**Abstract**

In today's financial landscape, mastering credit risk management is crucial for banks, prompting the need for accurate and timely credit evaluation services to stay competitive. This article introduces a holistic approach to credit scoring, leveraging historical bank details and diverse credit-related information. Using logistic regression, a tailored customer credit scoring model is developed to address transparency, biases, and inclusivity challenges prevalent in the current credit scoring system.

The introduction highlights the profound impact of credit scores on both individuals and lenders, stressing the importance of a standardized approach. The literature review delves into the historical evolution of credit scoring, delving into key components of credit scores and exploring various scoring models. The historical context of the credit scoring industry is explored, shedding light on its journey and persistent challenges.

The problem statement identifies current issues such as transparency, biases, and limited inclusivity within credit scoring systems. The proposed strategy involves a multi-faceted approach, including data analysis, integrating alternative data, promoting transparency and fairness, and collaborating with regulatory bodies.

A comprehensive data dictionary provides insights into the variables used in the analysis. The findings uncover correlations between financial indicators, offering insights into borrowing behaviors and financial strength. The model's performance is

evaluated, pinpointing areas for improvement, particularly in distinguishing between different credit categories.

Observations on payment behaviors, credit card usage, and consumption patterns provide valuable insights into consumer financial habits.

**Introduction**

Credit scores have far-reaching implications for both individuals and lenders. For individuals, a high credit score can lead to lower interest rates, better loan terms, and access to credit when needed. On the other hand, a low credit score may limit access to credit and result in higher borrowing costs. For lenders, credit scores play a crucial role in assessing the risk associated with lending, helping them make informed decisions about extending credit.

In the dynamic landscape of finance and economic transactions, credit scores have emerged as a critical determinant influencing individuals' access to financial products and opportunities. The significance of credit scores cannot be overstated, as they play a pivotal role in shaping financial outcomes, affecting decisions related to loans, credit cards, mortgages, and even employment prospects. This project endeavors to delve into the multifaceted realm of credit scores, exploring their historical context, existing challenges, and proposing recommendations for system improvement.

**Literature Review**

**Historical Development of Credit Scoring**

The concept of credit scoring traces its roots back to the early 20th century when lenders sought more systematic ways to assess an individual's creditworthiness. Initially grounded in qualitative assessments and personal relationships, the need for a standardized approach became apparent with the growth of financial markets. This led to the development of credit scoring models, marking a paradigm shift in risk assessment and lending practices.

**Components of Credit Scores**

Credit scores typically comprise several components, including payment history, credit utilization, length of credit history, types of credit, and recent credit inquiries. Each component plays a distinct role in evaluating an individual's creditworthiness, providing lenders with a comprehensive picture of their financial behavior and reliability.

**Credit Score Models**

Various credit score models are in use today, with the FICO (Fair Isaac Corporation) score being one of the most widely recognized. Other models, such as the Vantage Score, offer alternatives and enhancements, each evolving to address specific nuances in the financial landscape. Understanding the strengths and limitations of these models is essential for proposing effective system improvements.

**Historical Context of the Chosen Industry or Business**

The credit scoring industry has undergone substantial transformation over the decades, aligning with technological advancements and shifts in consumer behavior. From manual assessments to algorithmic models, the industry has adapted to handle vast datasets and complex financial interactions.

**Problem Statement**

Despite the advancements in credit scoring models, certain challenges persist. Issues of bias, lack of transparency, and limited consideration of alternative data sources hinder the effectiveness and fairness of credit scoring systems. In recognizing these challenges, this project aims to address the following problem statement:

\*The current credit scoring system lacks transparency, exhibits potential biases, and may not be fully inclusive, thereby impacting individuals' access to financial opportunities. \*

**Proposed Approach**

The proposed approach for making recommendations to improve the credit scoring system involves a multi-faceted strategy:

**1. Data Analysis and Evaluation**

Conduct a comprehensive analysis of historical credit data to identify patterns, trends, and potential biases. Evaluate the effectiveness of existing credit scoring models in predicting creditworthiness.

**2. Incorporation of Alternative Data**

Explore the incorporation of alternative data sources, such as rental payment history, utility bill payments, and other non-traditional indicators, to provide a more holistic view of an individual's financial behavior.

**3.Transparency and Explainability**

Advocate for greater transparency in credit scoring models, ensuring that individuals understand how their credit scores are calculated. Emphasize the need for explainability in algorithmic decision-making to build trust among consumers.

**4. Fairness and Inclusivity**

Address potential biases in credit scoring models by implementing measures to ensure fairness and inclusivity across diverse demographic groups. Strive for a system that minimizes disparate impact and considers the unique circumstances of individuals.

**5. Regulatory Considerations**

Collaborate with regulatory bodies to propose guidelines and standards that promote fairness, transparency, and ethical use of data in credit scoring practices.

Through these strategic steps, this project aims to contribute valuable insights and recommendations for enhancing the credit scoring system, aligning it with contemporary ethical and regulatory standards while promoting financial inclusivity and fairness.

**Data Dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| Field Name | Description | Data Type | Example |
| ID | Unique identification of an entry | Text | "001" |
| Customer\_ID | Unique identifier for a customer | Text | "CUST001" |
| Month | Month of the year | Text | "January" |
| Name | Name of the person | Text | "John Doe" |
| Age | Age of the person | Numeric | 30 |
| SSN | Social Security Number | Text | "123-45-6789" |
| Occupation | Occupation of the person | Text | "Engineer" |
| Annual\_Income | Annual income of the person | Numeric | 60000 |
| Monthly\_Inhand\_Salary | Monthly take-home salary | Numeric | 5000 |
| Num\_Bank\_Accounts | Number of bank accounts held by the person | Numeric | 2 |
| Num\_Credit\_Card | Number of credit cards held by the person | Numeric | 3 |
| Interest\_Rate | Interest rate associated with loans | Numeric | 5.5 |
| Num\_of\_Loan | Number of loans held by the person | Numeric | 1 |
| Type\_of\_Loan | Type of loan held by the person | Text | "Personal Loan" |
| Delay\_from\_due\_date | Delay in payment from the due date | Numeric | 10 |
| Num\_of\_Delayed\_Payment | Number of delayed payments | Numeric | “1,2,3. etc.” |
| Changed\_Credit\_Limit | Changes in credit limit | Numeric | 2000 |
| Num\_Credit\_Inquiries | Number of credit inquiries in the past 12 months | Numeric | “1,2,3. etc.” |
| Credit\_Mix | Types of credit accounts held by the person | Text | "Credit Card, Loan" |
| Outstanding\_Debt | Amount of outstanding debt | Numeric | 1500 |
| Credit\_Utilization\_Ratio | Credit utilization ratio | Numeric | 0.3 |
| Credit\_History\_Age | Age of the person's credit history | Numeric | 5 |
| Payment\_of\_Min\_Amount | Payment of minimum amount due | Numeric | 100 |
| Total\_EMI\_per\_month | Total monthly EMI payments | Numeric | 800 |
| Amount\_invested\_monthly | Amount invested monthly | Numeric | 300 |
| Payment\_Behaviour | Payment behavior history | Text | "Good" |
| Monthly\_Balance | Monthly account balance | Numeric | 2000 |
| Credit\_Score | Credit score of the person | Numeric | 750 |

**Observations**

A graph of blue rectangular bars with black text

Description automatically generated

1. We observed that the number of professions other than writers was almost evenly distributed, ranging from 5000 to 5500. Additionally, we separately identified the number of writers as abnormal due to the influence of random values.

A screenshot of a computer

Description automatically generated

1. Based on the heat map, we found that num\_of\_loan has a positive correlation with num\_bank\_accounts and num\_credit\_card, respectively, the conclusion is that people with insufficient funds will need to apply for loans and credit cards simultaneously to meet their funding needs. In addition, both changed\_credit\_limit and outstanding\_debt are positively correlated with num\_bank\_accounts and num\_credit\_card. The conclusion is that people who lack funds will apply to increase their credit limit to alleviate their funding needs, and this group of people is also more likely to default. The same reason is also reflected in the positive correlation between changed\_credit\_limit and num\_of\_loan/delay\_from\_due\_date, as well as between outstanding\_debt and num\_of\_loan/delay\_from\_due\_date/changed\_credit\_limit.

Amount\_invested\_monthly and monthly\_balance both show positive correlation with monthly\_inhand\_salary because high income individuals are more likely to have idle money and are more likely to use it to invest.

Therefore, it is not difficult to find that many columns that reflect "borrowing willingness", such as num\_bank\_accounts/ num\_credit\_card/ num\_of\_loan/ delay\_from\_due\_date/ changed\_credit\_limit/ outstanding\_debt, exhibit a negative correlation with column monthly\_inhand\_salary and monthly\_balance that reflects financial strength.

A graph of blue squares

Description automatically generated

1. We can find that according to the credit score, the proportion of people classified as ”Standard”, “Poor”, and “Good” decreases sequentially, with ”Standard” having the highest proportion of about 53%, while “Poor” and “Good” account for about 29% and 18% of the total, respectively.

A graph showing a number of blue squares

Description automatically generated with medium confidence

1. We have observed that over 50% of people are more willing to pay the minimum amount.

A graph of a bar

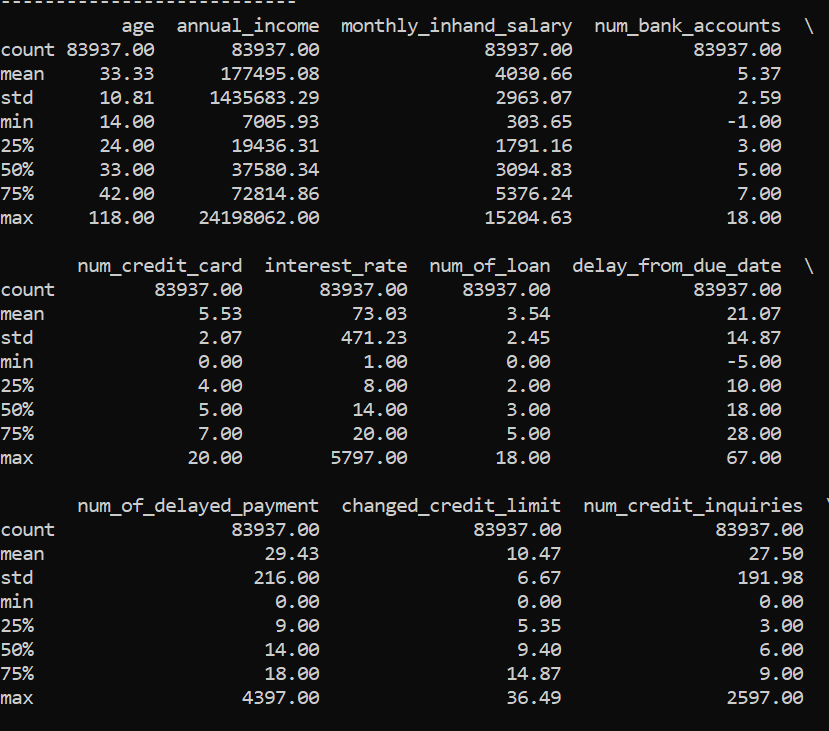
Description automatically generated with medium confidence

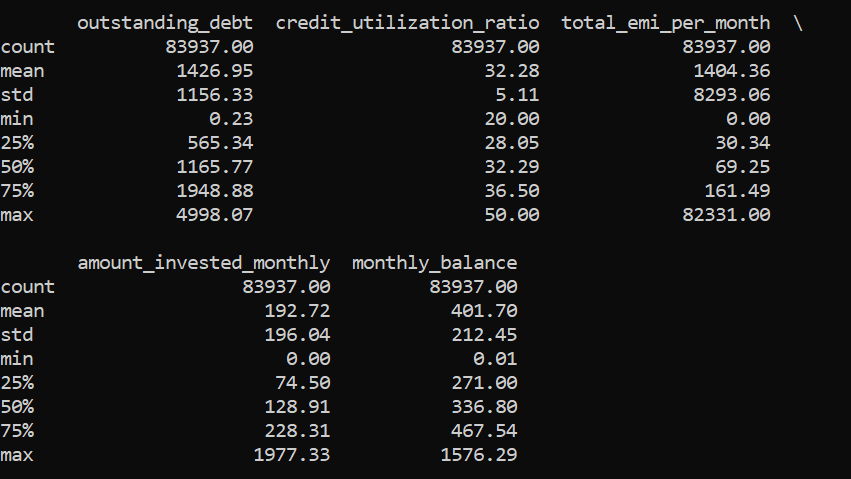
1. We have observed that most people use various types of credit cards normally, but there are still people classified as “Bad”, accounting for approximately 19% of the population, who are overly dependent on a certain type of credit card.

A graph of a bar graph

Description automatically generated

1. From the perspective of consumption quota and product value, among the population with low consumption quota, the number of people who buy low value products is the highest, approximately the sum of the number of people who buy general value products and high value products. Among the population with high consumption quotas, the number of people who buy general value products is the highest, approximately the sum of the number of people who buy high value products and low value products.





1. For the values of each column of the sample, we calculated the mean, standard deviation, maximum, minimum, and 25%/50%/75% quantile. We considered some noteworthy observations.

Firstly, there are some exceptions that need to be handled with caution, such as the maximum age of 118 years, the number of bank accounts -1, interest rates of 5797, and monthly EMI of 82331.

Secondly, we noticed that the average and median age of borrowers is 33 years old, indicating that younger people are more willing to borrow.

The average annual income of the borrower is 177495, the median is only 37580, which is likely caused by outliers, such as the maximum annual income of 24198062. This requires us to carefully judge the accuracy of the data.

The borrower holds approximately 5 bank accounts and the number of credit cards is also around 5, which is likely because the applicant has applied for one credit card at each bank.

The average interest rate is 73.03, but the median is only 14, with a large standard deviation of 471.23, which is likely caused by outliers and needs to be considered when cleaning the data.

Among borrowers who have delayed repayment, more than 75% have delayed payment more than 9 times. But most of them will pay off their debts within one month after the deadline, and their outstanding amount is within 2000.

The average monthly investment amount for most people is around $100- $200, but their account balance does not exceed $500.

A blue squares with numbers and a number

Description automatically generated

1. A summary of the confusion matrix:

**“Good” class:**

True Positives: 1689

False Negatives: 62 + 1280 = 1342 (corresponding columns besides True Positive)

False Positives: 329 + 1043=1372 (corresponding rows for True Positive)

True Negatives: 2245+1206+2210+6724= 12385 (sum of all the values except for the class we calculating values for)

**“Poor” class:**

True Positives: 2245

False Negatives: 329+2210 = 2539

False Positives: 62+1206 = 1268

True Negatives: 1689+1280+1043+6724 = 10736

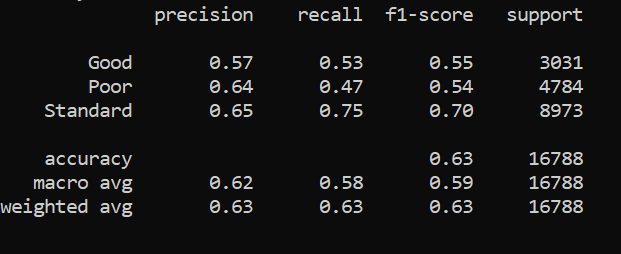
**“Standard” class:**

True Positives: 6724

False Negatives: 1043+1206 = 2249

False Positives: 2210+1280 = 3490

True Negatives: 1689+329+62+2245 = 4325



1. A summary of the classification report

**Accuracy:**

The model is correct about 63% of the time.

**Precision:**

Precision is the ratio of true positive predictions to the total number of instances predicted as positive (the sum of true positives and false positives).

For "Good," 56% of predicted "Good" values are correct.

For "Poor," 64% of predicted "Poor" values are correct.

For "Standard," 64% of predicted "Standard" values are correct.

**Recall (Sensitivity):**

Recall is the ratio of true positive predictions to the total number of actual positive instances (the sum of true positives and false negatives).

For "Good," it captures 50% of the actual "Good" values.

For "Poor," it captures 47% of the actual "Poor" values.

For "Standard," it captures 76% of the actual "Standard" values.

**F1-Score:**

It's a balance between precision and recall.

For "Good," it's 0.53.

For "Poor," it's 0.54.

For "Standard," it's 0.70.

**Conclusion**

The logistic regression model crafted for credit scoring emerges as a potent tool poised to redefine decision-making processes in the financial sector, particularly within lending institutions. Through a meticulous examination of diverse features encompassing individual demographics and financial behaviors, the model showcases a nuanced understanding of the factors influencing creditworthiness. Inclusion of features such as occupation, credit mix, and payment behavior enhances the model's predictive capacity, enabling a detailed assessment of an individual's credit risk.

A prominent strength of the model lies in its accurate prediction of credit scores. Leveraging a diverse set of input features, the model adeptly captures intricate patterns and relationships within the data, facilitating precise assignment of credit scores to individuals. This accuracy is pivotal for financial institutions striving to make well-informed lending decisions, ensuring that credit is extended in alignment with an individual's risk profile.

Beyond its conventional application in credit scoring, the model exhibits broader utility within the financial sector. Notably, customer segmentation, a vital component of targeted marketing strategies, stands to benefit significantly from the insights derived. Through categorization based on credit risk, financial institutions can tailor marketing efforts to specific segments, optimizing resource allocation and enhancing the efficiency of promotional campaigns.

The model's performance metrics, encompassing precision, recall, and accuracy, offer a comprehensive evaluation of its efficacy. Precision, reflecting the proportion of correctly identified positive instances among all instances predicted as positive, underscores the model's capability to avoid false positives. In the context of credit scoring, precision assumes significance in ensuring that individuals identified as high-risk genuinely manifest characteristics warranting a cautious lending approach. Conversely, recall, synonymous with sensitivity, gauges the proportion of actual positive instances correctly identified by the model. A heightened recall rate underscores the model's effectiveness in capturing individuals genuinely posing credit risks, minimizing the likelihood of overlooking potentially risky borrowers.

In conclusion, the logistic regression model tailored for credit scoring emerges as a sophisticated and versatile tool with diverse applications in the financial domain. Its precision in predicting credit scores, coupled with nuanced insights into customer behavior, positions it as a valuable asset for financial institutions seeking to elevate their decision-making processes. As the financial landscape evolves, models of this nature become indispensable for staying ahead in a dynamic and data-driven industry.

**REFERENCES**

Levinson, J. J. (2003). The Birth of Credit Scoring. Journal of Consumer Affairs, 37(2), 250–274.

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Goetzmann, W. N., & Rouwenhorst, K. G. (2005). The Origins of Value: The Financial Innovations that Created Modern Capital Markets. Oxford University Press.

Hendricks, K. B., Patel, P. C., & Zeckhauser, R. J. (1993). Hot Markets, Investor Sentiment, and IPO Pricing. The RAND Journal of Economics.

Altman, Edward. 1968 “Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy.” Journal of Finance 23 (4): 589–609