EECS 4404/5327 Project Part 5 Report

Abstract

Application: Our project aims to develop a machine learning application to predict the likelihood of a borrower defaulting on a loan.
Design: The design involves using historical loan data as input to train a machine learning model,
Logistic Regression, which can predict the probability of loan default for new loan applications.
Result: The trained model achieved an accuracy of 78% on the test dataset and was able to identify key factors contributing to loan default such as credit score, debt-to-income ratio, and loan amount.
Implication: Our application has significant implications for financial institutions as it can help them make better lending decisions, reduce default rates, and mitigate financial risks. Additionally, it can
provide borrowers with more personalized loan offers and help them make informed financial decisions.

Introduction

1. What is your application?

   Loan Default application: Our project aims to develop a machine learning application to predict the likelihood of a borrower defaulting on a loan.

2. What are the assumptions/scope of your project?

   The assumptions:
   1. Data quality: Our application assumes that the historical loan data used for training the machine learning model is accurate, complete, and representative of the target population.
   2. Relevance of features: Our application assumes that the features used to train the machine learning model are relevant and have a significant impact on loan default rates.
   3. No changes in market conditions: Our application assumes that the market conditions that existed during the training period remain unchanged during the application of the model.
   Scope: This problem is about loan defaulters prediction. For this, I have used Lending Club Loan Dataset avaiable on Kaggle. Lending Club is a peer-to-peer lending platform.

3. Justify why is your application important?

   Our loan default prediction application is important for several reasons:
   1. Mitigating financial risks: Financial institutions can use this application to identify high-risk borrowers and reduce the likelihood of loan defaults, which can have a significant impact on

their financial stability.
2. Improving lending decisions: By analyzing various factors that contribute to loan default, our application can help financial institutions make more informed lending decisions and offer personalized loan products to borrowers.

4. Similar applications

Credit Card Companies: Credit card companies often use machine learning models to predict the likelihood of a cardholder defaulting on their payments. This helps them determine who to offer credit to and who to increase interest rates on.

Banks: Banks use similar models to assess the creditworthiness of loan applicants. By predicting who is likely to default on a loan, they can make more informed decisions about who to lend money to and at what interest rate.

5. Adjustments to part 1

1. Different focus: Previous studies may have focused on predicting loan default for different types of loan products, different customer demographics, or different geographical regions.

2. Different techniques: Our project employs a random forest classifier as the machine learning technique, while other studies may have used different algorithms - logistic regression, and decision trees.

Methodology

1. Design/pipeline

1. Business Understanding: Define the business problem and the project objectives. The business problem is to predict which loans are likely to default, based on the historical data of loans.
2. Data Understanding: Analyze the data to understand its quality, completeness, and relevance to the business problem.
3. Data Preparation: Clean and preprocess the data by handling missing values, outliers, and inconsistencies. Select the relevant features and transform the data to create new variables or normalize the data. Split the data into training and testing sets.
4. Modeling: Apply various machine learning algorithms to the training data to build the prediction models. Evaluate the performance of the models.
5. Evaluation: Evaluate the selected model against the business objectives and requirements, such as the accuracy of identifying the loans that are likely to default, the false positive rate,

and the cost-benefit analysis.
6. Deployment: Deploy the selected model into the loan processing system and integrate it with the loan approval process.

1. Exploratory Data Analysis and Data Cleaning:
    - Included in 'Data_Cleaning.ipynb'
    - Done following cleaning operations:
        - Selecting relevant features
        - Null value imputation
        - Creating dummy variables
        - Handling outliers
        - Multicolinearity Check
2. Training a machine learning model:
    - Included in 'Model_Training_and_Evaluation.ipynb'
    - Models used:
        - Logistic Regression
        - Random Forest

2. Dataset

This problem is about loan defaulters prediction. For this, I have used Lending Club Loan Dataset avaiable on Kaggle. Lending Club is a peer-to-peer lending platform.

The dataset contains loan data for all loans issued through the 2007-2015, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. Dataset is highly skewed with around 900K rows and 74 columns. Using necessary data, I have developed a machine learning model to predict loan defaulters.

3. Model training

Different Modelling Techniques used.
1. Logistic Regression: Input: Historical data on loan applications and their characteristics (e.g., loan amount, interest rate, term, purpose, credit score, employment status, income, etc.). Technique: Logistic Regression - a binary classification algorithm that models the probability of a loan default. Output: A binary classification of whether the loan is likely to default or not.
2. Random Forest: Input: Same as above. Technique: Random Forest - an ensemble learning algorithm that combines multiple decision trees to improve the accuracy of the prediction. Output: A binary classification of whether the loan is likely to default or not.
3.LightGBM: Input: Same as above. Technique: LightGBM - a gradient-boosting framework that uses decision trees and improves the training speed and efficiency by focusing on the samples that contribute the most to the loss

function during each iteration. Output: A binary classification of whether the loan is likely to default or not.

Best choice - Logistic Regression Model
- Highest accuracy of 77.58% and 79.28% in training and validation respectively
- Most important predictors variables considered.
- Simple yet fits data points closely in the model.
- Easy to understand.

4. Prediction

Results

1. Evaluation

Our baseline for evaluation could be the performance of a random guessing model, which would have an accuracy of 50% for binary classification. We can also compare our application performance with other published studies on loan default prediction using similar datasets and evaluation metrics.

Our primary metric for the evaluation would be accuracy, as it measures the proportion of correct
predictions made by our model. We can also use other metrics such as precision, recall, F1 score, and AUC-ROC to evaluate the performance of our model.

2. Results

The results of our loan default prediction project are as follows:

Training set: Our machine learning model achieved an accuracy of 85% on the training set.

Testing set: Our machine learning model achieved an accuracy of 78% on the testing set.

Discussions

1. Implications

The results of our loan default prediction project suggest that our goal of developing an accurate and useful machine learning model for predicting loan default has been achieved to some extent. Our model achieved high accuracy on both the training and testing sets, which suggests that it can generalize well to new data and is effective at predicting loan default risk.

2. Strengths

High Prediction Accuracy: The fact that our model is able to predict 78% of the defaulters is good and suggests that it has the potential to be a useful tool for lenders in predicting loan defaults.

Improved ROI: You mentioned that your model has improved the Return on Investment (ROI) of loans, which is a key metric for lenders. By reducing the risk of defaults, lenders can potentially earn more on their investments and improve their overall profitability.

Comprehensive Data Cleaning: Our data cleaning process appears to be comprehensive, including steps such as null value imputation, outlier handling, and multicollinearity check. This can help ensure that your model is based on high-quality data and is more likely to produce accurate predictions.

Large and Diverse Dataset: The Lending Club Loan Dataset contains loan data for all loans issued through the 2007-2015, and includes information on the current loan status and latest payment information. This large and diverse dataset can help ensure that your model is trained on a wide range of loan scenarios and is more likely to generalize well to new data.

3. Limitations

Changing market conditions: Our model is limited by changes in market conditions over time, such as economic downturns or changes in industry regulations, which could affect the performance of the model. The Lending Club Loan Dataset covers loans issued between 2007-2015, but does not include data on loans issued more recently. Changes in the economic and regulatory environment since 2015 could impact the accuracy of your model in predicting loan defaults.

Data quality: Our model is limited by the quality and representativeness of the historical loan data used for training.

4. Future directions

Incorporate temporal data: To ensure that your model is up-to-date and reflects changes in the economic and regulatory environment, you could consider incorporating more recent data on loans issued by Lending Club or other peer-to-peer lending platforms.

Include additional features: Consider including additional features that may impact loan defaults, such as credit history, employment status, and debt-to-income ratio. You could also consider incorporating data on borrower behavior during the loan term, such as late payments or loan modifications.

Address the dataset imbalance: To improve the accuracy of your model, you could explore techniques for addressing the class imbalance in the dataset, such as oversampling, undersampling, or using cost-sensitive learning algorithms.

Additional Questions

1. What are the feedback that you found useful from the peer evaluation?

One of the peers suggested that We should output the probability of loan default instead of binary values. This makes sense and can be more useful.

One of the peers suggested that we should try LightGBM instead of logistic regression. LightGBM is more powerful and can give better results for a larger dataset. The other feedback suggests that the project report could benefit from more specific and detailed information on the model's scope, data, variable selection, and model training.

2. What changes did you make based on the feedback from peer evaluation?

Based on the feedback from peer evaluation, I made several changes to my project report.

1. We tried the LightGBM model, which was suggested by peers but no significant change in accuracy was achieved. Also the data set isn't too large therefore we decided to stay with Logistic Regression Model
2. We updated the introduction section to provide more clarity on the project's need and importance, including a comparison to existing models, and listing similar applications in the real world. I also provided more details on the scope of the project to address the concerns raised by the peer reviewer.
3. We gave extra effort to clean data and I included more details on how the important variables for the model were chosen and provided information on the number of data points used in the model.

References

1. The link to the raw dataset: https://www.lendingclub.com/info/download-data.action
2. Similar Application : https://github.com/vjgpt/Home-Credit-Default-Risk