“Prediction of Loan Eligibility Through Logistic Regression Algorithm”

Submission Date:

28/04/2024

Submitted by

Group Name:

ASDS 12 Bachelors

Group Members:

Khandoker Toufiq Amin Rumi (20231235)

Moshiur Rahman (20231233)

Md. Shaifullah (20231204)

Kazi Muhtasim Rafid (20231224)

Submitted to

Farhana Akter Bina

Assistant Professor

Department of Statistics & Data Science

Jahangirnagar University

1. Introduction

Lending money is a crucial aspect of the financial world, where individuals or organizations borrow funds for various purposes. However, ensuring that borrowers repay their loans is essential for lenders to sustain their operations and manage risks effectively. To address this challenge, lenders often analyze borrower characteristics and financial data to assess creditworthiness and predict loan repayment outcomes. Regression analysis, a statistical method, is commonly used for this purpose, allowing lenders to understand the relationship between borrower attributes and loan repayment behavior.

The dataset used in this analysis is sourced from LendingClub.com, a platform connecting borrowers with investors. It covers loan data from 2007 to 2010 and includes diverse borrower information, such as credit policy adherence, loan purpose, interest rates, income levels, and credit history. By examining this dataset and applying regression analysis techniques, we aim to develop predictive models that can assist lenders in making informed decisions regarding loan approval and risk management.

2. Objectives:

The objective of the study incorporates:

1. Model Development: Develop regression models to predict the likelihood of loan repayment based on borrower attributes. By analyzing historical loan data, it is aimed that the identification of significant factors influencing repayment behavior and construct predictive models to forecast future repayment outcomes.
2. Feature Importance Analysis: Determine the relative importance of different borrower characteristics in predicting loan repayment.

3. Methodology

3.1 Data Collection

The data used in this project is the Loan Eligible Dataset from Kaggle by Sara Mahdavi [1]. This data is data from the Lending Club company which is a financial services company headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission, and to offer loan trading on a secondary market. They are present in all urban, semi-urban, and rural areas. The customer first applies for a mortgage loan after which the company validates the customer's eligibility for the loan.

3.2.1 Dataset Features:

a. **The Credit Policy**: Indicates whether the customer meets the credit underwriting criteria of LendingClub.com. It's coded as 1 if they meet the criteria and 0 if they don't.

b. **Purpose**: Describes the reason for the loan. It includes categories such as "credit_card", "debt_consolidation", "educational", "major_purchase", "small_business", and "all_other".

c. **Interest Rate**: Represents the interest rate of the loan, given as a proportion. For instance, an interest rate of 11% is stored as 0.11. Higher rates are assigned to borrowers judged to be riskier by LendingClub.com.

d. Installment: Indicates the monthly installment amount owed by the borrower if the loan is funded.

e. **Logarithm of Annual Income**: Represents the natural logarithm of the borrower's self-reported annual income.

f. **Debt-to-Income Ratio (DTI):** Shows the ratio of the borrower's debt to their annual income. It's calculated by dividing the amount of debt by the annual income.

g. **FICO Credit Score**: Reflects the borrower's FICO credit score, a measure of creditworthiness.

h. **Days with Credit Line**: Represents the number of days the borrower has held a credit line.

i. **Revolving Balance**: Shows the borrower's revolving balance, which is the amount left unpaid at the end of the credit card billing cycle.

j. **Revolving Line Utilization Rate**: Indicates the borrower's revolving line utilization rate, which measures the proportion of the credit line used relative to the total credit available.

k. **Inquiries in the Last 6 Months**: Reflects the number of inquiries made by creditors on the borrower's credit report in the last 6 months.

l. **Delinquencies in the Last 2 Years**: Represents the number of times the borrower has been 30+ days past due on a payment in the past 2 years.

m. **Public Records**: Indicates the number of derogatory public records associated with the borrower, such as bankruptcy filings, tax liens, or judgments. e dataset contain these feature.

3.2 Data Preparation

Data preparation techniques used to prepare data before being processed into machine learning models include:

3.1.1 Identifies a Null and Missing value:

Null or missing values in the dataset can significantly affect the performance of machine learning models. Techniques such as identifying and detecting null values in the dataset are essential. This could involve using functions like .isnull() or .isna() in Python libraries like Pandas to identify where null values exist in the dataset.

Table 1: Null value status of the dataset

Column Name	Null Value Count
credit.policy	0
purpose	0
int.rate	0
installment	0
log.annual.inc	0
dti	0
fico	0
days.with.cr.line	0
revol.bal	0
revol.util	0
inq.last.6mths	0
deling.2yrs	0
pub.rec	0
not.fully.paid	0

As the dataset contain no missing value or null value there is no need to remove or modify the data.

3.1.2 Changing categorical values to numerical values

Many machine learning algorithms, including logistic regression, require numerical input data. Therefore, categorical variables need to be encoded into numerical values. This can be achieved through techniques like one-hot encoding or label encoding, depending on the nature of the categorical variables and the requirements of the algorithm. In this step we use the fit_transform() function from the LabelEncoder class, which is in the sklearn.preprocessing library. This step of changing to a numerical value is useful for several modeling algorithms that can only process data in numerical form, so that our data is safe and can be used regression analysis.

3.1.3 Data Splitting

Before training any machine learning model, it's essential to split the dataset into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance. Typically, data is split into a training set (used to train the model) and a testing set (used to evaluate the model's performance) using techniques like random sampling or cross-validation. At this stage we divide the data into a training portion of around 80% and a testing portion of around 20%.

3.3 Technologies and Tools

In case of data science and machine learning, Python serves as the most effective programming language due to its versatility, extensive libraries, and wide community support. When it comes to predictive modeling, especially for tasks like predicting loan eligibility through logistic regression, several key technologies and tools come into play. Here are some essential components that's used in this project:

a. Python:

Python stands as the primary programming language in the data science domain. Its readability, flexibility, and vast ecosystem of libraries make it an ideal choice for data manipulation, analysis, and modeling tasks.

b. Pandas:

Pandas is a powerful library for data manipulation and analysis. It provides data structures like DataFrame, which is particularly useful for handling structured data such as loan application datasets. With Pandas, you can easily clean, transform, and preprocess data before feeding it into a machine learning model.

c. NumPy:

NumPy is the fundamental package for numerical computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently. NumPy arrays serve as the building blocks for many other libraries, including Pandas.

d. Matplotlib:

Matplotlib is a popular plotting library that enables the creation of various types of visualizations, ranging from simple line plots to complex heatmaps. Visualizing data is crucial for understanding patterns, relationships, and trends, which is essential in the exploratory data analysis (EDA) phase of any data science project.

e. SciPy:

SciPy builds upon NumPy and provides additional functionality for scientific computing. It offers modules for optimization, integration, interpolation, and more. While not directly used in logistic regression, SciPy often complements the workflow by providing tools for advanced statistical analysis and optimization techniques.

f. Seaborn:

Seaborn is a statistical data visualization library that works closely with Matplotlib. It provides a high-level interface for creating attractive and informative statistical graphics. Seaborn simplifies the process of creating complex visualizations, making it easier to explore relationships between variables in the dataset.

g. scikit-learn (sklearn):

Scikit-learn is an extensive machine learning library in Python, offering tools for various tasks such as classification, regression, clustering, and dimensionality reduction. For logistic regression, sklearn provides a straightforward implementation along with utilities for data preprocessing, model evaluation, and hyperparameter tuning.

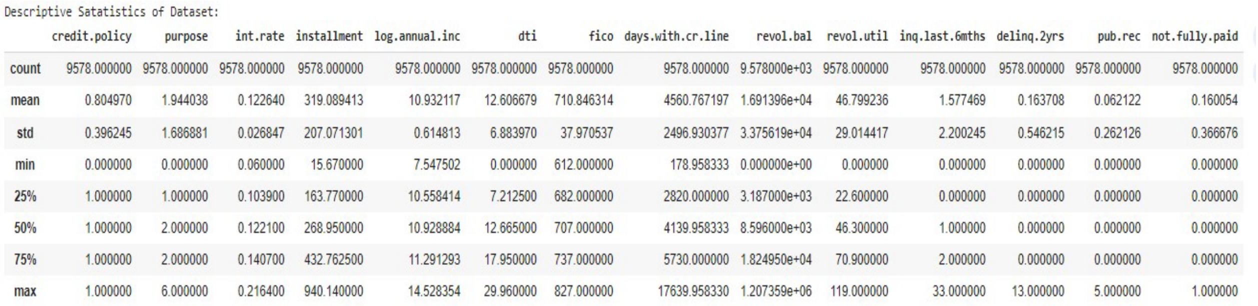
Key functionalities within scikit-learn for logistic regression include:

- **make_classification**: Generate synthetic datasets for classification tasks, which can be useful for testing and prototyping models.
- **train_test_split**: Split the dataset into training and testing sets to evaluate model performance on unseen data.
- **StandardScaler**: Standardize features by removing the mean and scaling to unit variance, ensuring that all features contribute equally to the model.
- **LogisticRegression**: Implement logistic regression, a widely-used algorithm for binary classification tasks. It estimates the probability that a given instance belongs to a particular class based on the input features.

4. Analysis of Dataset

4.1 Descriptive Statistics:

The figure 2 represent the overall statistics of the dataset.



Descriptive Satatistics of Dataset:

	credit.policy	purpose	int.rate	installment	log.annual.inc	dti	fico	days.with.cr.line	revol.bal	revol.util	inq.last.6mths	delinq.2yrs	pub.rec	not.fully.paid
count	9578.000000	9578.000000	9578.000000	9578.000000	9578.000000	9578.000000	9578.000000	9578.000000	9.578000e+03	9578.000000	9578.000000	9578.000000	9578.000000	9578.000000
mean	0.804970	1.944038	0.122640	319.089413	10.932117	12.606679	710.846314	4560.767197	1.691396e+04	46.799236	1.577469	0.163708	0.062122	0.160054
std	0.396245	1.686881	0.026847	207.071301	0.614813	6.883970	37.970537	2496.930377	3.375619e+04	29.014417	2.200245	0.546215	0.262126	0.366676
min	0.000000	0.000000	0.060000	15.670000	7.547502	0.000000	612.000000	178.958333	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	1.000000	0.103900	163.770000	10.558414	7.212500	682.000000	2820.000000	3.187000e+03	22.600000	0.000000	0.000000	0.000000	0.000000
50%	1.000000	2.000000	0.122100	268.950000	10.928884	12.665000	707.000000	4139.958333	8.596000e+03	46.300000	1.000000	0.000000	0.000000	0.000000
75%	1.000000	2.000000	0.140700	432.762500	11.291293	17.950000	737.000000	5730.000000	1.824950e+04	70.900000	2.000000	0.000000	0.000000	0.000000
max	1.000000	6.000000	0.216400	940.140000	14.528354	29.960000	827.000000	17639.958330	1.207359e+06	119.000000	33.000000	13.000000	5.000000	1.000000

Fig. 1: Overall Statistics of dataset

The dataset contains information related to credit policies and various financial attributes of borrowers:

- ➢ The majority of borrowers (approximately 80.5%) meet the credit policy criteria.
- ➢ Interest rates vary from 6.00% to 21.64%, with an average of 12.26%.
- ➢ Installment amounts range from $15.67 to $940.14, with an average of $319.09.
- ➢ The average log annual income is 10.93, with a standard deviation of 0.61.

- Debt-to-income ratios range from 0 to 29.96, with an average of 12.61.
- FICO scores range from 612 to 827, with an average of 710.85.
- The average number of days with a credit line is 4560.77, with a standard deviation of 2496.93.
- Revolving balances vary widely, from $0 to $1,207,359.00, with an average of $16,913.96.
- Revolving utilization rates range from 0% to 119%, with an average of 46.80%.
- On average, borrowers have 1.58 inquiries in the last 6 months and 0.16 delinquencies in the last 2 years.
- The average number of public records is 0.06.
- Approximately 16% of borrowers have not fully paid their obligations.

4.2 Purpose of initiating loan:

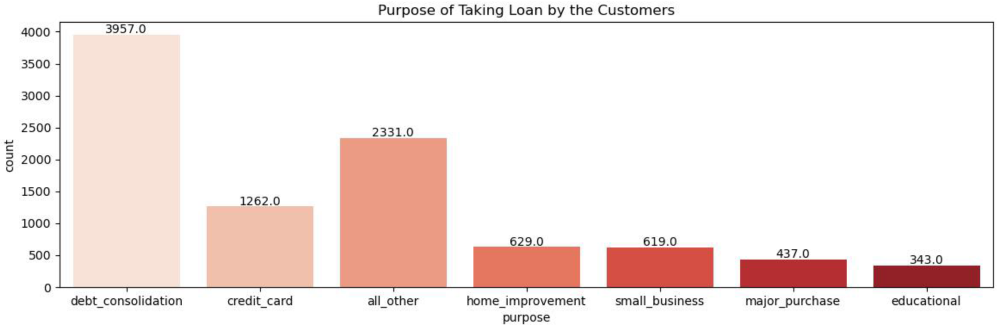


Fig. 2: Distribution of loan purpose among customers

From the figure 3 we can see the dept consolidation is the most common reason for taking loan followed by credit card debt and other unspecified uses where education is the least common reason. This trend suggests people are more likely to use loans to manage existing debt rather than for financing education

4.3 Interest Rate distribution among debtor:

Interest rate in the dataset may vary between 6% to 21.6% where average interest rate is 12.2%. Figure 4 represent the interest rate distribution among the debtor.

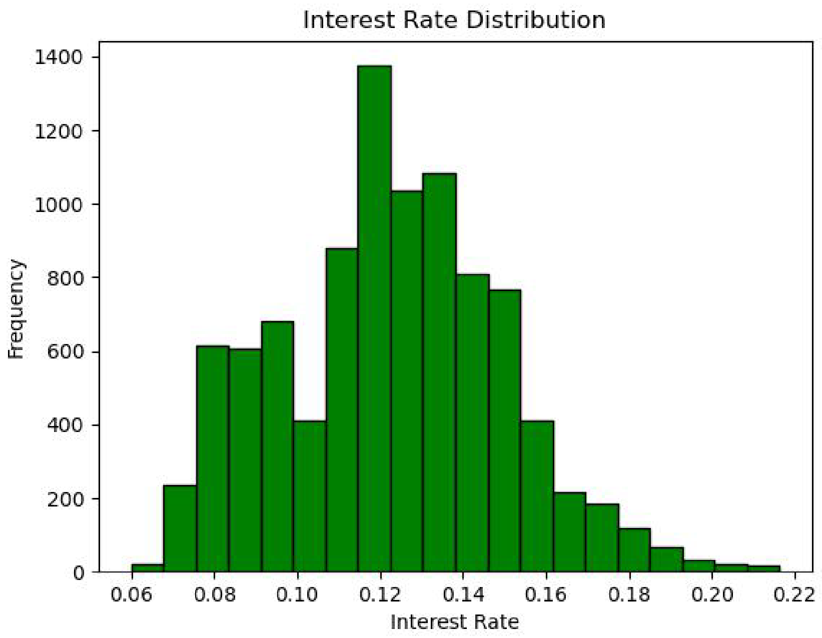


Fig. 3: Interest Rate distribution among the debtor

4.4 Annual income distribution

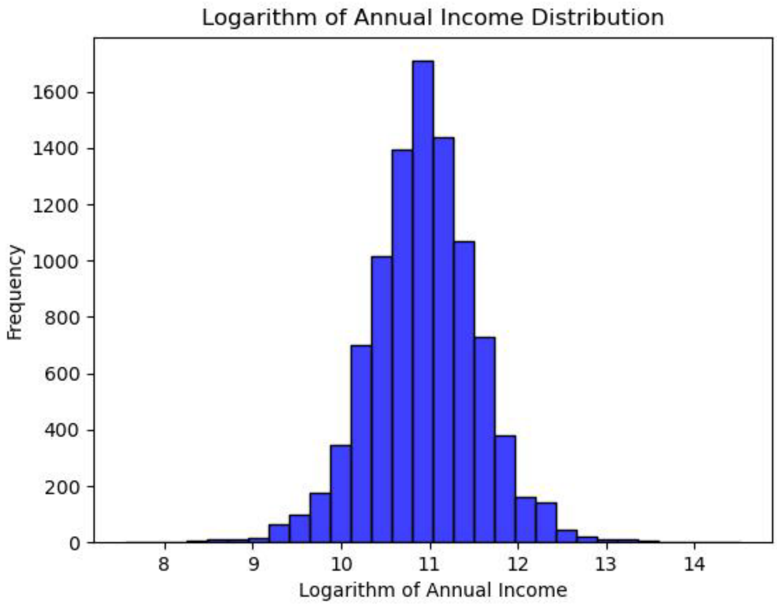


Fig. 4: Logarithm of annual income distribution

The distribution of income is skewed almost uniform. The most common log income appears to be in the 10 to 12 range. Fewer debtors have log incomes at the ends of the distribution, which are below 9 and above 13.

4.5 Relation between FICO score and Interest Rate:

FICO score is a measure of a person's creditworthiness, and interest rate is the percentage of a loan that a borrower is charged. In the scatter plot, each dot represents a borrower. The horizontal axis (x-axis) shows the borrower's FICO score, and the vertical axis (y-axis) shows the interest rate on the borrower's loan. The scatter plot shows a negative correlation between FICO score and interest rate. This means that borrowers with higher FICO scores tend to get lower interest rates on their loans. This is because borrowers with higher FICO scores are considered to be a lower risk to lenders, so lenders are willing to charge them a lower interest rate.

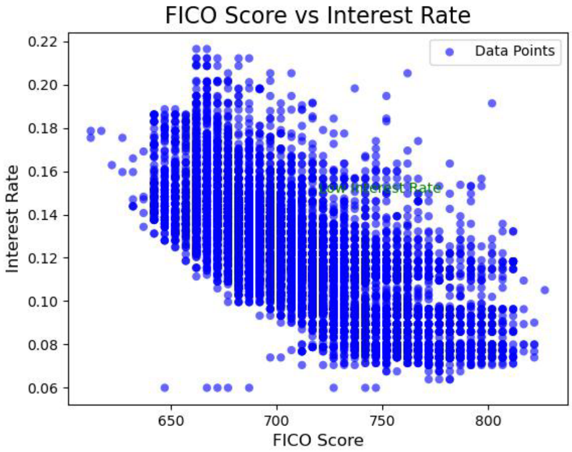


Fig. 5: Relationship between FICO Score and Interest Rate

4.6 Debt to income ratio:

Debt-to-Income Ratio (DTI) shows the ratio of the borrower's debt to their annual income. It's calculated by dividing the amount of debt by the annual income.

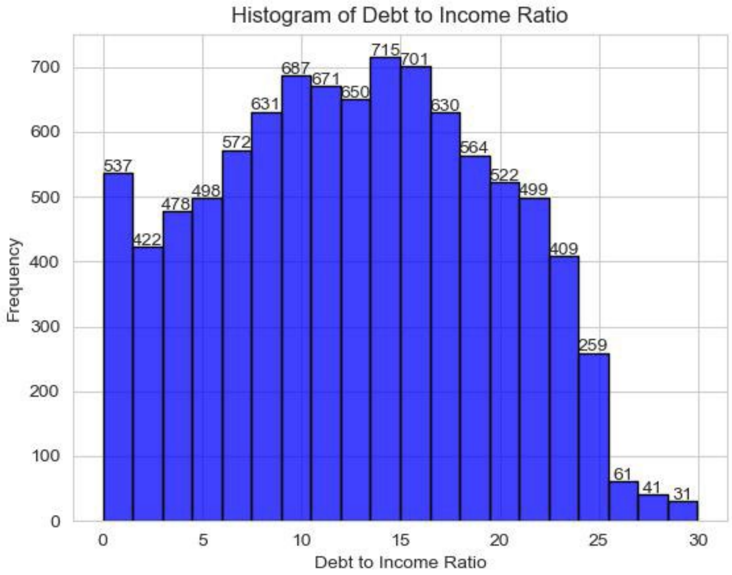


Fig. 6: Debt to income Ratio Histogram

The debt to income ratio vary by 0 to almost 30% with an average 12.6%. The maximum debt to income ratio appears to be clustered between 14 and 15.

4.7 -Correlation map

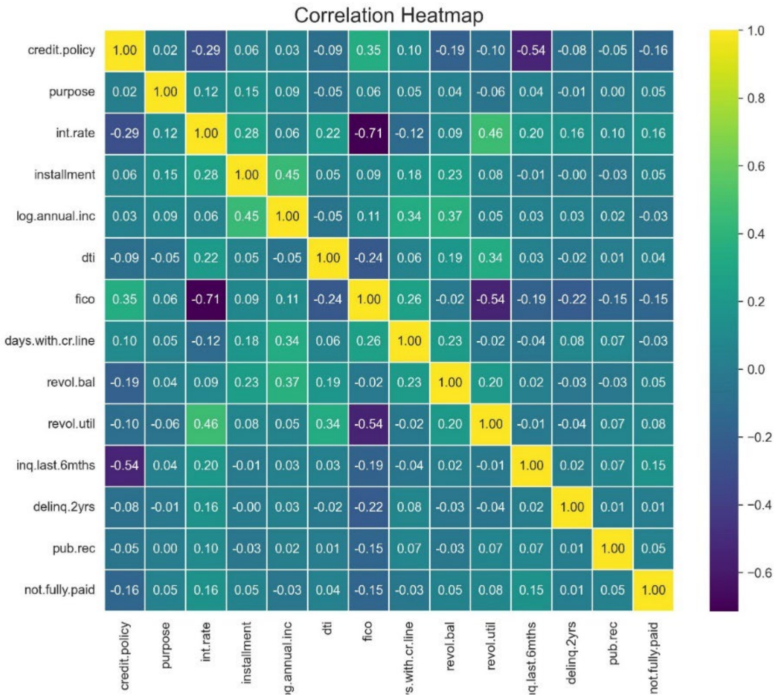


Fig. 7: Correlation Heatmap of dataset

The relationship between various variable to Credit Policy:

- **Purpose**: The correlation coefficient is close to zero (0.017569), indicating a very weak positive correlation between 'credit policy' and the purpose of the loan. This suggests that there is almost no relationship between whether a borrower meets the credit policy and the purpose of the loan they are applying for.

- **Interest Rate**: The correlation coefficient is negative (-0.294089), indicating a moderate negative correlation between 'credit policy' and the interest rate. This suggests that borrowers who meet the credit policy criteria tend to have lower interest rates on their loans.

- **Installment**: The correlation coefficient is positive (0.058770), indicating a very weak positive correlation between 'credit policy' and the installment amount. This suggests that there is a slight tendency for borrowers meeting the credit policy to have higher installment amounts.

- **Logarithm of Annual Income**: The correlation coefficient is positive (0.034906), indicating a very weak positive correlation between 'credit policy' and the logarithm of annual income. This suggests that there is a slight tendency for borrowers meeting the credit policy to have higher annual incomes.

- **Debt-to-Income Ratios**: The correlation coefficient is negative (-0.090901), indicating a weak negative correlation between 'credit policy' and the debt-to-income ratio. This suggests that borrowers meeting the credit policy criteria tend to have lower debt-to-income ratios.

- **FICO**: The correlation coefficient is positive (0.348319), indicating a moderate positive correlation between 'credit policy' and the FICO credit score. This suggests that borrowers who meet the credit policy criteria tend to have higher FICO credit scores.
- **Days with a Credit Line**: The correlation coefficient is positive (0.099026), indicating a very weak positive correlation between 'credit policy' and the number of days with a credit line. This suggests that there is a slight tendency for borrowers meeting the credit policy to have a longer credit history.
- **Revolving Balances**: The correlation coefficient is negative (-0.187518), indicating a moderate negative correlation between 'credit policy' and the revolving balance. This suggests that borrowers meeting the credit policy criteria tend to have lower revolving balances.
- **Revolving Utilization Rate**: The correlation coefficient is negative (-0.104095), indicating a weak negative correlation between 'credit policy' and the revolving utilization rate. This suggests that borrowers meeting the credit policy criteria tend to have lower revolving utilization rates.
- **Inquaries in Last 6 Months**: The correlation coefficient is negative (-0.535511), indicating a strong negative correlation between 'credit policy' and the number of inquiries in the last 6 months. This suggests that borrowers meeting the credit policy criteria tend to have fewer inquiries in the last 6 months.
- **Delinquencies in the Last 2 Years**: The correlation coefficient is negative (-0.076318), indicating a weak negative correlation between 'credit policy' and the number of delinquencies in the past 2 years. This suggests that borrowers meeting the credit policy criteria tend to have fewer delinquencies.
- **Public Record**: The correlation coefficient is negative (-0.054243), indicating a very weak negative correlation between 'credit policy' and the number of derogatory public records. This suggests that borrowers meeting the credit policy criteria tend to have fewer derogatory public records.
- **Not Fully Paid**: The correlation coefficient is negative (-0.158119), indicating a weak negative correlation between 'credit policy' and the target variable 'not.fully.paid'. This suggests that borrowers meeting the credit policy criteria are slightly less likely to be classified as 'not fully paid'.

5. Regression Analysis for Loan Prediction

5.1 Solution Specific:

The implementation of machine learning modeling to predict the eligibility value for credit applications this time has the following specifications:

- Implement modeling using algorithms: Logistic Regression, Random Forest and Boosting.
- The performance assessment of the three models will be made using several metrics/measurement methods, including the following:

  a. Mean Squared Error (MSE):

This metric squares the difference between the predicted and actual values, then takes the final average value (Bickel, 2015)[2].

The MSE formula is as follows:

$$\text{MSE} = 1/n \sum_{i=1}^{n} (y - \bar{y})^2$$

b. Confusion Matrix:

This matrix maps the prediction results into several categories, including:

Table 2: Confusion Matrix Table

Types	Model Prediction	Actual
True Positive	1	1
False Positive	1	0
True Negative	0	0
False Negative	0	1

c. Accuracy:

Accuracy is measured by the following formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

d. Precision is measured by the following formula:

$$\text{Precision} = \frac{TP}{TP+FP}$$

e. Sensitivity is measured by the following formula:

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

f. Area Under The Curve

The area under the curve (area under the curve) or also known as AUC is used as a measure to judge whether a model is good or bad. AUC close to 1 means that the model has good performance, while AUC close to 0.5 indicates that the model has poor performance.

5.2 Modeling:

The model chosen for this solution is a model that uses the Logistic Regression algorithm, because this algorithm is suitable for problems with many independent variables and produces binary output (0/1, Yes/No, Approve/Reject, etc.).

Pros of Logistic Regression:

- Easy to implement.
- Can accommodate multi-variables.
- Not only does it provide a measure of how precise a predictor is (a measure of the coefficient), but also the direction of the association (positive or negative)
- Very fast in classifying unknown records.
- Has good accuracy for simple data sets and performs well when data sets are linearly separable.

Cons of Logistic Regression:

- If the number of observations is smaller than the number of features, Logistic Regression cannot be used because it can cause overfitting.

- The main limitation of Logistic Regression is the assumption of linearity between the dependent and independent variables.
- Can only be used to predict discrete functions. Therefore, the Logistic Regression dependent variable is bound to a set of discrete numbers.
- Non-linear problems cannot be solved by Logistic Regression because they have a linear decision surface. Linearly separable data is rare in real world scenarios.
- Logistic Regression requires an average or no multicollinearity between independent variables.
- It is difficult to derive complex relationships using Logistic Regression. More powerful and concise algorithms such as Neural Networks can easily outperform these algorithms.

5.3 Procedure of Analysis:

In this process, the data is first split into features (x) and the target variable (y). The feature matrix is then further divided into training and testing sets using the train_test_split function. A logistic regression model is instantiated and trained on the training data. Predictions are made on the test set, and performance metrics such as accuracy, precision, recall, and the confusion matrix are calculated to evaluate the model's effectiveness. This systematic approach ensures the model's robustness and provides insights into its predictive capabilities for determining loan eligibility based on the given features.

5.4 Findings:

Evaluation of the performance of machine learning modeling is done in several ways. On this evaluation stage, modeling using Logistic Regression measures its performance with several metrics. Here's an explanation:

a. Accuracy (87.7%): The accuracy metric indicates the overall correctness of the model's predictions. In this case, the model achieves an accuracy of 87.7%, implying that it correctly predicts loan eligibility for approximately 87.7% of the test instances.

b. Precision (0.87): Precision measures the proportion of true positive predictions among all positive predictions made by the model. A precision of 0.87 indicates that out of all instances predicted as eligible for a loan, approximately 87% are indeed eligible.

c. Recall (0.89): Recall, also known as sensitivity, quantifies the model's ability to correctly identify all positive instances in the dataset. A recall of 0.89 suggests that the model captures approximately 89% of all eligible loan applicants.

d. Mean Square Error:
The MSE is a measure of the average squared difference between the actual and predicted values. In this case, the MSE is approximately 0.123. Since MSE values closer to 0 indicate better model performance, this indicates that, on average, the model's predictions have a squared error of about 0.123

e. Confusion Matrix:
Table 3: Confusion Matrix

Types	Logistic Regression Value count
True Positive (TP)	158
False Positive (FP)	25
False Negative (FN)	21
True Negative (TN)	170

The confusion matrix provides a detailed breakdown of the model's predictions. It reveals that the model correctly identifies 158 instances as eligible for a loan (TP) but incorrectly classifies 25 instances as eligible when they are not (FP). Additionally, the model fails to identify 21 instances that are actually eligible (FN), while correctly identifying 170 instances as not eligible (TN).

f. Area Under Curve:
The AUC is a performance metric commonly used for binary classification problem. It measures the ability of the model to distinguish between positive and negative classes. AUC ranges from 0 to 1, where a score closer to 1 indicates a better model. In this case, the AUC is approximately 0.877. This means that the logistic regression model has a good ability to distinguish between the positive and negative classes, with an AUC 0.877 being relatively high.

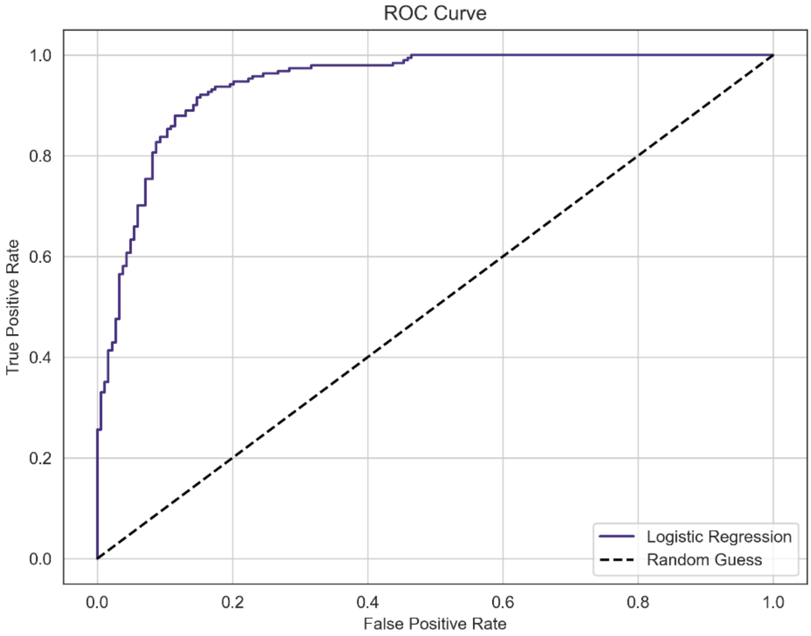


Fig. 8: Area Under Cover Curve

6. Conclusion

The analysis of loan eligibility prediction using machine learning models, particularly Logistic Regression, Random Forest, and Boosting algorithms, provided valuable insights into the relationship between various predictor variables and credit policy. Among the predictor variables examined, the FICO credit score demonstrated a moderate positive correlation with meeting the credit policy criteria, suggesting that applicants with higher FICO scores are more likely to meet the credit policy requirements. Furthermore, the number of inquiries in the last 6 months exhibited a strong negative correlation with credit policy adherence, indicating that applicants with fewer recent inquiries are more likely to meet the credit policy.

The implementation of logistic regression modeling revealed promising results in predicting loan eligibility, with an accuracy of 87.7%, precision of 0.87, and recall of 0.89. The confusion matrix provided a detailed breakdown of the model's performance, highlighting its ability to correctly classify instances as eligible or ineligible for a loan. However, it's essential to note the limitations of logistic regression, including its assumption of linearity between independent and dependent variables, and its inability to handle non-linear relationships.

In conclusion, the logistic regression model shows potential for accurately predicting loan eligibility based on the provided features. Future work could involve further refinement of the model, exploration of additional algorithms such as Random Forest and Boosting, and incorporation of additional features to enhance predictive performance and robustness. Overall, leveraging machine learning techniques offers a promising approach to optimizing loan approval processes and minimizing the risk of default for lending institutions.

7. References

1. Sara Mahdavi, Loan Data, 2024. [Online] Available: https://www.kaggle.com/code/saramah/loan-prediction/notebook [Accessed: 20-April-2024]
2. P. J. Bickel and K. A. Doksum, Mathematical Statistics: Basic Ideas and Selected Topics, vol. 1, 2 vols. CRC Press, 2015

Work Distribution:

Name	ID	Email	Contribution	Percentage	Signature
Leader Name	20231235	toufiqamin1995@gmail.com	Supervising & Leading	Work load (25%)	
Member	20231233	rahman.moshiur996@gmail.com	Data Analysis & Presentation	Work load (25%)	
Member	20231204	saifullahbinshiraj@gmail.com	Report & Presentation	Work load (25%)	
Member	20231224		Data Analysis & ML implementation	Work load (25%)	