers they serve.

Acknowledgement

"We would like to express our deep gratitude to all those who contributed to the success of this project. We would like to thank St. Clair College for adding the Capstone Project course to our academic curricula.

We are also grateful to our advisors and mentors, John Gerassimou and Manjari Maheshwari, who provided beneficial feedback, direction, and expertise throughout the duration of the project. We would like to thank classmates for their insightful comments and recommendations.

We would also like to recognize the contributions made by Rohan Paris, who provided data on STATSO for this project. This endeavor would have been impossible without his support.

Finally, we would like to extend our deepest appreciation to our family and friends, who provided emotional support and encouragement throughout the duration of the endeavor. We were motivated and inspired by their unwavering support.

Once more, we want to say thanks to everyone who contributed to the success of this endeavor."

Contents

List of Figures

1 Introduction

Credit scores are utilized by lending institutions as a decision-making tool for monetary matters such as loans and credit cards. It is difficult to determine which credit score category we fall into and which goods we have the best chance of qualifying for because we can check it with a credit card company or on a personal finance website, only to find that it differs on another. This makes it difficult for us to know which products we have the best chance of qualifying for. When a lender pulls your credit score, they may request it from a different credit agency, such as Experian, Equifax, or TransUnion, and/or request a specific version that differs from the one you checked. Most credit scores consider the same factors, such as On-time and Delay days of bill payment, New Credit, Credit History, Outstanding Debt, and Credit Mix.

We want to know how to generate credit score classifications using a Machine Learning model. In addition, we discovered case study data for Credit Score Classification on Statso.

In this project, we gave careful consideration to the readily available data, conducted comprehensive preprocessing, cleaning, visualization, feature engineering and selection, and data reduction, and selected appropriate machine learning models. The subsequent deployment of a machine learning model to a web application enables greater accessibility, interactivity, integration with other tools and platforms, scalability, and continuous improvement via real-time data collection and feedback. These are some of the questions that will hopefully be answered by the project:

1. What are the main factors that influence a person's credit score?
2. How can machine learning algorithms be used to predict credit scores more accurately?
3. How can we verify that the credit score classification machine learning model is fair and bias-free, especially for age, SSN, and socio-economic status?
4. How can we make the credit score categorization machine learning model accessible and explainable, especially the features and variables utilized to predict?
5. What measures and indicators can we use to evaluate the machine learning model's impact on credit scoring efficiency, accuracy, and inclusivity?

2 Description of dataset

The data was gathered from an online source known as STATSO by our team. STATSO is a Data Science Community to Find Case Studies, Datasets and more. This site generates and compiles data from many online sources, and its authors then use this information to develop case studies that show how various types of data can be used to address specific problems.

We have 100,000 banking data for credit score classification based on 28 features.

1 **ID:** The record's unique number.
2 **Customer_ID:** The customer's unique ID number
3 **Month:** A time of year
4 **Name:** The person's name
5 **Age**: The person's age
6 **SSN:** The person's Social Security Number
7 **Occupation**: What a person does for a living.
8 **Annual_Income:** The person's annual income
9 **Monthly_Inhand_Salary:** The person's monthly salary in cash
10 The **number of bank accounts** that the person has.
11 **Num_Credit_Card**: The number of credit cards someone has
12 **Interest_Rate**: The rate of interest on the person's credit card.
13 The **number of loans** the person has taken out from the bank.
14 **Type_of_Loan:** The different kinds of loans that a person has taken out from the bank.
15 **Delay_from_due_date**: The average number of days between the date of payment and when the person paid.
16 **Num_of_Delayed_Payment**: Number of payments that the person has been late on.
17 **Changed_Credit_Card_Limit**: The percentage increase or decrease in the person's credit card limit.
18 **Num_Credit_Inquiries:** The number of times the person has asked about their credit card.
19 **Credit_Mix:** People are put into three groups based on the types of credit accounts they have, such as mortgages, loans, credit cards, etc.
20 **Outstanding_Debt**: The amount a person still owes.
21 **Credit_Utilization_Ratio:** The customer's credit card's credit usage ratio
22 **Credit_History_Age:** How old a person's credit past is.

23 **Payment_of_Min_Amount:** Yes, if the person only paid the minimum amount to be paid. No, if more than the minimum amount was paid.
24 **Total_EMI_per_month** is the person's overall EMI per month.
25 **Amount_invested_monthly:** The amount a person invests each month.
26 **Payment_Behavior**: How a person pays for things.
27 **Monthly_Balance** is the amount of money left in a person's account at the end of each month.
28 **Credit_Score**: A person's credit score.

In this problem, the Credit Score column is the target variable. ID, Customer ID, Name, and SSN are not needed to figure out a credit score because they have a lot of unique numbers and don't make reasoning.

3 Method of Analysis

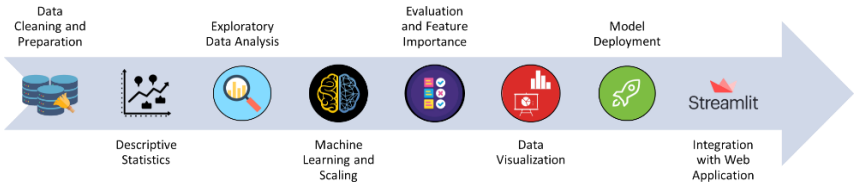

Figure 1 Workflow

The initial phase of analysis is to **clean data** for interpretation. In 100000 data records, there were nearly 60000 absent values. The absence of data can compromise the accuracy and trustworthiness of our findings, as well as introduce bias and weaken our analysis. Consequently, it is essential to handle absent data appropriately. The deletion of empty values is a common method used to clean data. We may choose to delete records with missing values, but only if the number of missing data is comparatively small. Nonetheless, if the absent data is substantial, this could result in the loss of vital information.

Strategy was to impute it using the other data that was presented. To achieve this goal, first partition the process into two unique categories for the columns, namely, numerical and categorical. When it comes to category NA values, the most recent non-null value that was supplied by the customer will be used in an effort to try and update the NA. When working with numerical examples, the average value of the variable is substituted for the null values that are present in the examples.

There was a column in the dataset referred to as "Credit_History_Age," which had data in the form "XX years and XX months." However, this information was unnecessary for both the analysis and the ML modeling. As a result, change was done on those columns to only contain the months.

To complete the process of preparing our data to be incorporated into the model, label encoder to the text-based column.

```
mdf = pd.concat([bdf.isnull().sum(), df.isnull().sum()], axis=1, keys=['Before', 'After'])
print(mdf)
```

	Before	After
ID	0.0	NaN
Customer_ID	0.0	NaN
Month	0.0	0.0
Name	9985.0	NaN
Age	0.0	0.0
SSN	0.0	NaN
Occupation	0.0	0.0
Annual_Income	0.0	0.0
Monthly_Inhand_Salary	15002.0	0.0
Num_Bank_Accounts	0.0	0.0
Num_Credit_Card	0.0	0.0
Interest_Rate	0.0	0.0
Num_of_Loan	0.0	0.0
Type_of_Loan	11408.0	NaN
Delay_from_due_date	0.0	0.0
Num_of_Delayed_Payment	7002.0	0.0
Changed_Credit_Limit	0.0	0.0
Num_Credit_Inquiries	1965.0	0.0
Credit_Mix	0.0	0.0
Outstanding_Debt	0.0	0.0
Credit_Utilization_Ratio	0.0	0.0
Credit_History_Age	9030.0	NaN
Payment_of_Min_Amount	0.0	0.0
Total_EMI_per_month	0.0	0.0
Amount_invested_monthly	4479.0	0.0
Payment_Behaviour	0.0	0.0
Monthly_Balance	1200.0	0.0
Credit_Score	0.0	0.0

The presence of **NaN** in this context suggests that those columns were eliminated during the data cleaning procedure.

Figure 2 Data Cleaning Result

The data were then analyzed using **descriptive statistics** for each column of data. And what we found was that some of the values contained within it were absolutely useless, while others were a lot further away from the majority of the data points. These individuals are known as outliers.

After gathering information from a wide number of sources, we have concluded regarding the likely range of values for the variable. The maximum number of credit cards, bank accounts, credit inquiries, and loans that an individual is permitted to have at any given time. Because the values were so high, it is necessary to do this step-in order to give the data any kind of sense.

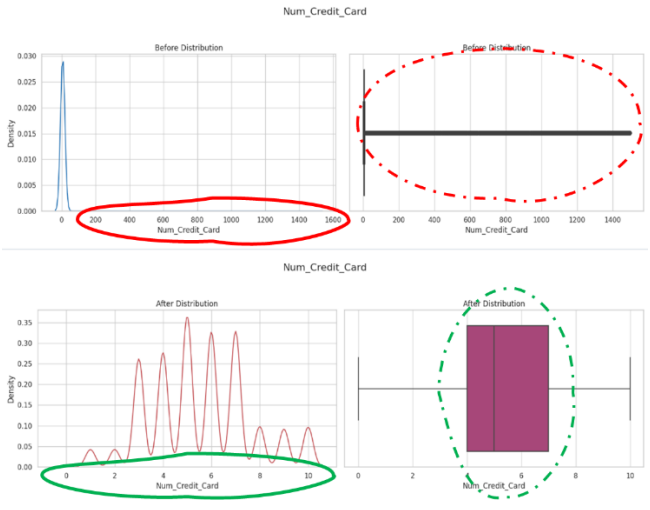

Figure 3 Outlier Handling Result

After dealing with the outliers, the next step in the EDA process was to take the columns that had been handled and determine the relationship between those columns and our target variable.

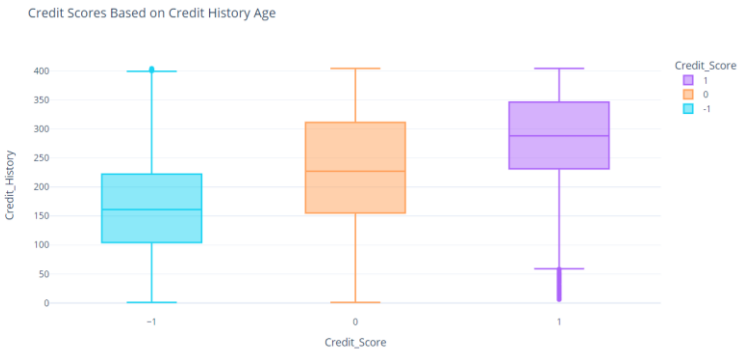

Figure 4 Credit History with Credit Score

Higher credit ratings are achieved by those who have a lengthy credit history.

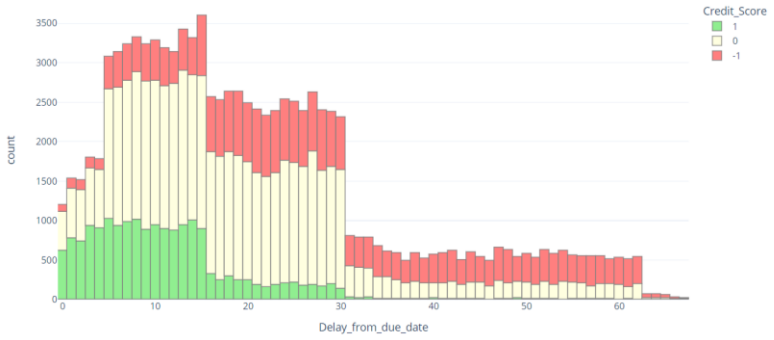

Figure 5 Delayed Day of Payment vs Credit Score

Features of the input data have varying scales; therefore, the machine learning algorithm may give greater weight to features with larger values and discard features with smaller values. This can impair the model's performance on features with lesser values, which may be essential to prediction. Having features that are extremely related can result in overfitting or decreased model performance.

Scaling down data is significant since it helps to normalize the data and bring the features into a comparable range. Prior to fitting the data to the model, we use MinMax Scaler. It scales the attributes of a dataset to a predetermined range, typically between 0 and 1, by subtracting the minimum value of each attribute and dividing by the attribute's range. It can enhance the performance of specific machine learning algorithms and simplify the comparison of the relative significance of different features.

This problem could entail classifying credit score into three distinct categories, such as "poor," "standard" and "good." In this situation, we are working with a multiclass classification problem, so it is not possible utilize a binary classification metric. Accuracy, Confusion Matrix, Correlation Plot, and F1 Score are going to be the performance measurements that we use. During this process, there seven classifiers : Random Forest, Decision Tree, XGBoost, AdaBoost, KNN, and Logistic Regression. Voting has also been utilized. (The purpose of the voting classifier is to aggregate the predictions from each individual model and generate a final prediction based on the majority vote of the individual predictions.) For each, performance indicators and plotted the prediction pattern and feature significance graphs are utilized. This gives us the ability to select the model that will work best for integrating it into the production environment so that it can make predictions based on new data as it is being collected.

4 Results

After analyzing the performance of each classifier using criteria such as accuracy, confusion matrix, correlation plot, and F1 score, as well as prediction pattern and feature importance graphs, the Random Forest Classifier emerged as the most successful model for the data.

	Train Score float...	Test Score float64	Macro - Averag...	Weighted - Aver...
	Accuracy		F1 Score	
Random Forest	0.7808	0.7409	0.7249	0.7423

Figure 6 Accuracy and F1 Score

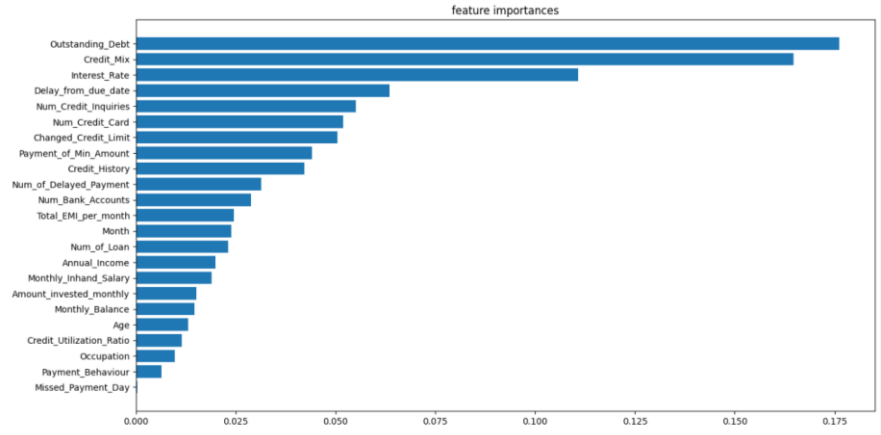

Figure 7 Feature Importance Graph

```
evaluate_classification(rf, "RandomForestClassifier", x_train,x_test,y_train,y_test)
```

```
Training Accuracy RandomForestClassifier 78.21285714285715  Test Accuracy RandomForestClassifier 74.08666666666667
```

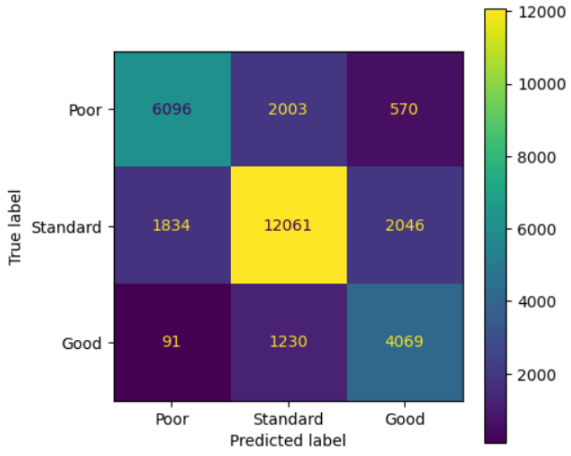

Figure 8 Confusion Matrix

Feature importance is a metric used in machine learning to identify the most significant input variables (or features) that contribute to a model's ability to predict. In other words, it helps in determining which features are most crucial for predicting the target variable. There are 23 features in our data that can determine the class of our objective variable. Nevertheless, some features have a negligible impact or have no real-world significance. Adding 23 fields to an application is also not an ideal choice. Therefore, we selected the 11 most important features for our model deployment. That will be used for application integration.

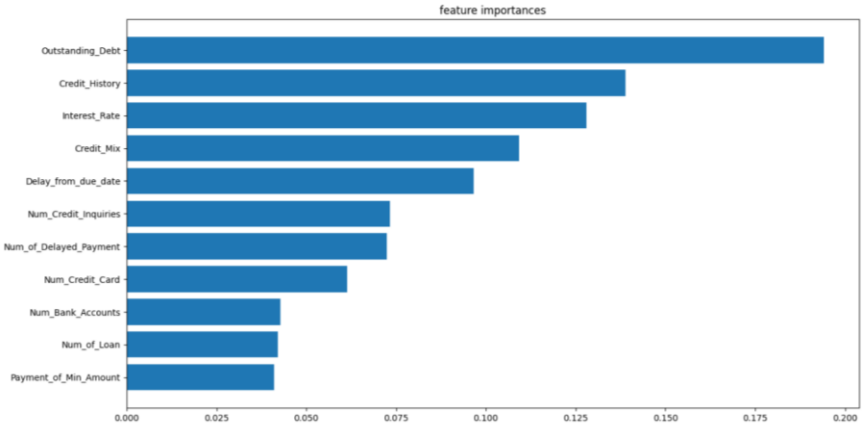

Figure 9 Feature Importance Graph with 11 Features

Based on these 11 features, we generate a few data visualizations to better comprehend the relationships among them and provide insights.

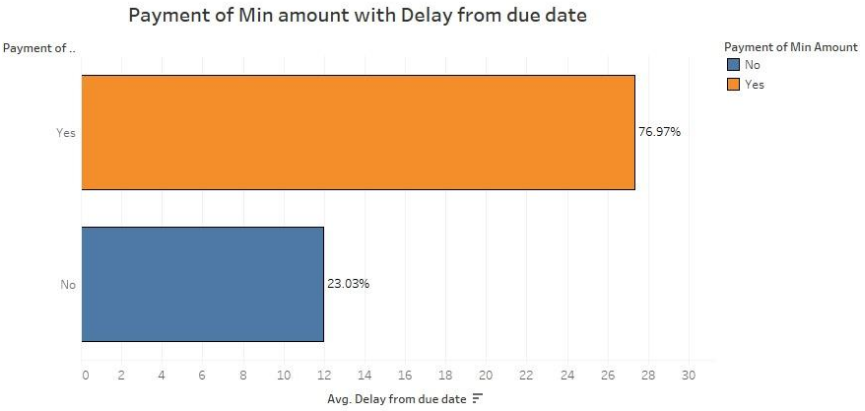

Figure 10 Pay of Min Amount with Delayed days of bill.

This visualization displays the customers' payment patterns over time based on whether they make the minimum payment by the deadline or postpone their payments.

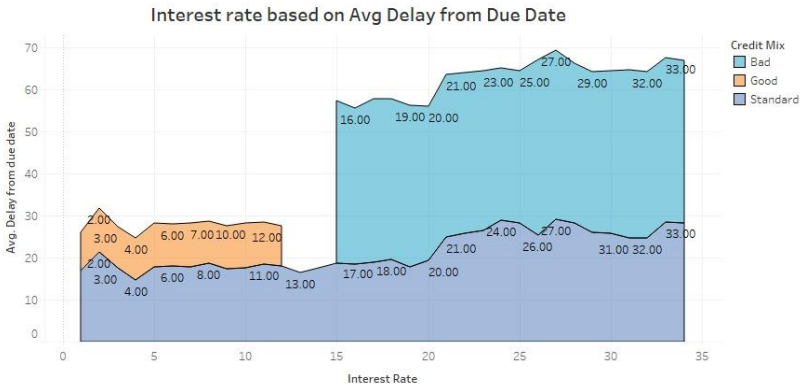

Figure 11 Interest Rate on Credit Card by Delayed Days

This chart displays the average interest rate over time for average Delay from due date. and shows the trends or correlations in the interest rate based on customers' historical payment behaviors and talks about their credit mix.

An entirely novel strategy with the model and put it into action. The.obj file type is used for the export of both the Random Forest model and the MinMax model. The trained model can be saved on a drive using this format, and then loaded back into memory whenever it is necessary to do so. It is simple to interface with other applications and may be used to transfer the trained model to production environments where it may be used. This is possible since it is easy to integrate. We additionally export Scaler from this part of the ML model because, in the case that the application receives input from the real world, we will need to scale those values down before feeding them to the Random Forest model. This is because the application will be able to handle the input better.

```
path = os.path.dirname(__file__)
folder_path = os.path.join(path,'..\\models')

@st.cache_resource()
def unzip_load(name):
    path_zip = os.path.join(path,'C:/Users/iamge/Desktop/code/models/' +name+ '.zip')
    ZipFile(path_zip).extractall(folder_path)
    path_obj = os.path.join(path,'C:/Users/iamge/Desktop/code/models/' +name+ '.obj')
    return  pickle.load(open(path_obj, 'rb'))

scaler = unzip_load('scaler')
rf = unzip_load('rf')


    if run:
        resp = {
            'Num_Credit_Card': Num_Credit_Card,
            'Num_Bank_Accounts': Num_Bank_Accounts,
            'Outstandin_Debt': Outstandin_Debt,
            'Interest_Rate' : Interest_Rate,
            'Delay_from_due_date': Delay_from_due_date,
            'Num_of_Delayed_Payment' : Num_of_Delayed_Payment,
            'Credit_mix': Credit_mix,
            'Num_of_Loan': Num_of_Loan,
            'Payment_of_Min_Amount': Payment_of_Min_Amount,
            'Credit_history': Credit_history,
            'Num_Credit_Inquiries': Num_Credit_Inquiries
        }
        output = transform_resp(resp)
        output = pd.DataFrame(output, index=[0])
        output.loc[:,:] = scaler.transform(output)

        credit_score = rf.predict(output)[0]

        if credit_score == 1:
            st.success('Your credit score is **GOOD**! Congratulations!', icon="✅")
            st.markdown('This person's credit score suggests they are likely to make their loan payments on time, making them a low credit risk.
        elif credit_score == 0:
            st.warning('Your credit score is **REGULAR**', icon="⚠")
            st.markdown('This person's credit score shows that they are likely to pay back a loan, but they may sometimes miss a payment. This s
        elif credit_score == -1:
            st.error('Your credit score is **POOR**.', icon="🚨")
            st.markdown('This person's credit score indicates they are unlikely to redeem a loan, so the risk associated with extending them cre

        prob_fig, ax = plt.subplots()

        with st.expander('Click to see how certain the algorithm was'):
            plt.pie(rf.predict_proba(output)[0], labels=['Poor', 'Regular', 'Good'], autopct='%.0f%%')
            st.pyplot(prob_fig)
```

Figure 12 Model Deployment Code

The Resp {} begins by gathering data from the real world, after which the scaler scales this data down, draws conclusions from the model, and generates output based on three if-elif conditions.

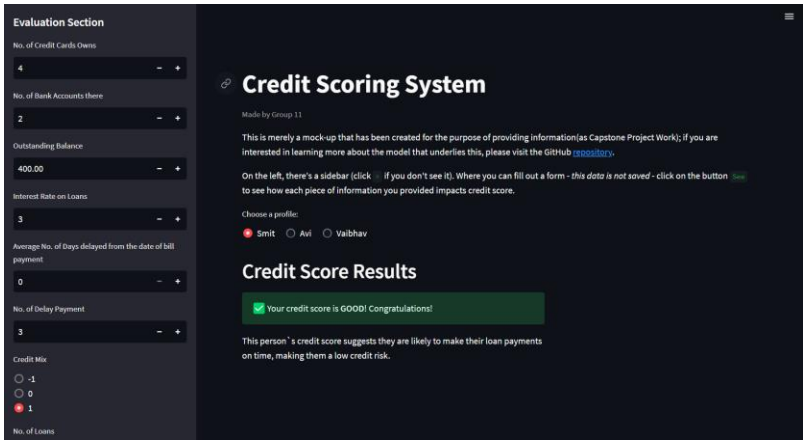

Figure 13 Model Integration with Streamlit (1)

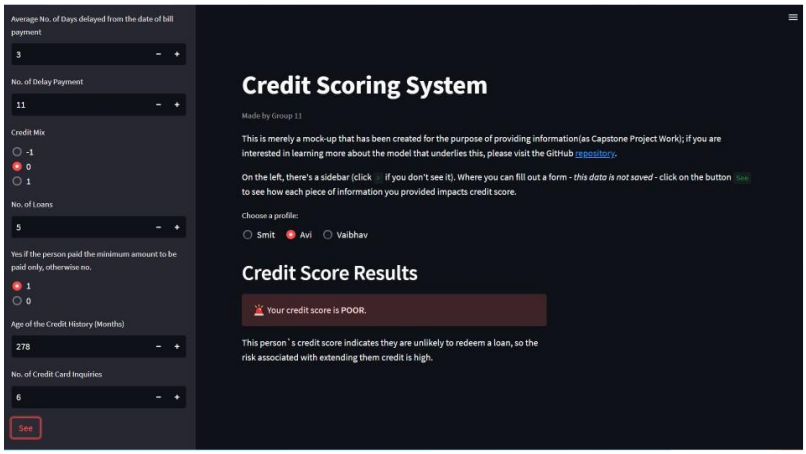

Figure 14 Model Integration with Streamlit (2)

FICO (Fair Isaac Corporation) is a lead in the development of a method for calculating credit scores using credit reporting agency data. FICO is used in the United States, and our data is from the United States because it contains the SSN variable. We therefore compare our final conclusion regarding the weighting of credit score factors with FICO factors.

Our Credit Score System

- 20% outstanding debt
- 15% Credit History
- 14% Interest Rate
- 10% Credit Mix
- 10% Delay days of bill payment
- 38% (Credit Inquires, No. of Credit Card, Bank Account, Loan and Payment of Min Amount)

FICO Score System

- 35% On-time and Delay days of bill payment
- 30% Outstanding Debt
- 15% Credit History
- 10% Credit Mix
- 10% New Credit

Figure 15 Weighting Factors Comparison

5 Discussion

Data Cleaning and Preparation was one of the key features of our analysis because we utilized several different approaches to clean the data without removing a single record from the data set. Another highlight was the ability to select features based on the model being used. We explored with seven distinct classifiers and evaluated each one based on several significant metrics for performance. The Features were selected for the efficient model that we decided to go with. (Based on feature importance graph). The model deployment and integration with the web-based application "Streamlit" was the standout element of our study. Results of this project as well as the approaches that we have employed confirm the initial assumptions.

As we progress with our analysis, we discover that our data was somewhat dummy Because one of the reasons for using that data source was for case study purposes. Other values in the dataset did not adequately reflect the actual data's properties. We experienced difficulties as well. One was that we had missing and false values, which affected the analysis. Preparing the data for analysis required extensive preprocessing, including feature scaling, feature engineering, and outlier detection. In addition, target variable classes are unbalanced. This may skew data and predictions. Finally, when Features are highly interrelated, the model may be overfit to the training data and perform poorly on new data.

After doing analysis, it is essential to ensure that the data is accurate, complete, and up to date before beginning the task at hand. Try to collect data that is representative of the actual world. Use a diverse dataset. Also, ensure that the project complies with ethical and legal standards, including privacy regulations and anti-discrimination laws, if the data is from the real world.

At last, If someone wants to develop our project in the future, they should consider new data of high quality. They can execute our processes and workflow to complete the project. When the application receives new real-time data, the individual may experiment with dynamic displays or retrain the model using the new data.

6 Conclusion

In a nutshell a credit score classification project entails performing data analysis on borrowers and developing a model that can predict the creditworthiness of individuals. The purpose of this project is to develop a credit score system that is both fair and precise, with the intention of assisting financial institutions and credit card businesses in rapidly issuing loans to consumers who have an excellent creditworthiness.

Moreover, credit score classification initiatives can have substantial social and economic ramifications, as they influence the ability of individuals and businesses to access credit and attain their financial objectives. Consequently, it is essential to ensure that the project is executed in a transparent and ethical manner with recommendations which are provided and that the model does not discriminate against certain groups of borrowers.

Lastly, a credit score classification project is a process that is iterative, and it is crucial to continuously evaluate and refine the model over time to ensure that it remains accurate and up to date by doing so, financiers can remain ahead of forthcoming developments and risks in the lending industry and make more effective credit extension decisions.

References

[1] https://statso.io/credit-score-classification-case-study/
[2] https://www.myfico.com/credit-education/whats-in-your-credit-score#:~:text=FICO%20Scores%20are%20calculated%20using,and%20credit%20mix%20(10%25)
[3] https://www.equifax.com/personal/education/credit/score/what-is-a-credit-score/#:~:text=300-579%3A%20Poor,740-799%3A%20Very%20good
[4] http://www.diva-portal.org/smash/get/diva2:1664698/FULLTEXT01.pdf
[5] https://www.scitepress.org/papers/2018/68236/68236.pdf
[6] https://en.wikipedia.org/wiki/Criticism_of_credit_scoring_systems_in_the_United_States
[7] https://www.datrics.ai/credit-scoring-using-machine-learning
[8] https://user-images.githubusercontent.com/33239902/183321842-be97fb04-f00b-4b62-8e6e-2b53d25335a0.gif