PROJECT AND DATA MANAGEMENT PLAN BY NSUDE ALBERT-22020699

PROJECT OVERVIEW

Github: https://github.com/Albertnsude/CreditDefaultPrediction_22020699_2024-10-24

Analysing and Comparing Machine Learning Approaches for Credit Default Prediction

Summary

In the financial sector, credit default prediction is crucial for managing risk, helping firms identify high-risk clients and take preventive actions. This project evaluates the effectiveness of three machine learning algorithms—XGBoost, Random Forest, and Logistic Regression—in predicting credit defaults. It assesses the predictive accuracy and feature selection of these models to determine the best method for credit risk modelling. Using datasets from Credit Risk Analytics and Kaggle, the study contributes to the growing knowledge in credit risk management and machine learning applications.

Background

To minimize losses, optimize lending strategies, and assess the likelihood of customer default, banks and financial institutions rely on credit risk prediction models. While logistic regression has traditionally been used, the advent of machine learning (ML) algorithms like Random Forest and XGBoost has introduced more accurate and efficient alternatives. These models can process large datasets and identify complex patterns, making them vital for risk management decision-making (Noriega et al., 2023). Several researchers have contributed to this field. Bahnsen et al. (2015) introduced cost-sensitive logistic regression to reduce the misclassification of high-risk customers, while Ngai et al. (2010) and Lessmann et al. (2015) demonstrated ML's superiority over traditional techniques in credit scoring and fraud detection. Consoli et al. (2021) emphasized the role of data science and ML in improving the accuracy and efficiency of financial analytics.

XGBoost, Random Forest, and Logistic Regression remain the most popular algorithms for credit default prediction. Logistic regression estimates binary outcomes, but Random Forest and XGBoost offer more advanced methods. Random Forest builds multiple decision trees for enhanced accuracy, while XGBoost excels with large datasets through its gradient-boosting technique (Noriega et al., 2023). Comparing these models and employing feature selection are essential for improving predictive accuracy and identifying key predictors in credit risk management

Research Questions:

1. How do machine learning algorithms such as Logistic Regression, Random Forest, and XGBoost compare in terms of predictive accuracy for credit card default prediction?
2. What is the effect of key feature selection on the performance of machine learning models?

Objectives are: To compare the predictive performance of XGBoost, Random Forest, and Logistic Regression in credit default prediction.

To assess the impact of key feature selection on each model’s performance.

To evaluate the significance of key features in predicting defaults.
To identify the most accurate machine learning algorithm for credit default prediction.

PROJECT PLAN

Task List and Timeline

Task	Description	Start Date	End Date
Research and Literature Review	Review existing literature on ML algorithms for credit default prediction	24-10-2024	31-10-2024
Data Collection and Preparation	Acquire the relevant dataset, clean, preprocess, and perform feature selection for credit default prediction.	01-11-2024	06-11-2024
Algorithm Selection and Model Design	Finalize selection of machine learning algorithms Models and design.	07-11-2024	12-11-2024
Model Training and Implementation	Train selected algorithms on the preprocessed dataset, ensuring optimal performance of each model	13-11-2024	17-11-2024
Model Evaluation and Performance Testing	Evaluate the models based on metrics like accuracy, precision, recall, and F1-score, and analyze the effect of feature selection on performance	18-11-2024	20-11-2024
Results Analysis and Algorithm Comparison	Compare the Models performance to determine the most accurate model for credit default prediction	22-11-2024	03-12-2024
Key Feature Identification	Identify the key features driving credit default prediction as revealed by the MLM.	04-12-2024	08-12-2024
Report Writing	Detailed report, including research question, methodology, findings, and conclusions	09-12-2024	14-12-2024
Final Review, Editing and Submission	Review, Submit the final report, with all key insights and conclusions.	03-01-2025	06-01-2025

DATA MANAGEMENT PLAN

Dataset Link: Kaggle , Credit Risk Analytics

Datasets Overview: This project uses two datasets: one from Kaggle, focusing on credit card defaulters in Taiwan, commonly used for credit risk evaluation, and the other from Credit Risk Analytics, offering comprehensive credit risk data. Both datasets are anonymized, ethically compliant, and provide valuable insights for machine learning algorithms to predict credit defaults effectively

Data Collection

This project uses two anonymized datasets: the Kaggle Dataset, with 30,000 records on credit limits, balances, payments, and default status, and the Credit Risk Analytics Dataset, offering over 100,000 records of real financial data for credit risk research. Both are freely accessible under educational licenses, ensuring privacy and ethics compliance

Metadata

Both datasets can be used with machine learning models because they are in CSV format. The code files add less than 1MB to the total size of the datasets, which is about 10MB.

Document Control

The project's code will be tracked via a GitHub repository. All files will have the same naming convention, CreditDefaultPrediction_22020699_2024-10-24, and weekly commits are used to guarantee progress.

There will be a **README file** with an overview of the project, datasets, machine learning models, and important conclusions.

Safety and Storage

Every week, GitHub and Google Drive will back-up the code and data, with OneDrive offering extra redundancy. Sharing with stakeholders will be accomplished through secure links. Data protection is ensured by weekly backups.

Ethical Conditions

GDPR Compliance: The anonymized datasets comply with GDPR, protecting personal data.

UH Ethical Policy: The publicly available, anonymized datasets meet the University of Hertfordshire's ethical standards.
Permission to Use: Both datasets have no restrictions for research use.

Ethical Collection: The datasets were ethically sourced from trusted institutions, adhering to financial research standards.

REFERENCE LIST

- Bahnsen, A.C., Aouada, D. & Ottersten, B. (2015). Example-dependent cost-sensitive logistic regression for credit scoring. *Expert Systems with Applications*, 42(6), 2453-2466.
- Consoli, S., et al. (2021). *Data Science for Economics and Finance: Methodologies and Applications*. Cham: Springer.
- Lessmann, S., et al. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030
- Ngai, E.W.T., et al. (2010). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006
- Noriega, J.P., Rivera, L.A. & Herrera, J.A. (2023). Machine Learning for Credit Risk Prediction: A Systematic Literature Review. *Data*, 8(11), p.16https://doi.org/10.3390/data811016