7PAM2002-0901-2024 Data Science Project

Assignment: PDM Plan

Student Name: VASANTH KUMAR MEKALA

Student ID: 22058159

Course: MSC Data Science with Advanced Research

Module: Data Science Project

Submission Date: 18 Oct 2024

**Project Overview**

**Project Title:**  
Evaluation of financial risk for loan approval using machine learning

**Summary of Project and Background**

Financial risk assessment is important in the loan approval process as it helps banking institutes and ledgers avoid granting loans to borrowers who are likely to default. The implication of technological solutions such as Machine Learning (ML)-based predictive modeling systems allows banking institutes as well as ledgers to evaluate the demographic attributes (age, occupation, marital status, family dependence) as well as financial attributes (such as income, credit score, and loan amounts) for estimating credit risk scores for applicants (Noriega, Rivera and Herrera, 2023). Thus, identifying borrowers who are more likely to repay their debts on time is possible, and can help banking institutes to generate more interest income from borrowers.

**Building on previous studies:**

* With the emergence of external factors like fluctuation in interest rates and an increase in inflation rates, risks in the loan-providing sector have increased. Based on a report published by Statista, the level of uncertainty in loan repaying has increased by 10% in 2024 across the world (Statista, 2024). In fact, the average credit risk scores of borrowers have risen from 5.94 in 2023 to 6.36 in 2024.
* Reflecting the severity of the issue of increasing credit risks within the financial service sector (Statista, 2024). In this context, it becomes important for banking organizations or ledgers to estimate credit risk scores before approving a loan to reduce overall credit risks and enhance the revenue generated from interest.

This study aims to develop Machine Learning models to predict credit risk scores of borrowers and classify loan approval status.

**Research Question:**  
What is the importance of machine learning in the prediction of financial risks within the loan approval process?

**Objectives:**

1. **Data Collection and Preprocessing:**
   * Firstly, the dataset is going to be downloaded into CSV (Comma Separated Value) format from Kaggle, after which the dataset is going to be loaded into the Python environment in the Jupyter Notebook platform. Jupyter Notebook IDE can be considered as the software platform in this research for writing Python scripts for data preprocessing, exploratory data analysis, and ML models. After loading the dataset into Python, preliminary data exploration such as exploration of shape and info of the dataset is going to be performed to evaluate the data types, valid observations, and the number of columns in the dataset.
   * After, preliminary data exploration, preprocessing steps like identification and treatment of missing values, and duplicate observations are going to be performed. This can help in minimizing data errors from the dataset, which can lead to effective training of ML models. After data cleaning, feature engineering (selecting appropriate colons) and data transformation steps like Normalisation of skewed numerical columns and encoding for categorical columns are going to be performed. Based on the transformed data, firstly data splitting (80-20% or 70-30%) will be performed and then ML models will be developed.
   * The ML models are going to be segregated into two parts, Regression and Classification, where Regression models will be developed for predicting continuous risk scores associated with the likelihood of loan default or financial instability. The target variable ‘RiskScore' is continuous, thus, the Regression model is going to be appropriate for predicting credit risk. On the other hand, Binary Classification models can also be developed for determining the binary outcome of loan approval, which can be helpful in predicting whether an applicant is likely to be approved or denied for a loan. The target variable for this is going to be the 'Loan Approved’, which has two distinct categories (0: No, 1: Yes). Finally, the predictive accuracy of the models (for both Regression models and Classification models) will be performed to identify the optimal models
2. **Model Development:**

* To develop Regression models (state-of-the-art Machine Learning models such as Random Forest regressor, Support Vector Regression, and Neural Networks) for predicting continuous risk scores associated with each individual’s likelihood of loan default or financial instability.
* To develop classification models for predicting the loan approval status of applicants, indicating whether an applicant is likely to be approved or denied a loan.

1. **Model Evaluation:**

* To evaluate the demographic and financial factors influencing credit risks in the context of loan approval process

1. **Visualization and Interpretation:**
   * Visualize key findings (e.g., actual vs. changing values, feature importance).
   * Interpret results to identify significant factors and evaluate demographic and financial attributes.
2. **Conclusion and Recommendations:**

* To recommend effective strategies for financial institutes to minimize credit risks

**References:**

* Kaggle (2024). *Financial Risk for Loan Approval*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval?select=Loan.csv [Accessed 4 Oct. 2024].
* Noriega, J.R., Rivera, L. and Herrera, A. (2023). Machine Learning for Credit Risk Prediction: a Systematic Literature Review. *Data*, [online] 8(11), pp.169–169. Doi: https://doi.org/10.3390/data8110169.
* Statista (2024). *Main risks in financial services 2024 | Statista*. [online] Statista. Available at: https://www.statista.com/statistics/1441154/top-risk-financial-services-industry/ [Accessed 4 Oct. 2024].

**Project Timeline**

| **Task** | **Start Date** | **End Date** | **Duration** |
| --- | --- | --- | --- |
| Literature Review & Project Selection | 26th Sept 2024 | 7th Oct 2024 | 1.5 weeks |
| Data Collection | 08th Oct 2024 | 22nd Oct 2024 | 2 weeks |
| Data Preprocessing | 23rd Oct 2024 | 30th Oct 2024 | 1 week |
| Model Implementation | 31st Nov 2024 | 14th Nov 2024 | 2 weeks |
| Application of ML Techniques | 15th Nov 2024 | 25th Nov 2024 | 1.5 weeks |
| Model Evaluation | 36th Nov 2024 | 3rd Dec 2024 | 1 week |
| Refinement & Optimization | 4th Dec 2024 | 6th Dec 2024 | 3 days |
| Interim Assessment Preparation | 7th Dec 2024 | 9th Dec 2024 | 2 days |
| Final Model Testing | 9th Dec 2024 | 10th Dec 2024 | 2 days |
| Report Writing | 11th Dec 2024 | 15th Dec 2024 | 5 days |
| Final Assessment Preparation | 3rd Jan 2025 | 5th Jan 2025 | 2 days |
| Submission | 6th Jan 2025 | 6th Jan 2025 | 1 day |

**Data Management Plan:**

**Data Overview and Meta Data:**

The research is going to be based on secondary quantitative data; thus, secondary quantitative data can be collected from reliable databases like Kaggle. Kaggle is a reliable database for secondary quantitative data, appropriate for developing Machine Learning models and Statistical analysis, thereby, Kaggle has been considered as the main database in this study. The ‘Financial Risk for Loan Approval’ dataset has been planned to be considered in this research as the dataset contains both demographic attributes (such as age, marital status, home ownership, and family dependence) as well as a diverse range of financial attributes like Income, loan purpose, credit amount, previous loan history, account balance, net worth, total assets, and total liabilities. The inclusion of these financial factors can be crucial in training the ML models effectively, which can help in accurately predicting the credit risk of borrowers.

**Dataset Source:** <https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval?select=Loan.csv>

**Data Collection:**

**Source**: Downloaded from Kaggle (*evaluation of financial risk for loan approval processes*).

**Format**: CSV file containing demographic and financial attributes.

**Document Control:**

**Code Storage & Version Control**

**GitHub**: Code will be versioned through a GitHub repository with frequent, clearly labeled commits.

**Data Storage:**

**Local Storage**: Organized in directories (e.g., data/, code/, outputs/).

**Format**: CSV or Excel files.

**Data & Code Backup:**

**Backup Methods**: Code will be regularly backed up on GitHub, with dataset files stored on OneDrive.

**README FILE**

**Contents**: The ReadMe file will outline the project, setup steps, dataset description, and usage instructions, including licensing and contributors.

**GitHub Repository**: https://github.com/Vasi0796/Vasanth-project.git

**Data Ethics:**

* **GDPR Compliance:**  
  The dataset is anonymized, containing no personal or sensitive information, so GDPR requirements are not applicable.
* **University of Hertfordshire Ethics Compliance:**  
  This project complies with UH’s ethical standards, as it does not involve personal or sensitive data.
* **Data Usage Permission:**  
  The dataset is publicly available on Kaggle, explicitly approved for research purposes.
* **Ethical Data Collection:**  
  The data is hosted on Kaggle, a reputable platform, that ensures ethical standards in data gathering.