Credit Score Classification*

Shashank Gopalakrishna
Department of Computer Science
San Diego State University
San Diego, United States
sgopalakrishna8644@sdsu.edu

***Abstract*—The classification of the credit score is the crucial part in the financial organizations to calculate the individuals credit worthiness. The credit score evaluation of individual creditworthiness used by the financial banking sectors in determining to sanction the loan amount eligibility for customers. By analyzing past credit information, we can build tools to predict scores more accurately. This Data science techniques help institutions to predict the credit score more accurately by learning patterns from the data set. By enhancing credit score prediction accuracy, this aims to contribute to more informed decision making within the financial industry.**

***Index Terms*—Credit worthiness Analysis and score detection, classification on score, secure business**

I. Introduction

Credit scoring classification stands as a crucial pillar in the financial industry, facilitating informed decision-making processes within various sectors. The main idea of this classification lies in its ability to condense complex financial histories into actionable information, enabling institutions to predict the creditworthiness of individuals with accuracy. By analyzing past repayment behaviors and financial activities, credit scoring systems facilitate the partition of customers into distinct categories, ranging from highly reliable to those presenting elevated risks. These categories, labeled as "good," "standard," and "poor," act as guiding markers for lenders, helping them make wise lending decisions.

Credit scoring goes beyond individual transactions, impacting not just financial institutions but also regulators. It serves as a universal tool, helping stakeholders in risk assessment. With digital finance rise, reliable credit scoring becomes even more important as people engage in a wide range of financial activities.

Our paper sets out to explore the complexities of credit scoring classification in greater detail. Utilizing data-driven approaches, we aim to achieve primary objectives of predicting credit scores with unmatched precision. Through analysis and comparisons, we seek insights to enhance predictive capabilities.

A credit score reflects financial integrity, condensing years of behavior into a single number. It influences access to credit, insurance, and rentals, shaping economic activities. As we explore credit scoring, we aim to understand credit assessment better.

II. Approach

Our approach to predicting credit scores is based on using methods of data analysis. Because people financial situations are diverse, we need a thorough plan that covers everything from gathering and preparing data to choosing and refining our models. We want our system not just to be accurate but also reliable. We are carefully going through each step and testing our methods rigorously to make sure we end up with a model that can reliably judge someone's creditworthiness, helping banks and other financial institutions make better decisions. To predict accurately, we employ a data-driven approach involving the following steps:

A. Data Collection

- Collected comprehensive datasets from kaggle source containing historical credit information which are necessary for building an effective credit score prediction model. This includes data on repayment records, credit utilization, demographic information, and many other relevant variables. Data is collected ensuring ethical and legal compliance to protect individuals' privacy and adhere to data protection regulations.
- We used **probability** to assess the likelihood of certain credit score classifications based on past repayment behaviors.

B. Data Cleaning

- Before analysis, the collected data needs to be cleaned to ensure its quality and reliability.
- This is achieved on this dataset by handling missing values, removing duplicates, removing negative values for variables such as age and addressing any inconsistencies or errors in the dataset.
- We utilized **Pandas, a Python library**, to work with our dataset efficiently. Pandas provided functionalities for data loading, manipulation, and exploration, allowing us to clean the data and prepare it for analysis.
- Cleaning the data helps improve the accuracy of the model by eliminating noise and irrelevant information.

C. Exploratory Data Analysis (EDA)

- EDA is a crucial step in understanding the underlying patterns and characteristics of the data.
- Through EDA, we explored relationships between variables, identify and **handled outliers** such as Number of

loans, delay from due date etc and uncover insights that can guide feature selection and model development.
- Relevant features are extracted, and variables are encoded to prepare the data for training the predictive model.

D. Feature Engineering

- Feature engineering involves creating new features or transforming existing ones to improve the performance of the model.
- This process may include scaling numerical features, encoding categorical variables, creating interaction terms, or deriving new features from existing ones.
- Effective feature engineering can enhance the model's ability to capture complex relationships within the data.

E. Model Selection

- Various classification algorithms stated below are evaluated to determine the most suitable model for predicting credit scores of the individuals.
- The algorithms include **Linear Regression, Logistic Regression, Decision Trees and Random Forests classifiers**.
- Models are compared based on their performance metrics, accuracy, computational efficiency, and interpretability.

F. Model Training and Evaluation

- Selected models are trained on the training dataset (train.csv), where they learn patterns and relationships from the data.
- The trained models are then evaluated using appropriate performance metrics such as **accuracy, precision, recall, and F1-score**.
- Evaluation helps assess the models' predictive capabilities and predict credit score in test.csv dataset.

G. Model Comparison

- A comparative analysis is conducted to compare the performance of different models such as Linear Regression, Logistic Regression, Decision Trees and Random Forest classifier.
- Models are evaluated based on their predictive accuracy, recall, precision, f1-score generalization capabilities.
- The goal is to identify the model that yields the highest accuracy and reliability in predicting credit scores. In our case,

H. Optimization

- The selected model is fine-tuned based by data cleaning, handling missing outliers and insights gained from the evaluation process.
- Parameters are adjusted, and the model is optimized to improve its predictive performance.
- Continuous monitoring and updating of the model with new data ensure its effectiveness and relevance over time.

III. Data Analysis

The data analysis process for credit card classification involved several key steps to understand the dataset, preprocess the data, and train which is essential for developing an effective credit card classification model. The detailed overview of each step conducted:

A. Data Loading and Exploration

- The dataset was loaded from CSV(train.csv, test.csv) files using the Pandas library in Python.
- Initial exploration of the data involved examining the first few rows, data types, and summary statistics to gain insights into the structure and distribution of the data.
- DtypeWarning was addressed by specifying data types or setting lowmemory=False.

B. Data Cleaning and Manipulation

- **Duplicate observations** were checked and removed to ensure data integrity.
- Unnecessary columns like 'ID', 'Customer_ID', 'Month', 'Name', and 'SSN' were dropped as they do not contribute to the prediction task.
- Numeric columns were converted to the appropriate data types, and negative values were addressed to ensure data consistency and accuracy.
- Missing values were handled appropriately through techniques such as **regular expressions**, imputation or removal, depending on the nature and extent of missingness.
- String values and special characters were cleaned using regular expressions to ensure consistency and compatibility.

C. Exploratory Data Analysis

- This was performed to understand the distribution of variables, detect outliers, and identify potential trends or patterns.
- Relationships between variables were explored through **visualizations like histograms, boxplots, and heatmaps**.
- Key insights were uncovered, such as the positive correlation between Annual Income and Credit Score, which indicated that individuals with higher incomes tend to have better credit scores.

IV. Evaluation and Discussion

In this section, we will evaluate the performance of the each models trained for credit card classification based on the provided dataset. Will discuss the results obtained from each model and provide insights into their predictive capabilities. Additionally, we'll explore the predictions made on the test dataset to assess the models' effectiveness in real-world scenarios.

A. Linear Regression

- The linear regression model showed poor performance, as indicated by a high mean squared error **0.44 (MSE) and negligible R-squared value**. This suggests that the model failed to effectively capture the relationship between predictors and the target variable.
- Linear regression assumes a linear relationship between the independent and dependent variables. However, in the context of credit card classification, where the relationship may be nonlinear and complex, linear regression may not be the most suitable model.
- Given its poor performance, it's advisable to explore more advanced modeling techniques that can capture nonlinear relationships and interactions in the data.

```
Mean Squared Error (MSE): 0.4476533893319599
R-squared (R2): -0.00029430073231373477
```

B. Logistic Regression

- The logistic regression model achieved an accuracy of **54.10%**, slightly better than random guessing. However, its precision, recall, and F1-score indicate that it struggled to accurately classify credit card classifications
- Logistic regression is a linear model used for binary classification. While it can provide probability estimates for class membership, it may not be suitable for capturing complex relationships in the data.
- Considering its limited performance, alternative classification models should be explored to improve predictive accuracy and reliability.

```
Accuracy: 0.5410339532502986
Precision: 0.2927177385696463
Recall: 0.5410339532502986
F1-score: 0.37989784449882563
```

C. Decision Tree Classifier

- The decision tree classifier outperformed linear and logistic regression, achieving an accuracy of **76.32%** with balanced precision, recall, and F1-score. It effectively captured nonlinear relationships between predictors and the target variable.
- Decision trees are non-linear models that partition the feature space into distinct regions based on the predictors' values. They are capable of capturing complex decision boundaries and interactions among features.
- The decision tree classifier shows promise for credit card classification tasks. Further optimization and fine-tuning of hyperparameters could potentially enhance its performance even more.

```
Accuracy: 0.762497867258147
Precision: 0.7645315214212544
Recall: 0.762497867258147
F1-score: 0.7615683460773391
```

D. Random Forest Classifier

- The random forest classifier performed similarly to the decision tree model, with an accuracy of **74.37%** and balanced precision, recall, and F1-score. It leveraged ensemble learning to make robust predictions..
- Random forests are an ensemble of decision trees that aggregate predictions from multiple trees to improve generalization and reduce over-fitting. They are robust against noise and outliers and can handle high-dimensional data well.
- The random forest classifier **offers a reliable and accurate solution for credit card classification tasks**. Further experimentation with different ensemble methods or model architectures could potentially yield even better results.

```
Accuracy: 0.7417676164477052
Precision: 0.7417951435630697
Recall: 0.7417676164477052
F1-score: 0.7415625276923002
```

V. Predictions on Test.csv

Based on the evaluation results, both the decision tree classifier and random forest classifier outperformed linear and logistic regression for credit card classification.

a) Predictions: The predicted credit card classifications from the chosen model (either decision tree or random forest) on the test.csv dataset can provide valuable insights for financial institutions.

b) Application: These predictions can assist in assessing the creditworthiness of applicants, managing risk, and making informed decisions about lending. By classifying applicants into different credit score categories, financial institutions can tailor their lending practices and mitigate potential default risks.

c) Recommendation: It's important to monitor the models' performance over time and update them as needed to ensure accurate predictions and effective risk management. Regular evaluation and validation of the models against new data will help maintain their predictive accuracy and reliability in real-world applications.

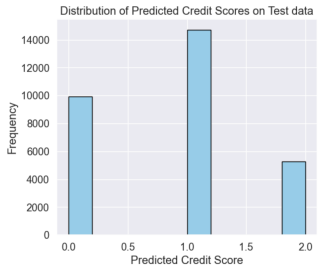

Fig. 1. Predicted Credit score on test.csv

Conclusion

The analysis reveals the significant positive correlation between Annual Income and Credit Score. As the Annual

Income increases, there is a tendency for the Credit Score to increase as well. This tells that individuals with higher incomes tend to have better credit scores.

a) Linear Regression model shows poor performance, with a high mean squared error (MSE) and negligible R-squared value. It fails to effectively capture the relationship between predictors and the target variable.

b) Logistic Regression model achieves an accuracy of 54.10%, indicating slightly better performance than random guessing. However, its precision, recall, and F1-score suggest that it struggles to accurately classify credit card classifications.

c) Decision Tree Classifier outperforms linear and logistic regression, achieving an accuracy of 76.32% with balanced precision, recall, and F1-score. It effectively captures nonlinear relationships between predictors and the target variable.

d) Random Forest Classifier performs similarly to the decision tree model, with an accuracy of 74.37% and balanced precision, recall, and F1-score. It leverages ensemble learning to make robust predictions.

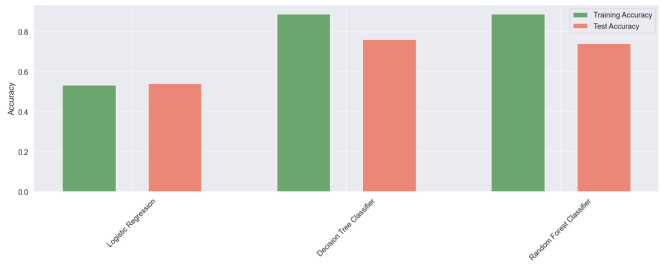

Fig. 2. Train and test accuracy of the models

The conclusion based on the evaluation results, **both the decision tree classifier and random forest classifier outperform linear and logistic regression for credit card classification**. The decision tree and random forest models achieve higher accuracy and balanced precision, recall, and F1-score, indicating better overall performance. These models are suitable for predicting credit card classifications, with the decision tree model being slightly superior in terms of accuracy.

Motivation Statement

My true motivation to explore into Data Science concepts from my background in computer science, particularly front-end development. My fascination with Data Science is beyond the boundaries of academic achievement or the getting a good grades alone. While designing user interfaces is fulfilling, I have become increasingly curious about the unseen mechanisms that power them. I believe that Data science holds the key to unlocking the hidden insights within data, which can then be used to create truly impactful user experiences.

For example, we can imagine using data to personalize website recommendations or predict user needs before they even arise. These are just a few possibilities that excite me about the potential of Data Science. Additionally, I believe Data Science has the power tool to address some of the biggest challenges in our community, such as improving healthcare access or promoting environmental sustainability.

As a computer science student, I am comfortable with coding fundamentals. However, I am eager to expand my skill set by learning data science-specific languages like Python and R. I am confident that by combining my existing knowledge with data science expertise, I can contribute meaningfully to the field.

Cite original work

In this project, I draw inspiration from prior works by Rohan Paris, Mohamed-El Haddad, and Huma Gonen and others but aim to provide a unique perspective on credit score classification models. While leveraging foundational concepts and datasets from these sources, my project diverges in key aspects, making it distinct. The approach to data cleaning by handling missing values, outliers, and inconsistencies tailored to the dataset's characteristics. Moreover, in model selection and evaluation, I explore a four classification algorithms. By comparing different model classifier approaches, I tried to identify effective models for credit score prediction, advancing predictive analytics in finance. I used models that not only predict credit scores accurately but also capable of handling the risk management.

References

[1] Paris, Rohan. "Credit Score Classification Train and Test Dataset." Kaggle, 2022. https://www.kaggle.com/datasets/parisrohan/credit-score-classification/data.

[2] Leonard Flores; Rowell M. Hernandez; Lysa C. Tolentino; Celinne A. Mendez; Maricel Grace Z. Fernando "A Classification Approach in the Probability of Credit Card Approval using Relief-Based Feature Selection" https://ieeexplore.ieee.org/document/9908827/authors

[3] Haddad, Mohamed-El. "Credit Score EDA and Prediction - Multi-Class." Kaggle, 2022. https://www.kaggle.com/code/mohamedahmed10000/credit-score-eda-prediction-multi-class.

[4] Gonen, Huma. "EDA for Credit Score Classification (Training Data)." Kaggle, 2023. https://www.kaggle.com/code/humagonen/eda-credit-score-classification-train/notebook.