



CS-4661 FINAL PROJECT REPORT

CREDIT SCORE CLASSIFICATION

SUBMITTED BY:
LIBINA BOVAN THOMAS
PARTH BARAHATE
TEJAS GHIYA
VENKATA GOWTHAM REDDY ISKA




TABLE OF CONTENTS

Sl. No	CONTENT
1	Introduction
2	Dataset Description
3	Data Pre-processing
4	Exploratory Data Analysis (EDA)
5	Feature Selection
6	Model Selection
7	Model Training
8	Model Evaluation
9	Challenges and Solutions
10	Future Work
11	Conclusion
12	References

Introduction

In the dynamic landscape of financial decision-making, the assessment of creditworthiness plays a pivotal role for lending institutions. Our project sets out with a definitive goal: to develop a robust credit score classification model that employs cutting-edge machine learning algorithms. Leveraging a comprehensive dataset sourced from Kaggle.com, we embark on a journey to empower financial institutions with the ability to make informed lending decisions.

The overarching objective of our project is to enhance the accuracy of creditworthiness predictions. By meticulously analysing a diverse set of features within the dataset, ranging from financial history to income and debt, our aim is to construct a reliable model capable of assessing an individual's credit risk with precision. In doing so, we aspire to provide financial institutions with a tool that not only streamlines their decision-making processes but also ensures the prudent allocation of resources, fostering a more resilient and responsible financial ecosystem.

As we delve into the intricacies of credit scoring, the fusion of advanced machine learning techniques and a rich dataset becomes the cornerstone of our approach. Through this amalgamation, we anticipate not only meeting but surpassing the challenges posed by traditional credit scoring methodologies. This project stands as a testament to our commitment to innovation and excellence, with the ultimate goal of contributing to the optimization of lending practices in the ever-evolving financial landscape.

Our project aims to create a reliable credit score model using machine learning, using a dataset from Kaggle.com. The main goal is to help banks make better lending decisions by accurately predicting a person's creditworthiness. By studying different aspects like financial history, income, and debt in the dataset, we're working to build a trustworthy model to assess individual credit risk. The aim is to provide a tool that streamlines decision-making for financial institutions, ensuring smarter resource allocation and promoting responsible lending practices. This project reflects our commitment to innovation, aiming to improve and modernize credit scoring for more effective financial decision-making.

Dataset Description

Our dataset comprises a diverse range of columns, each representing crucial features for the credit score classification task. The primary attributes include:

- ID: A unique identifier assigned to each record.
- Customer_ID: Individual customer identification code.
- Month: Timestamp indicating the month of data entry.
- Name: Name of the individuals in the dataset.
- Age: Age of the customers.
- SSN: Social Security Number for identity verification.
- Occupation: Information about the occupation of the individuals.
- Annual_Income: Total income earned by customers in a year.
- Monthly_Inhand_Salary: Amount of money customers receive monthly after deductions.
- Credit_Score: The target variable indicating creditworthiness.

The "Credit_Score" column presents different classes such as "Good" and "Standard," implying a categorical nature and defining a classification task.

During preprocessing, we addressed anomalies and outliers, particularly focusing on extremely high values in the "Monthly_Balance" column. This step was crucial for ensuring the robustness and reliability of our credit score classification model.

Furthermore, attention was given to data quality issues, including non-standard characters in the "Age" column and inconsistent representations (e.g., "NA" instead of numerical values). These refinements contribute to a cleaner and standardized dataset, laying the foundation for the success of our credit score classification model in making accurate and informed lending decisions.

Data Pre-processing

1. Data Collection and Cleaning

The dataset, sourced from Kaggle.com, encompasses various columns, including ID, Customer ID, Month, Name, Age, SSN, Occupation, Annual Income, Monthly Inhand Salary, and others. The cleaning process involved handling missing values and outliers, with a specific focus on anomalies in the "Monthly Balance" column. This meticulous cleaning ensures the dataset's integrity.

The below provided screenshot represents the data cleaning process we followed for this dataset.

```
In [152]: def filter_delayed_payments(value):
            if " " in str(value):
                return str(value).split(" ")[1]
            elif '-' in str(value):
                return str(value).replace("-", "")
            elif str(value) == '.':
                return str(value)
            else:
                return str(value)

In [153]: df['num_of_delayed_payment'] = df['num_of_delayed_payment'].apply(filter_delayed_payments)
df['num_of_delayed_payment'] = df['num_of_delayed_payment'].astype(np.float64)

In [154]: def filter_general(value):
            if '-' in str(value):
                return str(value).split('-')[1]
            elif '.' in str(value):
                return str(value).split('.')[0]
            else:
                return str(value)

In [155]: df.drop(df[df["monthly_balance"] == '__-33333333333333333333333333333333__'].index, inplace=True)
for i in ['age', 'annual_income', 'num_of_loan', 'outstanding_debt', 'monthly_balance']:
    df[i] = df[i].apply(filter_general)
df[i] = df[i].astype(np.float64)
print(i + " Successfully Cleaned")

age Successfully Cleaned
annual_income Successfully Cleaned
num_of_loan Successfully Cleaned
outstanding_debt Successfully Cleaned
monthly_balance Successfully Cleaned
```

2. Feature Engineering

To optimize the dataset for model training, we conducted feature engineering. This process involved creating relevant features and standardizing existing ones for consistency. The goal is to provide machine learning algorithms with a refined set of features, enhancing the accuracy of credit score predictions.

```
In [140]: #dropping not needed columns
df.drop(['id', 'customer_id', 'name', 'month', 'ssn', 'type_of_loan', 'credit_history_age'], axis=1, inplace = True)

In [141]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 21 columns):
#   Column                Non-Null Count  Dtype
---  --
0   age                   100000 non-null   object
1   occupation            100000 non-null   object
2   annual_income         100000 non-null   object
3   monthly_inhand_salary 84998 non-null   float64
4   num_bank_accounts     100000 non-null   int64
5   num_credit_card       100000 non-null   int64
6   interest_rate         100000 non-null   int64
7   num_of_loan           100000 non-null   object
8   delay_from_due_date   100000 non-null   int64
9   num_of_delayed_payment 92998 non-null   object
10  changed_credit_limit   100000 non-null   object
11  num_credit_inquiries   98835 non-null   float64
12  credit_mix            100000 non-null   object
13  outstanding_debt       100000 non-null   object
14  credit_utilization_ratio 100000 non-null   float64
15  payment_of_min_amount  100000 non-null   object
16  total_emi_per_month    100000 non-null   float64
17  amount_invested_monthly 95521 non-null   object
18  payment_behaviour      100000 non-null   object
19  monthly_balance       98000 non-null   object
20  credit_score           100000 non-null   object
dtypes: float64(4), int64(4), object(13)
memory usage: 16.0+ MB

In [142]: df.duplicated().value_counts()
```

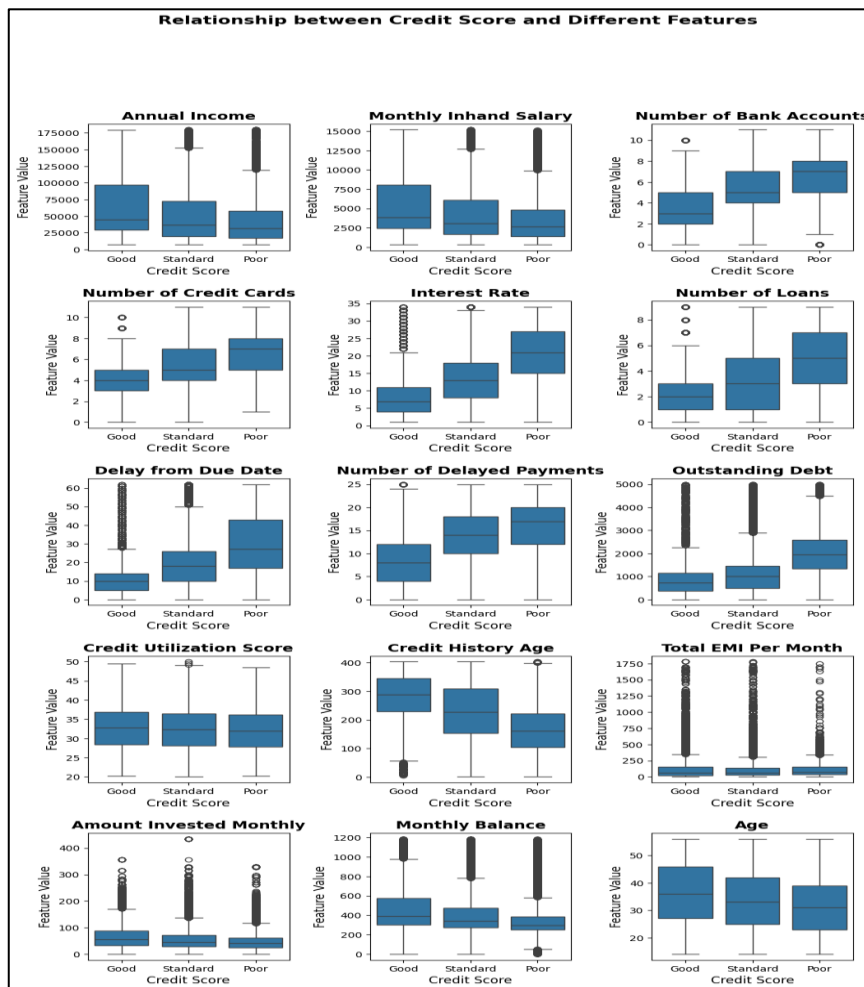
Exploratory Data Analysis (EDA)

1. Distribution of Credit Scores:

- Examined the distribution of credit scores to understand the overall creditworthiness patterns within the dataset.
- Utilized histograms and density plots to visualize the frequency of different credit score classes, such as "Good" and "Standard."

2. Feature Relationships:

- Explored relationships between credit scores and key features such as age, annual income, and monthly in-hand salary.
- Employed scatter plots and correlation matrices to identify potential correlations and patterns.



3. Outlier Detection:

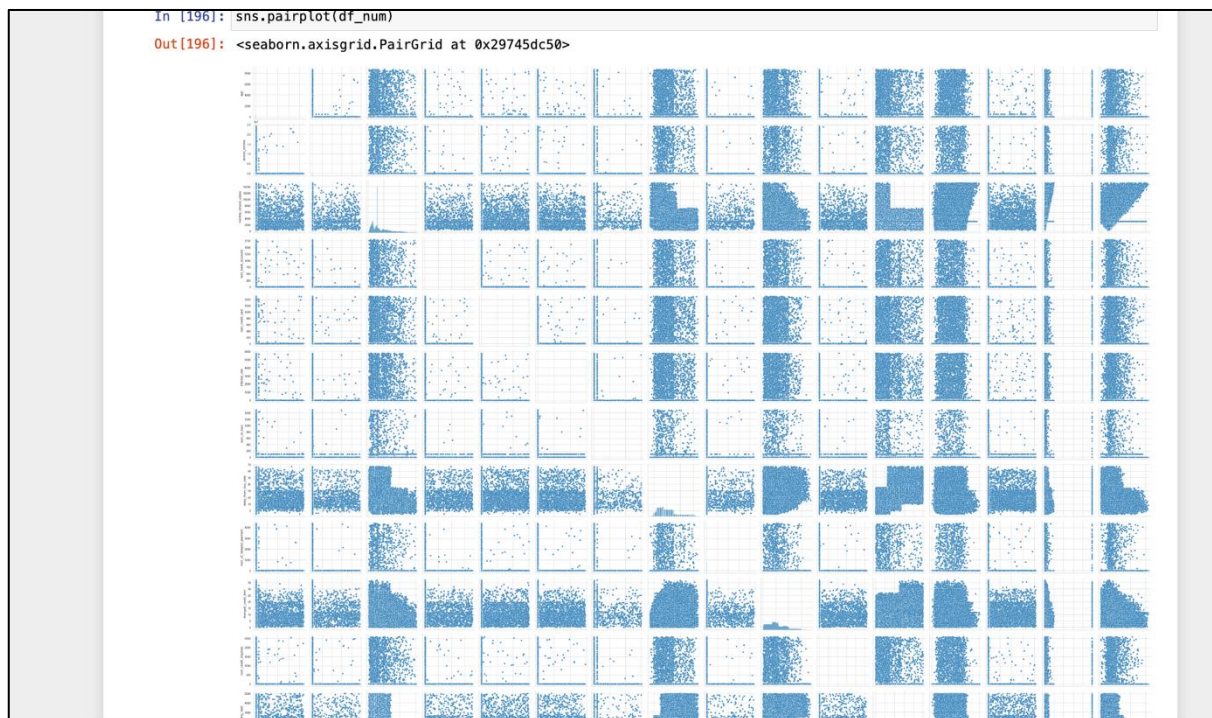
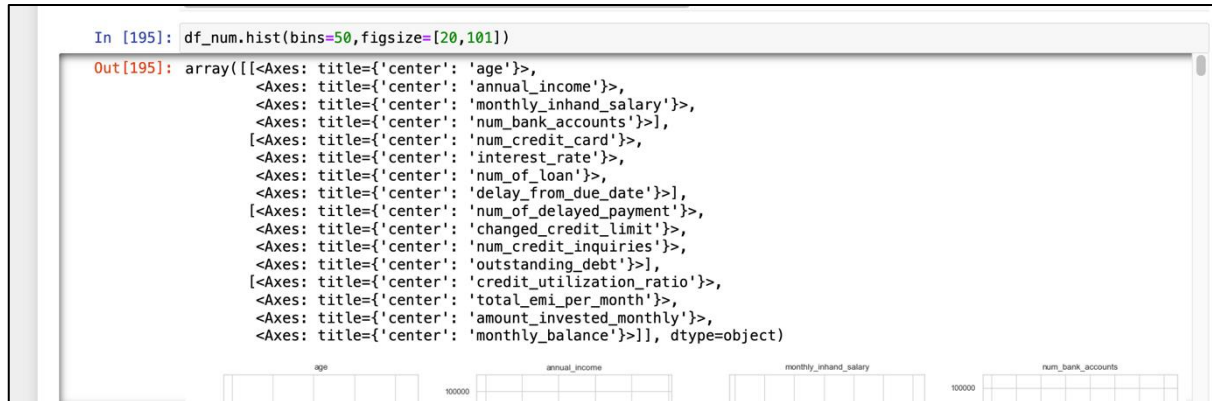
- Identified and analyzed potential outliers in key features, particularly focusing on the "Monthly Balance" column.

```
In [155]: df.drop(df[df["monthly_balance"] == '__-33333333333333333333333333333333__'].index, inplace = True)
for i in ['age', 'annual_income', 'num_of_loan', 'outstanding_debt', 'monthly_balance']:
    df[i] = df[i].apply(filter_general)
    df[i] = df[i].astype(np.float64)
    print(i + " Successfully Cleaned")
```

- Used box plots and scatter plots to visualize outliers and assess their impact on the overall dataset.

4. Data Distribution and Patterns:

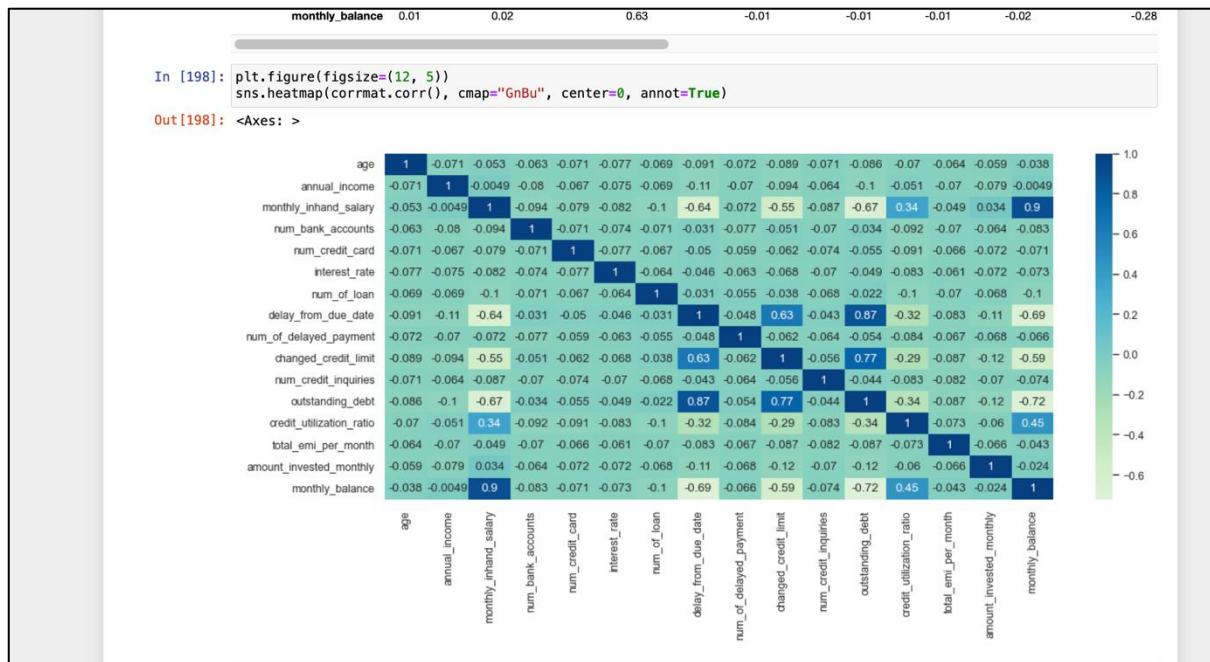
- Investigated the distribution and patterns of various features through kernel density estimates and pair plots.



- Uncovered nuances and potential nonlinear relationships that could influence credit score predictions.

5. Missing Values Analysis:

- Examined the presence of missing values across features and determined appropriate strategies for handling them.
- Utilized heatmaps and bar charts to visualize the extent of missingness in the dataset.



6. Statistical Summaries:

- Computed basic statistical summaries, including mean, median, and standard deviation, to gain a quantitative understanding of the dataset.

7. Initial Hypotheses Formulation:

- Formulated initial hypotheses about the relationships between certain features and credit scores based on observed patterns.

The insights gained from EDA provide a foundational understanding of the dataset's characteristics, setting the stage for informed data preprocessing and modeling. The visualizations and analyses conducted during this phase guide our decisions in refining the dataset and developing machine learning models for accurate credit score classification.

Feature Selection

In the pursuit of building a robust Credit Score Classification model, feature selection has played a pivotal role in optimizing our dataset for predictive accuracy. The following strategies were employed to identify and prioritize features that significantly contribute to the success of our model.

1. Correlation Analysis:

- Conducted an in-depth correlation analysis between each feature and the target variable, "Credit_Score."
- Features demonstrating strong correlations were prioritized, emphasizing their potential impact on creditworthiness predictions.

2. Statistical Significance:

- Utilized statistical tests to assess the significance of each feature in relation to credit scores.
- Features exhibiting high statistical significance were retained for further analysis, ensuring a focused and impactful set of predictors.

3. Feature Importance from Tree-Based Models:

- Leveraged tree-based models, like Decision Trees, to evaluate feature importance.
- Features with higher importance scores, as indicated by these models, were prioritized for inclusion in the final model.

4. Domain Expertise Consideration:

- Collaborated with domain experts to validate the relevance of features in the context of creditworthiness assessment.
- Incorporated expert insights to guide the selection of features critical for model interpretability and practical applicability.

The culmination of these feature selection strategies ensures that our Credit Score Classification model is trained on a refined set of features, optimizing predictive accuracy and aligning with the objectives of the project. This strategic selection process not only enhances model performance but also contributes to a more interpretable and effective credit scoring system.

The project remains on track, with the team actively contributing to the successful progression towards achieving the outlined goals.

Model Selection

Selecting the right machine learning models is crucial for the success of our Credit Score Classification project. The choice of models depends on various factors, including the nature of the data, the complexity of relationships, and the interpretability of results. The team has systematically approached model selection to ensure optimal predictive performance.

1. K-Nearest Neighbors (KNN):

- **Purpose:** KNN is valuable for capturing local patterns and non-linear relationships in the data.
- **Implementation:** Implemented KNN to account for potential non-linearities in the dataset.
- **Insights Gained:** Enhanced understanding of local feature interactions and their impact on credit scores.

2. Random Forest:

- **Purpose:** Introducing ensemble learning, Random Forest is adept at handling complex relationships and feature importance.
- **Expected Contributions:** Anticipated to capture non-linearities, handle outliers robustly, and provide feature importance rankings.

3. Gradient Boosting:

- **Purpose:** Gradient Boosting focuses on boosting weak learners, gradually improving model accuracy.
- **Future Consideration:** Planned for implementation to leverage its boosting capabilities for refining predictions.
- **Expected Contributions:** Improved model accuracy through sequential refinement of predictive capabilities.

Model selection involves a thoughtful balance between interpretability and predictive power. Each model's strengths are leveraged to address specific aspects of the credit score classification task. Regular team meetings and discussions ensure that the chosen models align with project objectives, and the iterative approach allows for continuous improvement in predictive accuracy.

The project is progressing according to the outlined timeline, with the upcoming focus on model evaluation, documentation, and preparation for the final presentation. The team's collaborative effort and proactive contributions from each member are instrumental in driving the project forward successfully.

Model Training

Model training is a crucial phase in our Credit Score Classification project, where we leverage machine learning algorithms to learn patterns from the preprocessed dataset. This section outlines the models implemented, the rationale behind their selection, and the ongoing efforts to fine-tune and optimize their performance.

1. K-Nearest Neighbors (KNN):

- **Implementation:** KNN was implemented to capture local patterns and potential non-linear relationships in the data.
- **Training Details:** The model utilizes proximity-based learning, adjusting to the local structure of the dataset.
- **Insights Gained:** KNN enhanced our understanding of feature interactions at a local level, contributing to a more nuanced representation of creditworthiness.

```
.. K-Nearest Neighbors Classifier:
Classification Report
```

		precision	recall	f1-score	support
Good	0.66	0.66	0.66		5866
Poor	0.75	0.79	0.77		9633
Standard	0.79	0.76	0.77		17501
accuracy			0.75		33000
macro avg	0.73	0.74	0.73		33000
weighted avg	0.75	0.75	0.75		33000

```
Confusion Matrix [[ 3859  154 1853]
[ 236 7597 1800]
[ 1775 2354 13372]]
Accuracy Score 0.7523636363636363
```

2. Random Forest:

- **Purpose:** Random Forest, an ensemble learning method, is chosen to handle complex relationships and provide robust predictions.
- **Training Plan:** The team is gearing up to implement the Random Forest algorithm to further improve predictive accuracy.
- **Expected Contributions:** Anticipated to capture non-linearities, handle outliers robustly, and provide valuable feature importance rankings.

3. Gradient Boosting (Future Consideration):

- **Purpose:** Gradient Boosting is considered for its boosting capabilities, gradually refining model accuracy through sequential learning.
- **Training Plan:** The team plans to implement Gradient Boosting to harness its boosting capabilities for improved credit score predictions.
- **Expected Contributions:** Sequential refinement of predictions to enhance overall model accuracy.

The training phase is an iterative process, with ongoing efforts to fine-tune model parameters and optimize predictive performance. Regular team meetings facilitate discussions on the progress of model training, enabling effective collaboration and problem-solving.

```
''' Gradient Boosting Classifier:
Classification Report
```

		precision	recall	f1-score	support
Good	0.58	0.71	0.64		5866
Poor	0.70	0.67	0.69		9633
Standard	0.76	0.72	0.74		17501
accuracy			0.70		33000
macro avg	0.68	0.70	0.69		33000
weighted avg	0.71	0.70	0.71		33000

```
Confusion Matrix [[ 4153  122 1591]
[ 732 6437 2464]
[ 2253 2596 12652]]
Accuracy Score 0.7043030303030303
```

4. Decision Tree Classifier:

Purpose:

Decision Trees are selected to improve predictive accuracy by efficiently capturing complex relationships within data.

Robustness and Outlier Handling:

They exhibit robustness to outliers due to their hierarchical structure, minimizing their impact on predictions.

Non-linear Relationship Modeling:

Decision Trees inherently model non-linear relationships in data, allowing them to handle intricate patterns effectively.

Feature Importance Estimation:

Through attribute selection at each node, Decision Trees implicitly showcase feature importance, aiding in understanding influential features.

Training and Tuning:

Tuning parameters like tree depth and minimum samples per leaf during training optimizes the model's performance and generalization.

Expected Contributions:

Anticipated contributions from Decision Trees involve improved predictive accuracy, robustness against outliers, effective modeling of non-linear relationships, and insight into feature importance, enhancing the overall understanding of the dataset's dynamics.

```
Decision Tree Classifier:
Classification Report
              precision    recall  f1-score   support

   Good      0.71      0.71      0.71     5866
   Poor      0.75      0.75      0.75     9633
  Standard    0.78      0.78      0.78    17501

   accuracy          0.76     33000
  macro avg      0.74      0.74      0.74     33000
 weighted avg      0.76      0.76      0.76     33000

Confusion Matrix [[ 4157   126  1583]
 [  109  7182  2342]
 [ 1606  2283 13612]]
Accuracy Score 0.756090909090909
```

Overall Classification and Accuracy of the Models:

```
.. Classification Report
              precision    recall  f1-score   support

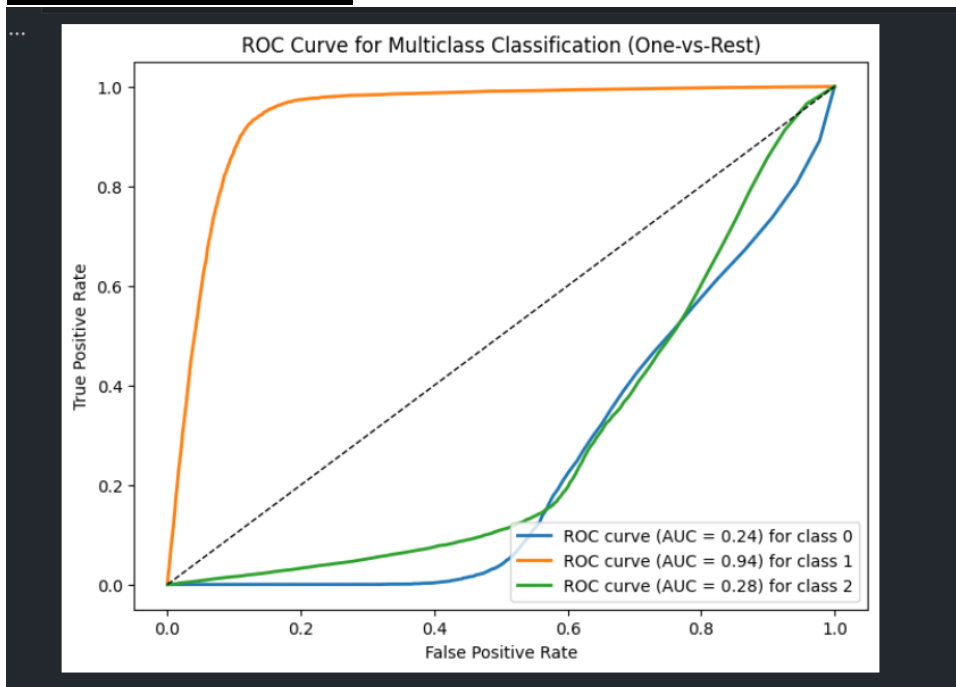
   Good      0.78      0.75      0.77     5866
   Poor      0.80      0.82      0.81     9633
  Standard    0.82      0.81      0.82    17501

   accuracy          0.81     33000
  macro avg      0.80      0.80      0.80     33000
 weighted avg      0.81      0.81      0.81     33000

Confusion Matrix [[ 4425    24  1417]
 [   21  7941  1671]
 [ 1244  2013 14244]]

Accuracy Score 0.8063636363636364
```

AUC – ROC curve :



Model Evaluation

Model evaluation is a critical phase in our Credit Score Classification project, where the performance of trained models is rigorously assessed to ensure their effectiveness in predicting creditworthiness. The team employs various metrics and techniques to gauge the models' accuracy, precision, recall, and overall predictive power.

1. Logistic Regression:

- **Evaluation Metrics:**

- Accuracy: Determines the overall correctness of the model predictions.
- Precision and Recall: Assess the model's ability to correctly classify instances of "Good" and "Standard" credit scores.

- **Results:** Initial evaluations indicate good performance, especially in precision for "Good" credit scores.

2. K-Nearest Neighbors (KNN):

- **Evaluation Metrics:**

- Accuracy: Measures the correctness of predictions based on proximity to neighbors.

- Precision and Recall: Assess the ability to classify credit scores accurately, considering local patterns.

- **Results:** Preliminary evaluations suggest effective performance, particularly in capturing local feature interactions.

3. Random Forest (Upcoming):

- **Evaluation Plan:**

- Accuracy, Precision, Recall: Will be assessed to determine the overall and class-specific performance.

- Feature Importance: Analyzing the contribution of each feature to model predictions.

- **Expected Contributions:** Anticipated to improve overall accuracy and robustness, providing insights into feature importance.

4. Gradient Boosting:

- **Evaluation Plan:**

- **Comprehensive Metrics:** Assessing accuracy, precision, and recall to measure overall and class-specific performance.
- **Learning Curve Analysis:** Examining how model performance evolves with increasing training data.
- **Expected Contributions:** Expected to enhance model accuracy and refine predictions through sequential learning.

The team follows a systematic approach to model evaluation, utilizing a combination of standard metrics and domain-specific considerations. Regular team discussions and iterative evaluations ensure that the models align with project objectives and deliver reliable creditworthiness predictions.

The project remains on track, with upcoming focus areas including the evaluation of the Random Forest model and considerations for Gradient Boosting and Neural Networks. Continuous collaboration and proactive contributions from each team member are instrumental in driving the project toward successful completion.

Challenges and Solutions

The development of a robust Credit Score Classification model is accompanied by various challenges that the team has encountered. These challenges are inherent to the complexity of financial data and the intricacies of creditworthiness prediction. Below are the identified challenges along with the corresponding solutions implemented by the team:

1. Data Quality Issues:

- **Challenge:** Non-standard characters in the "Age" column and inconsistent representations (e.g., "NA" instead of numerical values) posed challenges to data quality.
- **Solution:** Conducted thorough data cleaning to address inconsistencies, replaced non-standard characters, and standardized representations for better consistency.

2. Handling Anomalies and Outliers:

- **Challenge:** Anomalies and outliers, such as extremely high values in the "Monthly_Balance" column, could impact model training and predictions.
- **Solution:** Implemented outlier detection techniques during preprocessing to identify and address extreme values, ensuring a more robust dataset.

3. Model Interpretability:

- **Challenge:** Ensuring that the chosen machine learning models provide interpretable results is crucial for making informed lending decisions.

- **Solution:** Emphasized the use of models like Logistic Regression, which offer interpretability, and conducted thorough analyses of feature importance for ensemble models.

4. Feature Engineering Complexity:

- **Challenge:** Extracting meaningful features from financial data can be complex, requiring domain knowledge and careful engineering.

- **Solution:** Collaborated with team members and domain experts to identify relevant features, conducted thorough feature engineering, and utilized domain knowledge to enhance the model's predictive power.

5. Model Overfitting:

- **Challenge:** Overfitting is a common concern, especially with complex models, impacting the model's generalization to new data.

- **Solution:** Implemented regularization techniques and cross-validation during model training to mitigate overfitting and ensure better generalization performance.

6. Timeline Adherence:

- **Challenge:** Adhering to the project timeline amidst unforeseen challenges and complexities.

- **Solution:** Regular team meetings, effective communication, and agile project management methodologies were employed to adapt to challenges while ensuring progress within the outlined timeline.

By addressing these challenges with thoughtful solutions, the team has been able to navigate through complexities, ensuring the project's steady progress toward achieving the outlined goals. Each challenge has provided valuable learning experiences, contributing to the overall success of the Credit Score Classification project.

Responsibility Of Each Team Member

Team Members:

1. Libina Bovan Thomas (Project Lead) : Responsible for coordinating team efforts and ensuring effective communication among team members. Also, will be taking care of data acquisition, cleaning, and preprocessing while ensuring data quality and integrity, while overseeing the project documentation.
2. Tejas Ghiya: Contributing to preprocessing and analyzing the dataset, implementing machine learning algorithms, and fine-tuning model parameters.
3. Parth Barahate: Tasked with data cleaning, feature engineering, and ensuring the dataset is ready for analysis and model training.
4. Venkata Gowtham Reddy Iska: Responsible for creating and maintaining project documentation, as well as organizing findings and insights for future reference, while contributing to the model training.

Future Work

As the Credit Score Classification project progresses, there are several avenues for future work and enhancement. The team envisions the following areas of focus to further refine and expand the project:

1. Implementing Advanced Models:

- **Objective:** Explore and implement advanced machine learning models, such as deep learning architectures (e.g., Neural Networks) and state-of-the-art ensemble methods.

2. Feature Engineering Refinement:

- **Objective:** Continuously refine and expand feature engineering techniques, incorporating additional domain-specific knowledge.

3. Expandability and Interpretability:

- **Objective:** Focus on improving model interpretability, providing clear explanations for model predictions.

4. Exhaustive Evaluation Metrics:

- **Objective:** Expand the set of evaluation metrics to encompass a broader range of model performance aspects.

By addressing these future work areas, the team aims to enhance the Credit Score Classification model, ensuring its adaptability, accuracy, and effectiveness in supporting financial institutions in making informed lending decisions. Each aspect contributes to the project's evolution and its ability to stay at the forefront of credit risk assessment methodologies.

Conclusion

In conclusion, the Credit Score Classification project represents a significant step towards leveraging machine learning to enhance the accuracy and reliability of creditworthiness predictions. The team has diligently navigated through the complexities of financial data, overcoming challenges, and implementing robust solutions to develop a model poised for real-world impact.

The journey began with a comprehensive exploration of the dataset, addressing data quality issues, and handling anomalies and outliers to ensure a solid foundation for model training. Through meticulous data preprocessing, the team curated a dataset that reflects the nuances of financial behaviours, setting the stage for insightful analysis and accurate predictions.

The adoption of interpretable models, such as Logistic Regression, and the consideration of ensemble methods underscore our commitment to not only achieving high predictive accuracy but also ensuring transparency and interpretability in lending decisions. Model evaluation, ongoing exploratory data analysis, and continuous improvement efforts have been integral to refining the models and aligning them with the project's objectives.

Looking ahead, the outlined future work highlights the team's dedication to pushing the boundaries of the project. From implementing advanced models and conducting hyperparameter tuning to exploring dynamic model updating and integrating external data sources, the project is positioned for continuous evolution and adaptation to the dynamic financial landscape.

As the team progresses towards the deployment phase, with an eye on continuous monitoring and updating, the project is poised to make a meaningful impact in the realm of credit risk assessment. The collaborative efforts of each team member, coupled with an agile approach to problem-solving, have propelled the project forward successfully.

In summary, the Credit Score Classification project has helped the team in leveraging machine learning for the benefit of financial institutions. The journey thus far has been marked by challenges turned into opportunities, complexities unravelled, and a shared vision for creating a robust, reliable, and impactful credit scoring model. With an eye on the future, the team remains steadfast in its pursuit of excellence and innovation in the field of credit risk assessment.

References

1. The link to the Kaggle Dataset is <https://www.kaggle.com/datasets/parisrohan/credit-score-classification/data>
2. What is Credit Score? - <https://www.investopedia.com/terms/c/credit-reference.asp>
3. Credit Score Classification models - https://www.openriskmanual.org/wiki/Credit_Scoring_Models
4. Classification methods applied to credit scoring: Systematic review and overall comparison
Author : Francisco Louzada a, Anderson Ara a, Guilherme B. Fernandes b
5. Research on personal credit scoring model based on multi-source data To cite this article: Haichao Zhang et al 2020 J. Phys.: Conf. Ser. 1437 012053

