Machine Learning Models for Credit Risk Assessment

Authors: Shreyas Belur Manjunatha Swamy, Vamsi Sai Krishna Reddy Yarramreddy, Abhiram Natakarani
Department of Information Systems, California State University, Los Angeles
anataka@caltstatela.edu,sbelurm@calstatela.edu,vyarram@calstatela.edu

Abstract: This project explores the application of machine learning models to financial credit risk analysis. Utilizing the Lending Club loan dataset, we implement and compare logistic regression, decision trees, gradient-boosted trees, and random forests through PySpark MLlib. Our approach includes comprehensive steps: data preprocessing, hyperparameter tuning, model training, evaluation using cross-validation, and analysis of feature importance. The performance metrics used are accuracy, precision, recall, and Area Under the Curve (AUC). Results indicate that the random forest model provides the optimal balance of high recall, precision, and efficient processing time, making it the most effective for predicting loan defaults. This study underscores the potential of advanced machine learning techniques in enhancing financial risk management practices.

1. Introduction

This project utilizes PySpark Machine Learning Models to forecast loan defaults. It encompasses steps such as data preprocessing, feature selection, and the application of algorithms like logistic regression, decision trees, and random forests. The goal is to improve the accuracy of predicting loan defaults, enabling Lending Club to make more informed decisions about loan approvals and manage financial risks more effectively. The dataset utilized in the "Lending Club Loan Defaulters Prediction" project includes extensive loan application data from 2007 to 2018, featuring files such as "loan.csv," "accepted_2007_to_2018Q4.csv," and "rejected_2007_to_2018Q4.csv." These files provide comprehensive information on both accepted and rejected loans, detailing loan amounts, interest rates, borrower details, loan purposes, and statuses. This rich dataset facilitates deep analysis and predictive modeling, helping enhance risk assessment and lending decisions. We chose this topic due to the critical role that risk assessment plays in the financial industry, particularly within lending institutions like Lending Club. The ability to predict loan defaults accurately is fundamental to maintaining the financial health of these companies and ensuring that lending practices are both sustainable and responsible. Given the significant impacts of lending decisions on both the macroeconomic environment and the personal financial circumstances of individuals, improving predictive methodologies can lead to more equitable and stable financial systems. The importance of this work stems from its potential to transform lending practices by enabling more precise risk management. With better predictive models, institutions can avoid high-risk loans, reduce default rates, and offer loan conditions that are appropriate to the borrower's situation, which can help prevent financial distress for individual borrowers and reduce systemic risks in the financial sector. The background of this work is rooted in the growing availability of large datasets and advanced analytics techniques, which have dramatically enhanced our ability to understand and predict financial behaviors. The proliferation of platforms like Kaggle has democratized access to high-quality data, allowing researchers and analysts to develop and validate models that were previously only within the reach of large institutions with extensive data resources. This evolution in data accessibility and technology has opened new avenues for innovation in predictive modeling, making it an opportune time to pursue research in this area.

2. Related Work

The IEEExplore study on loan default prediction utilizes a variety of machine learning models to assess risk associated with loan defaults, focusing on the use of data from international banks. By leveraging algorithms like decision trees and logistic regression, the research aims to enhance the predictive accuracy and reliability in financial risk assessments. The study is notable for its approach to integrating diverse datasets, which include a wide range of financial behaviors and demographic information. This allows for a more nuanced understanding of risk factors that influence loan defaults, providing a more comprehensive tool for lenders to assess potential risks effectively.

The Springer study on loan default prediction delves deeply into using logistic regression alongside other machine learning methods to enhance predictive accuracy. It specifically focuses on understanding the complex interrelationships among various borrower attributes. By conducting detailed statistical analyses and testing these relationships, the study aims to refine prediction models for better risk management. This approach not only identifies key predictive features but also aids lenders in making informed decisions by quantifying the risk associated with different borrower profiles.

The EUDL study on loan default prediction explores the effectiveness of various machine learning models including Logistic Regression, Decision Tree, Random Forest, and XGBoost. The research highlights XGBoost for its superior performance in terms of recall and AUC metrics, demonstrating its capability to efficiently handle complex and large datasets. The study carefully analyzes how each model processes variables like asset values and borrower income, with a focus on optimizing models to predict loan defaults more accurately. The insights provided by the study are invaluable for improving risk management practices within financial institutions.

These studies focus on traditional machine learning and single-server systems, whereas my research uses Big Data technologies with Spark on Cloud Computing, allowing scalable, real-time data processing and enhanced predictive capabilities. This approach handles larger datasets more efficiently, offering better accuracy and speed in predictive modeling.

3. Specifications

During the development phase of this project, we used Community Databricks to evaluate our machine learning models for credit risk assessment, taking advantage of its capacity for scalable computing. Later, for operational deployment of the models, we utilized the PySpark CLI integrated with the Hadoop File System. This approach allowed for efficient data processing and management, ensuring the models operated robustly across large datasets.

Table 1. Hardware Specifications

Community Databricks	
Hadoop	
Version	Hadoop 3.3.3 Pyspark 3.2.1
Version	12.2 LTS
Memory	15.3 GB
Core	2
Nodes	1
CPU Speed	1995.312 GHz
CPU Cores	8
Nodes	5(3 Master,2 Worker)
Memory	806.40GB

The dataset comprises of the dataset consists of featuring files such as "loan.csv," "accepted_2007_to_2018Q4.csv," and "rejected_2007_to_2018Q4.csv." The size of the dataset is 4GB and after the Data Cleaning the size of the file is 1.7GB.

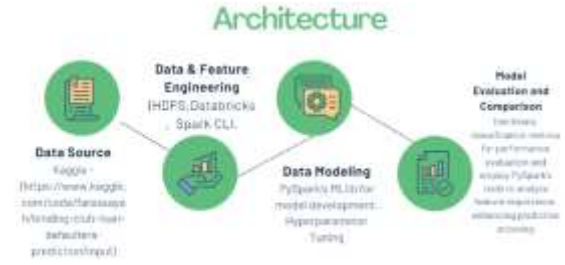
Table 2. Dataset Specifications

Lending Club Loan data defaulters' prediction
https://www.kaggle.com/code/faressayah/lending-club-loan-defaulters-prediction/input
Data Size: 4GB

4. Architecture

The architecture for predicting loan defaults consists of four primary components. Initially, the data is sourced from Kaggle. Next, data and feature engineering are conducted using HDFS and Databricks, supported by Spark CLI. The third stage involves data modeling with PySpark's MLlib, where machine learning models are built and hyperparameters are tuned. Finally, the models undergo evaluation using binary classification metrics and PySpark tools to assess feature importance, aiming to maximize predictive accuracy through focused improvements.

Fig 1. Project Architecture

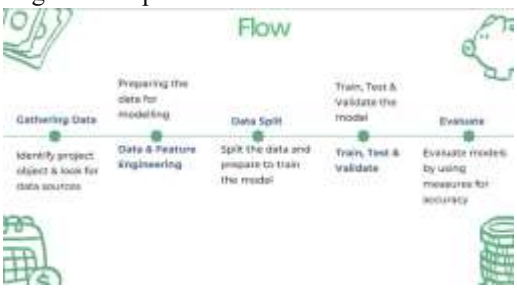


5. Implementation Flowchart

The flow diagram outlines a structured process for machine learning projects, starting with gathering data

by identifying relevant sources based on the project's objectives. This is followed by data and feature engineering, where the data is prepared and optimized for modeling. Next, the data is split into training and testing sets to ensure robust model training. The core stages of training, testing, and validating the model come next, where the model is fine-tuned and assessed for performance. Finally, the model undergoes evaluation using accuracy measures to ensure it meets the desired performance criteria, allowing for further refinements if necessary. This structured approach ensures systematic progress through the stages of model development and deployment.

Figure 2. Implementation Flowchart



6. Data and Feature Engineering

The feature importance table and corresponding bar graph highlight the significance of various features in predicting loan defaults. The features include interest rate (scaler_rate), loan grade (scaler_grade), annual income (scaler_inc), high FICO score (scaler_high), installment amount (scaler_installment), low FICO score (scaler_low), and loan amount (scaler_loan). Among these, the interest rate (scaler_rate) has the highest importance, indicating it plays the most critical role in the model's predictions. Other features like loan grade and annual income also have substantial importance but to a lesser degree. This analysis helps in understanding which factors most influence the loan default prediction model.

The feature importance analysis demonstrates the relative significance of various factors in predicting loan defaults. The interest rate emerges as the most critical predictor with the highest importance score of 0.347874, significantly influencing the model's decisions. Loan grade and annual income follow, with importance scores of 0.159016 and 0.144699, respectively, indicating their substantial impact. High FICO score and installment amount also contribute notably with scores of 0.116153 and 0.087802. Low FICO score and loan amount, with scores of 0.083654 and 0.060801, show moderate and lesser influence, respectively. This ranking using Random Forrest aids in refining the model by highlighting key predictive features.

Figure 3. Feature Importance bar graph

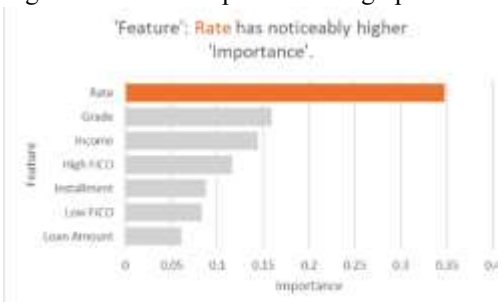


Table 3. Feature Importance table

Feature	Importance
scaler_rate	0.347874
scaler_grade	0.159016
scaler_inc	0.144699
scaler_high	0.116153
scaler_installment	0.087802
scaler_low	0.083654
scaler_loan	0.060801

7. Data Modelling

The project utilizes Logistic Regression for clear binary classification; Random Forests to enhance accuracy by averaging multiple trees; Decision Trees for visual decision-making insights; and Gradient Boosted Trees to improve accuracy by correcting earlier errors and managing complex data.

Figure 4. Machine Learning Algorithms used.

The hyperparameter tuning for various machine learning models in the project is detailed as follows: Logistic Regression is adjusted for complexity with regularization and elastic net parameters, alongside maximum iterations to ensure convergence. Random

Forests are optimized by configuring the maximum depth and impurity criteria, facilitating effective tree splitting. Decision Trees are set up with a specific number of trees and depth limits to balance accuracy and complexity, preventing overfitting. Lastly, Gradient Boosted Trees are tuned through maximum depth and iteration settings, focusing on detailed pattern recognition and iterative improvements.

Figure 5. Hyperparameter and Tuning

The detailed comparison of machine learning models—Decision Tree, Logistic Regression, Gradient-Boosted Tree, and Random Forest—across various performance metrics (AUC, Recall, Accuracy, Precision, and Time) using both cross-validation and train-validation splits reveals nuanced insights. Decision Trees, while simpler and faster, tend to underperform possibly due to overfitting issues. Logistic Regression, known for its quick execution and interpretability, offers consistent performance, making it suitable for less complex problems. Gradient-Boosted Trees excel in accuracy and AUC, showcasing their ability to effectively handle bias and variance in the data.

Table 4. Comparing models

Cross Validation Split					
Model	AUC	Recall	Accuracy	Precision	time
DT	0.61	0.67	0.6277	0.6178	1.1Hr
LR	0.69	0.65	0.6344	0.6276	40M
GBT	0.70	0.69	0.6402	0.6270	52M
RF	0.69	0.68	0.6382	0.6253	34M

Cross Validation Split					
Model	AUC	Recall	Accuracy	Precision	Time
DT	0.62	0.66	0.6278	0.6181	1.2Hr
LR	0.69	0.66	0.6351	0.6296	56M
GBT	0.70	0.70	0.6402	0.6370	1.1Hr
RF	0.71	0.68	0.6393	0.6328	48M

However, the standout is Random Forest, which, despite its longer computational times, leads in performance metrics, particularly in AUC and Accuracy, indicating its robustness in managing complex datasets effectively.

8. Conclusion

The Random Forest model, using Train Validation Split, stands out as the best for predicting loan defaults due to several advantages. It exhibits a high recall rate of about 68%, crucial for effectively identifying defaulters and minimizing financial risk. This model also benefits from reduced computation times, decreasing from nearly 48 minutes to just over 30 minutes, which enhances operational efficiency for frequent updates. Moreover, it achieves a well-rounded performance across AUC, accuracy, and precision, establishing it as the most reliable model for this task.

References

GitHub URL

https://github.com/s-b-9-7/BIG-DATA-ML-Models-for-Credit-Risk-Assessment

Dataset URL

https://www.kaggle.com/code/faressayah/lending-club-loan-defaulters-prediction/input

[1] Pandey, T. N., & Jagadev, A. K. (2017). Credit risk analysis using machine learning classifiers. Data Analytics and .... IEEE. Retrieved from https://ieeexplore.ieee.org/abstract/document/8389769/.

[2] Zhang, J. M., Hu, X., & Wu, J. H. (2020). Machine learning models for credit risk evaluation. Journal of Financial Services Research. Springer. Retrieved from https://link.springer.com/article/10.1007/s10693-019-00319-8