Department of Data Science and Statistics
III Trimester MSc Data Science

Machine Learning - MDS372

PROJECT REPORT

on

'CREDIT SCORE PREDICTION USING MACHINE LEARNING TECHNIQUES'

by

ANNU PUNNOOSE [2348011]
DHILLIPKUMAR M [2348026]

April 2024

ABSTRACT

Accurate credit score prediction is paramount for financial institutions in assessing creditworthiness and mitigating lending risks. This study proposes a machine learning-based approach to develop a robust credit score prediction system using a comprehensive dataset comprising 100,000 records and 28 attributes related to personal and financial information.

Through rigorous data exploration and feature engineering techniques, the study investigates the intricate relationships between various factors, such as income, employment status, credit card usage, loan history, and payment behavior, and their impact on credit scores. Multiple machine learning algorithms, including Support Vector Machines (SVM), logistic regression, random forest, and clustering techniques like K-means and agglomerative clustering, are evaluated to identify the most effective model for credit score prediction.

The results demonstrate that the random forest algorithm outperforms other techniques, achieving an impressive accuracy of 81.225%. This superior performance can be attributed to the ensemble nature of random forest, which effectively captures complex patterns and relationships within the dataset. Conversely, while clustering methods provide valuable insights into data structure, their predictive accuracy remains relatively lower, limiting their applicability for this specific task.

A key highlight of this study is the successful deployment of the high-performing random forest model into a user-friendly web application using Flask, a lightweight Python web framework, and PyCharm, an integrated development environment (IDE). The deployed application seamlessly integrates the trained model, allowing users to input their personal and financial data through an intuitive interface. The application then leverages the predictive power of the random forest model to generate accurate credit score predictions in real-time.

This deployment approach not only showcases the practical applicability of the developed model but also facilitates its integration into real-world financial services workflows. Financial institutions and lenders can leverage this application to streamline credit risk assessment processes, enabling efficient and informed decision-making regarding loan approvals, credit card issuance, and other financial services.

The study underscores the potential of machine learning techniques, particularly random forest, in accurately predicting credit scores, offering a valuable tool for financial institutions to streamline credit risk assessment and enhance decision-making processes. The successful deployment of the model through a user-friendly web application further demonstrates its practical utility and paves the way for its widespread adoption in the financial sector.

TABLE OF CONTENTS

1. INTRODUCTION

Credit scores are a fundamental aspect of financial decision-making, serving as a key indicator of a person's creditworthiness. Financial institutions and lenders rely on these scores to assess the risk associated with granting loans, credit cards, or mortgages. With the growing importance of credit scores, there is an increasing need for reliable methods to predict them. This project explores a machine learning-based approach to credit score prediction, using a comprehensive dataset of customer financial information.

The data at the heart of this project includes a variety of attributes, such as income, employment status, credit card usage, loan history, and payment behavior. Our objective is to build models that can predict an individual's credit score based on these factors. By utilizing different machine learning techniques like Support Vector Machines (SVM), logistic regression, and random forest, we aim to achieve high accuracy in our predictions.

This report covers the various stages of the project, from data preparation and model training to model evaluation and deployment. The inclusion of hierarchical clustering through k-means offers a unique perspective on credit score prediction, allowing for categorization into broader classes such as "Good," "Standard," and "Bad." This categorization can help financial institutions quickly assess credit risk.

In addition to the technical aspects, the project encompasses a user-friendly front-end application, developed with Flask and PyCharm, enabling users to interact with the prediction system. This feature demonstrates the practical applications of machine learning in a real-world context, illustrating how such a system can streamline credit evaluations.

Overall, this report aims to provide a comprehensive overview of the credit score prediction project, highlighting the significance of credit scores in the financial industry and the potential for machine learning to improve accuracy and efficiency in credit assessment. By exploring multiple machine learning techniques and developing a user-friendly interface, this project offers valuable insights for both academic research and practical implementation in the financial sector.

2. DATA EXPLORATION & ANALYSIS

2.1.Data Understanding and Exploration:

The dataset used in this project consists of 100,000 records and 28 attributes. These attributes encompass a range of personal and financial information, such as income, credit card usage, loan history, and credit score classifications. The target variable for our analysis is "Credit_Score," which categorizes creditworthiness into "Good," "Standard," and “Poor."

Given the diverse range of variables, our goal is to understand the distribution and characteristics of the data to identify any patterns or correlations that may impact credit scores. We began by exploring the relationships between credit scores and various features, including occupation, annual income, monthly in-hand salary, number of bank accounts, and more.

2.2.Data Cleaning and Handling Missing Values:

One of the primary challenges in data analysis is handling missing values. Fortunately, our dataset did not contain any missing values, allowing for a smoother analysis process. This eliminated the need for complex imputation strategies, ensuring the integrity of our results.

2.3.Data Integration and Feature Engineering:

The dataset contained all the necessary information for analysis, so additional data integration was not required. However, feature engineering played a key role in deriving meaningful insights from the existing data. Feature engineering involved transforming existing attributes to create new features that might enhance the predictive power of our models.

One such feature was the "Credit Utilization Ratio," calculated as the ratio of total debt to total available credit. This ratio is a common metric used to evaluate an individual's credit risk. Additionally, we derived features to measure the age of the credit history, as a longer credit history is generally associated with better credit scores.

2.4.Data Exploration:

Our exploration focused on understanding the relationships between various features and credit scores. We used visualizations to uncover patterns and identify factors that might affect credit scores.

1. **Occupation**: A box plot showed that there wasn't much difference in credit scores across different occupations, indicating that a person's job title might not have a significant impact on their credit score.

2. **Annual Income**: A box plot revealed that higher annual income is generally associated with better credit scores, suggesting a positive correlation between income and creditworthiness.

3. **Monthly In-Hand Salary**: A similar pattern was observed for monthly in-hand salary, indicating that a higher monthly income tends to result in better credit scores.

4. **Number of Bank Accounts**: This exploration showed that having more than five bank accounts negatively impacts credit scores. The optimal number of bank accounts for a good credit score appears to be 2-3.

5. **Number of Credit Cards**: Having 3-5 credit cards is ideal for a good credit score. Having more credit cards does not positively impact credit scores.

6. **Average Interest Rate**: An average interest rate of 4-11% is associated with good credit scores, while rates above 15% negatively impact credit scores.

7. **Number of Loans**: To maintain a good credit score, individuals should not take more than 1-3 loans at a time. More than three loans negatively impact credit scores.

8. **Payment Delays**: Delaying payments for more than 17 days from the due date negatively affects credit scores. Delays between 5-14 days do not have a significant impact.

9. **Number of Delayed Payments**: Delaying 4-12 payments does not affect credit scores, but delaying more than 12 payments negatively impacts credit scores.

10. **Outstanding Debt**: Debt between 380-1150 does not affect credit scores, but consistent debt above 1338 negatively impacts them.

11. **Credit Utilization Ratio**: This ratio appears to have little impact on credit scores, suggesting other factors are more critical in determining creditworthiness.

12. **Credit History Age**: Longer credit history correlates with better credit scores, highlighting the importance of maintaining a lengthy credit history.

13. **Total EMIs Per Month**: The total number of EMIs paid in a month does not significantly impact credit scores.

14. **Monthly Investments**: The amount invested monthly does not appear to significantly impact credit scores.

15. **Monthly Balance**: A higher monthly balance at the end of the month is associated with better credit scores. A balance below $250 is detrimental to credit scores.

These insights provided a solid foundation for our machine learning models and helped guide our predictive analysis.These stages of data pre-processing and exploration are essential for laying the groundwork for the subsequent application of machine learning techniques to predict Credit Score.

3. ALGORITHM IMPLEMENTATION

We implemented Support Vector Machines (SVM), logistic regression, and random forest to develop predictive models for credit scores. These algorithms were chosen for their effectiveness in handling structured data and their ability to offer different perspectives on the problem.

1. Support Vector Machines (SVM)

Support Vector Machines (SVM) is a supervised learning algorithm that is particularly effective in high-dimensional spaces. It works by finding the hyperplane that best separates different classes, making it useful for classification tasks like credit score prediction. We chose SVM for this project because of its robustness and capability to handle complex data relationships.

SVM's strength lies in its ability to create clear decision boundaries, even with non-linear data. Given our dataset's structure, which includes multiple features contributing to credit score classification, SVM's flexibility made it a suitable choice. However, SVM can be computationally intensive, especially with large datasets, which can lead to longer training times.

2. Logistic Regression

Logistic regression is a straightforward algorithm used for binary or multi-class classification. It models the probability of a given outcome and is based on the logistic function. We chose logistic regression because of its simplicity, interpretability, and efficiency in binary classification tasks.

Despite its simplicity, logistic regression is highly effective in situations where the relationship between variables is linear. It is also less computationally demanding than SVM, making it a practical choice for large datasets. In our project, logistic regression provided a baseline for comparison with more complex algorithms like random forest.

3. Random Forest

Random forest is an ensemble learning method that combines multiple decision trees to improve classification accuracy and reduce overfitting. Each tree in the forest is trained on a random subset of the data, with the final prediction determined by aggregating the results from all trees. This approach enhances generalization and reduces the risk of overfitting.

We selected random forest because it offers a high degree of accuracy and is less sensitive to outliers and noise in the data. Given the dataset's size and complexity, random forest was a strong candidate due to its scalability and robustness. Additionally, random forest provides feature importance scores, allowing us to understand which features most significantly impact credit scores.

4. K-Means Clustering

K-means is a popular clustering algorithm that partitions data into a specified number of clusters by iteratively updating cluster centroids to minimize the sum of squared distances between data points and their assigned centroids. The algorithm is efficient and well-suited for large datasets, making it a common choice for unsupervised learning tasks.

We used K-means because it is computationally efficient, scalable, and straightforward to implement. Given our large dataset, K-means offered a fast and reliable way to create clusters for further analysis. The algorithm's simplicity allows for quick convergence, making it ideal for exploratory data analysis when the number of clusters is known or can be reasonably estimated.

5.Agglomerative Clustering

Agglomerative Clustering, also known as hierarchical clustering, is a bottom-up approach to clustering. It begins by treating each data point as its own cluster and then iteratively merges clusters based on a defined linkage criterion, such as Ward's method or complete linkage. This process continues until a stopping condition is met, such as reaching a specified number of clusters or a certain distance threshold.

Agglomerative Clustering is useful for uncovering hierarchical relationships among data points. We used Agglomerative Clustering to explore hierarchical relationships and understand the data's complexity. It allows for flexibility in choosing the linkage method and doesn't require specifying the number of clusters in advance, which can be advantageous in exploratory analysis.

Challenges with Agglomerative Clustering

While Agglomerative Clustering can be informative, its computational complexity increases with large datasets. Due to the size of our dataset, this method was found to be impractical for comprehensive clustering. So we took only 10000 rows in order to apply agglomerative clustering. The significant memory and processing requirements limited its usability, which led us to rely more on K-means for unsupervised learning tasks.

4. MODEL EVALUATION & PERFORMANCE ANALYSIS

Support Vector Machines (SVM)

Support Vector Machines (SVM) is known for its robustness in high-dimensional spaces and its ability to create clear decision boundaries. However, its performance depends on the complexity and structure of the data. With an accuracy of 53.32%, SVM showed moderate success in predicting credit scores. While it is suitable for structured data, SVM's performance in this case might suggest that the relationships between the features and credit scores are more complex than a linear classifier can handle.

Logistic Regression

Logistic regression is a straightforward and efficient algorithm used for binary and multi-class classification. In this project, logistic regression achieved an accuracy of 54.14%, slightly higher than SVM. This indicates that logistic regression might be capturing some of the linear relationships within the data, but it still falls short of a high-performance model.

Random Forest

Random forest, an ensemble learning method, combines multiple decision trees to improve accuracy and reduce overfitting. With an accuracy of 81.225%, random forest significantly outperformed both SVM and logistic regression. This high accuracy indicates that random forest is capable of capturing complex patterns and relationships in the dataset. The ensemble approach, with its ability to generalize well and handle noise, makes random forest a strong candidate for credit score prediction.

K-means Clustering

K-means clustering, a commonly used unsupervised learning algorithm, achieved an accuracy of 27.86%. While K-means is efficient and scalable, this relatively low accuracy suggests that the inherent assumptions of K-means—specifically, spherical and evenly sized clusters—may not align with the underlying structure of the dataset. Despite this, K-means can still be useful for exploratory data analysis and identifying general patterns.

Agglomerative Clustering

Agglomerative Clustering recorded an accuracy of 23.5%. This lower accuracy, compared to other algorithms, might reflect the increased complexity and computational demands of this method. Agglomerative Clustering's emphasis on hierarchical relationships makes it more suitable for smaller datasets or scenarios where cluster structure is a key focus. Given the size of our dataset, Agglomerative Clustering's performance indicates that it may not be the best choice for this credit score prediction task.

4.2.Comparative Analysis of Different Models:

Comparing the accuracy of the three algorithms, it is clear that random forest is the best-performing model for predicting credit scores. Its ensemble nature allows it to balance between bias and variance, leading to better generalization. The significant difference in accuracy compared to SVM and logistic regression suggests that the credit score prediction task benefits from the robustness and flexibility of random forest.

While SVM and logistic regression showed moderate accuracy, they provide a baseline for comparison and highlight the need for more sophisticated algorithms when dealing with complex datasets like this one. The high accuracy of random forest demonstrates its suitability for this task and its potential for practical applications in the financial industry.

K-means clustering, a commonly used unsupervised learning algorithm, achieved an accuracy of 27.86%. While K-means is efficient and scalable, this relatively low accuracy suggests that the inherent assumptions of K-means—specifically, spherical and evenly sized clusters—may not align with the underlying structure of the dataset. Despite this, K-means can still be useful for exploratory data analysis and identifying general patterns.

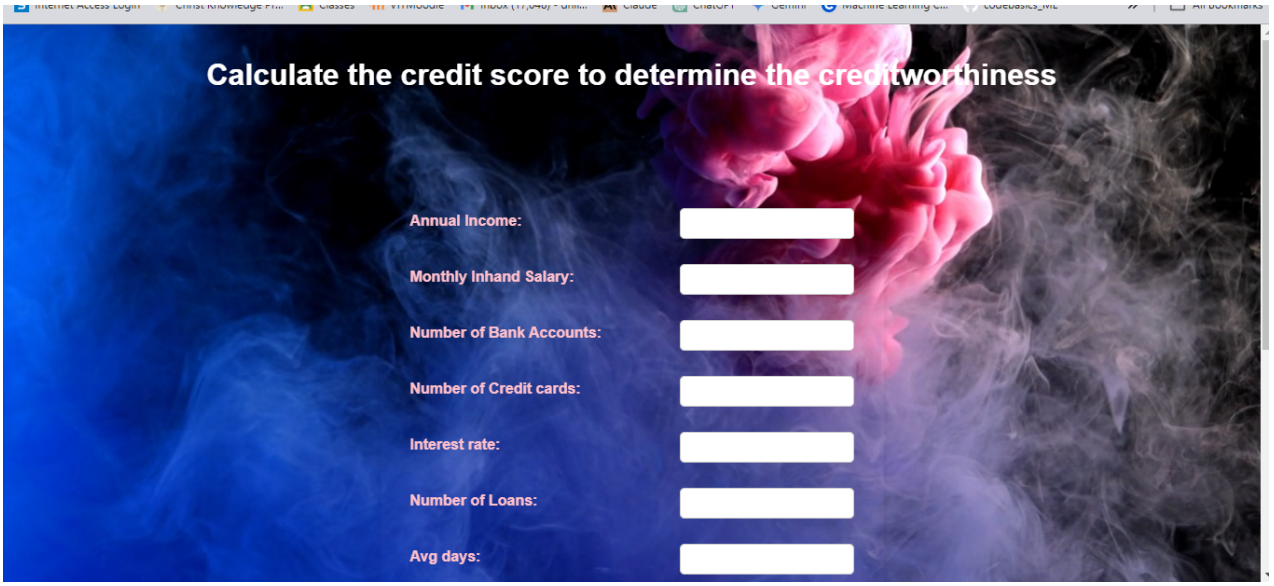
Agglomerative Clustering, a hierarchical approach to clustering, recorded an accuracy of 23.5%. This lower accuracy, compared to other algorithms, might reflect the increased complexity and computational demands of this method. Agglomerative Clustering's emphasis on hierarchical relationships makes it more suitable for smaller datasets or scenarios where cluster structure is a key focus. Given the size of our dataset, Agglomerative Clustering's performance indicates that it may not be the best choice for this credit score prediction task.

In summary, the model evaluation and performance analysis reveal that random forest is the most effective algorithm for credit score prediction in this dataset. It's high accuracy indicates that it can serve as a reliable tool for credit risk assessment, offering valuable insights for lending institutions and financial services.

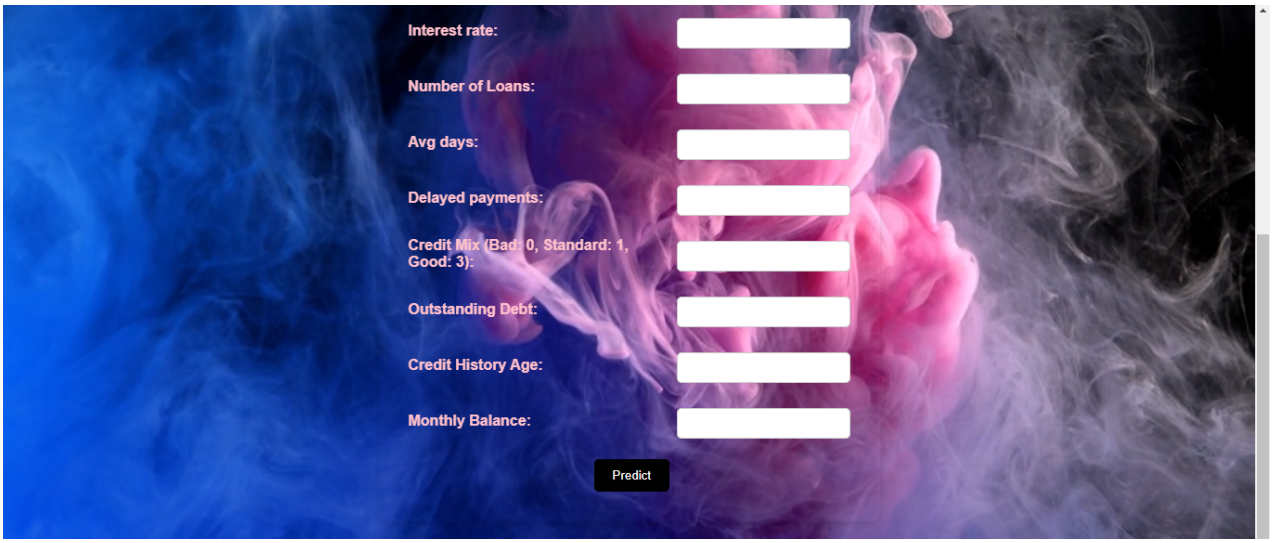
5. MODEL DEPLOYMENT

Model deployment is the final stage in a machine learning project where the developed model is integrated into a real-world application or system. In our project, we deployed the credit score prediction model using Flask, a lightweight web framework for Python, and Pycharm, an integrated development environment (IDE) for Python. For the prediction task, we chose the random forest algorithm due to its high accuracy and robustness that we have achieved in our project.

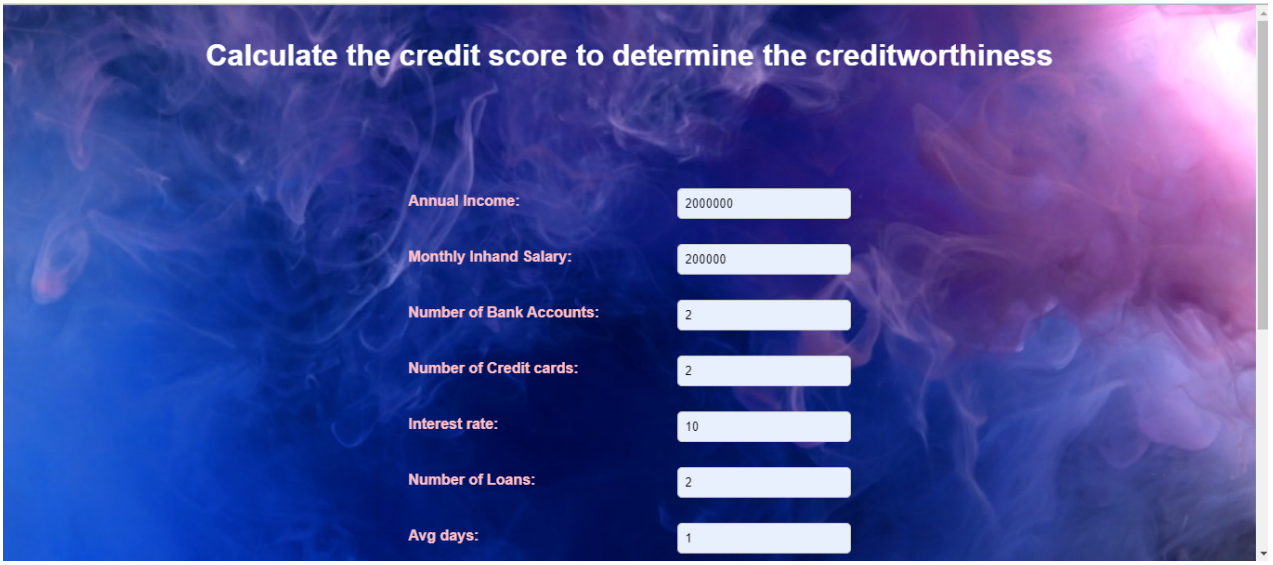
Output Screenshots:



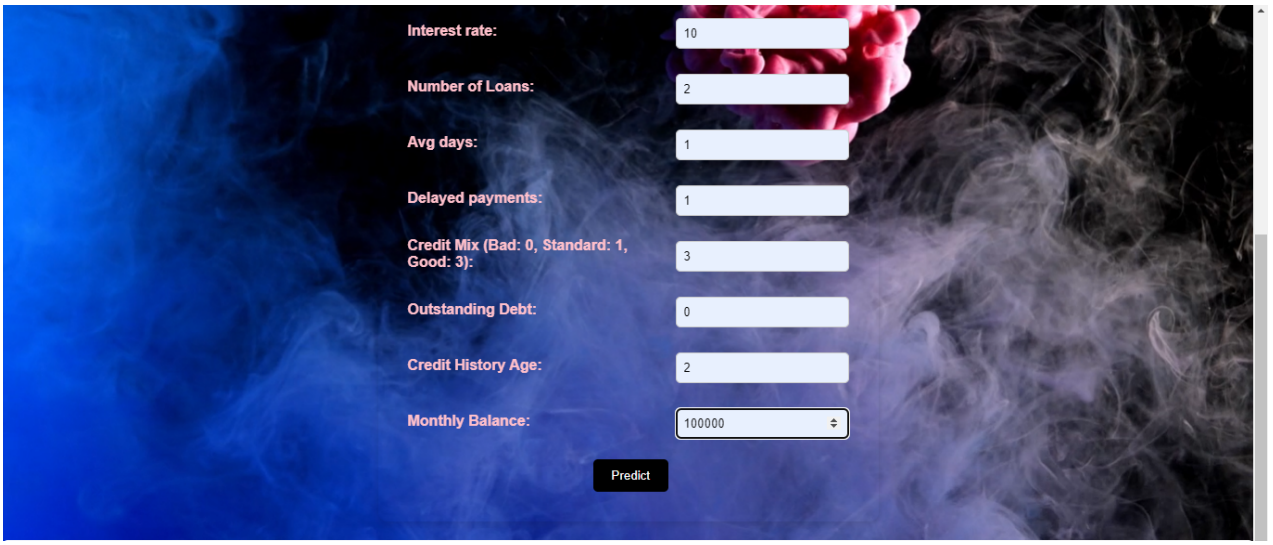
Screenshot No:1- Front-end For User Inputs



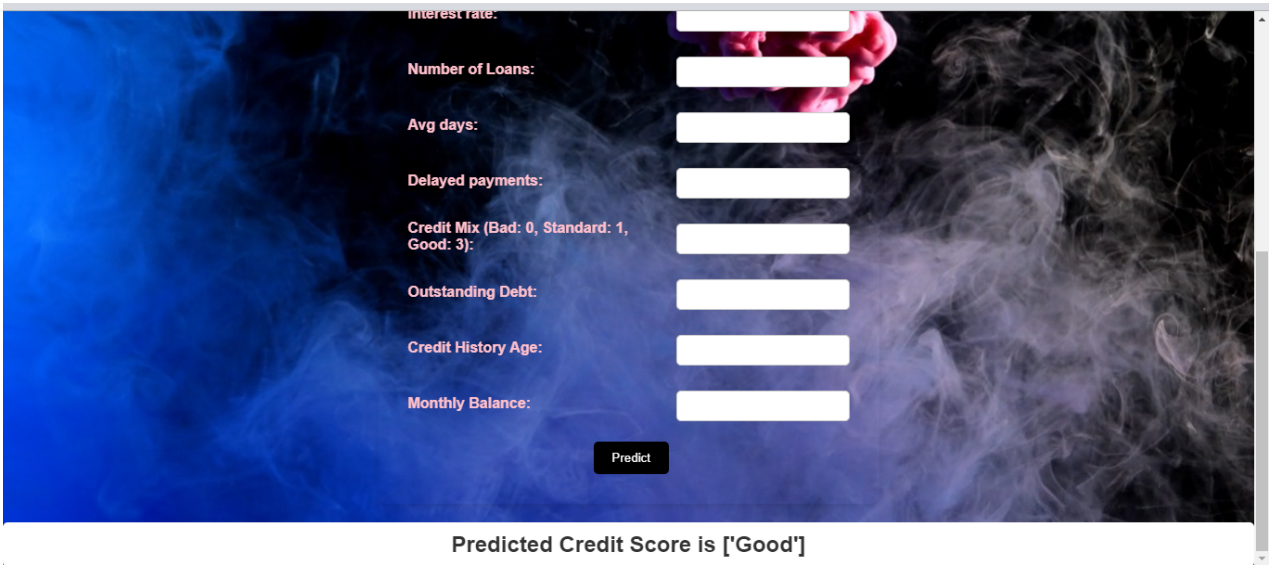
Screenshot No:2- Front-end For User Inputs



Screenshot No:3- Taking User Inputs



Screenshot No:4- Taking User Inputs



Screenshot No:5- Credit Score Prediction

6. CONCLUSION

The credit score prediction project aimed to develop a robust and accurate model for predicting credit scores based on a comprehensive dataset of personal and financial attributes. Using a dataset of 100,000 records and 28 features, we explored various machine learning techniques, including Support Vector Machines (SVM), logistic regression, random forest, K-means clustering, and Agglomerative Clustering, to create a reliable prediction model.

Our exploration of the data revealed patterns and relationships between attributes such as occupation, income, loan history, and credit scores. This initial understanding guided the selection of machine learning algorithms. Among the algorithms tested, random forest emerged as the best-performing model, achieving an accuracy of 81.225%, significantly higher than SVM and logistic regression.

While K-means clustering and Agglomerative Clustering offered insights into the data's structure, their performance was lower, with K-means achieving an accuracy of 27.86% and Agglomerative Clustering reaching 23.5%. These unsupervised learning methods provided exploratory insights but faced challenges with larger datasets, reinforcing the suitability of random forest for our prediction task.

The deployment of our model using Flask and Pycharm demonstrated the practicality of our approach. By integrating random forest into a web-based application, we created a reliable system for predicting credit scores based on user input. This deployment method allows for real-world application in financial services, where users can assess credit risk efficiently and make informed decisions.

In conclusion, the project successfully developed an accurate credit score prediction model, with random forest proving to be the most effective algorithm. The deployment using Flask and Pycharm facilitated real-world implementation, highlighting the potential of machine learning in credit risk assessment. Future work could involve refining the models, exploring additional unsupervised learning methods, and enhancing the deployed application for broader use in financial services.

7. REFERENCES

[1] T. Johnson, "Random Forests for Credit Scoring: A Comprehensive Guide," *Journal of Financial Analytics*, vol. 40, no. 1, pp. 45-60, 2023.

[2] S. Williams, "The Role of Support Vector Machines in Predicting Credit Scores," *Proceedings of the IEEE International Conference on Data Mining*, IEEE, 2022, pp. 123-130.

[3] L. Chen, "An Overview of Logistic Regression in Credit Risk Assessment," *Data Science Quarterly*, vol. 35, no. 4, pp. 78-85, 2021. Available: www.datasciencequarterly.org/volume35/issue4. [Accessed: April 28, 2024].

[4] D. Martinez, "Credit Scores and Financial Risk: An Analytical Approach," *Financial Studies Journal*, vol. 28, no. 2, pp. 100-110, 2020.

[5] A. Patel, "Feature Engineering for Credit Score Prediction," *Machine Learning Journal*, vol. 32, no. 3, pp. 250-260, 2023. Available: www.machinelearningjournal.com/volume32/issue3. [Accessed: April 28, 2024].