7PAM2002-0509-2024 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: 16 July 2025

**Project Overview**

**Project Title:**  
Early warning system for bankrupt prediction by using machine learning models.

**Summary of Project and Background**

* Bankruptcy is a financial event that can have harsh consequences for stakeholders, including creditors, regulatory institutions, employees, and investors.
* In such circumstances, bankruptcy prediction plays a significant role in supporting economic growth and maintaining financial stability.

**Building on previous studies:**

* Altman Z-score: Combines five financial ratios to estimate bankruptcy risk.
* Asif et al. (2024): Applied Altman model on NSE-listed firms.
* Zhu (2024): Advocated ML in advancing financial decision-making.

These works establish the limitations of linear models and support ML's superiority in analyzing nonlinear relationships for bankruptcy prediction.

**Research Question:**  
Can Machine learning models accurately predict company bankruptcy based on corporate governance indicators and the financial ratios?

**AIM:**

The **Aim of bankruptcy prediction** is to **identify companies that are at risk of financial failure** before it happens, using financial and operational data. This enables early intervention by stakeholders to mitigate losses and take corrective actions.

**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 (75-25% or 70-30%) will be performed and then ML models will be developed.
2. **Model Development:**

* **Algorithms: Random Forest, XGBoost, Neural Networks**
* **Feature Engineering: Ratio transformations, interaction terms, PCA**

1. **Model Evaluation:**

* **Metrics:** Accuracy, Precision, Recall, F1-score, ROC-AUC, Confusion Matrix
* **Validation:** K-fold cross-validation, test set evaluation
* **Comparison:** Traditional vs. ML model performance

1. **Visualization and Interpretation:**

Visual tools will be used to interpret results and model behavior:

* Feature importance plots
* Confusion matrix
* ROC curves
* SHAP (SHapley Additive exPlanations) values for model interpretability

These insights will assist stakeholders in understanding which features are most predictive of bankruptcy.

1. **Conclusion and Recommendations:**

The project concludes with:

* Summary of key findings
* Evaluation of ML models’ effectiveness in bankruptcy prediction
* Practical recommendations for financial institutions to implement ML-based early warning systems
* Suggestions for continuous monitoring and improvement of models

1. **References:**

* Asif, M., Saxena, A., Bhardwaj, S., Tiwari, S. and Chaturvedi, S. (2024) ‘A STUDY OF ALTMAN Z-SCORE BANKRUPTCY MODEL FOR ASSESSING BANKRUPTCY RISK OF NATIONAL STOCK EXCHANGE-LISTED COMPANIES’, *Proceedings on Engineering Sciences*, 6(2), pp. 789–806. (Available at: https://doi.org/10.24874/PES06.02A.006.)
* Di Quirico, R. (2023) ‘Uncertainty, Anxiety and the Post-Pandemic Economic Environment’, *Clinical Neuropsychiatry*, 20(4), p. 227. (Available at: https://doi.org/10.36131/cnfioritieditore20230401.)
* Statista (2025) *Largest U.S. bankruptcies as of June 2019*, *Statista*. (Available at: https://www.statista.com/statistics/1096794/largest-bankruptcies-usa-by-assets/) [Accessed: 29 May 2025].
* Tierney, A. (2024) *Number of business bankruptcies U.S. 2000-2020*, *Statista*. (Available at: https://www.statista.com/statistics/817918/number-of-business-bankruptcies-in-the-united-states/) [Accessed: 29 May 2025].

**Project Timeline**

|  |  |  |  |
| --- | --- | --- | --- |
| **Task** | **Start Date** | **End Date** | **Duration** |
| **Project Selection** | **14th May 2025** | **21st May 2025** | **1 week** |
| **Data Collection** | **22nd May 2025** | **5th Jun 2025** | **2 weeks** |
| **Data Preprocessing** | **6th Jun 2025** | **13th Jun 2025** | **1 week** |
| **Model Implementation** | **14th Jun 2025** | **28th Jun 2025** | **2 weeks** |
| **Application of ML Techniques** | **29th Jun 2025** | **13th July 2025** | **2 weeks** |
| **Model Evaluation** | **14th July 2025** | **21st July 2025** | **1 week** |
| **Refinement & Optimization** | **22nd July 2025** | **25th July 2025** | **3 days** |
| **Interim Assessment Preparation** | **26th July 2025** | **28th July 2025** | **2 days** |
| **Final Model Testing** | **29th July 2025** | **31st July 2025** | **2 days** |
| **Report Writing** | **1st Aug 2025** | **14th Aug 2025** | **1.5 weeks** |
| **Final Assessment Preparation** | **15th Aug 2025** | **17th Aug 2025** | **2 days** |
| **Submission** | **18th Aug 2025** | **18th Aug 2025** | **1 day** |

**Data Management Plan:**

**Data Overview and Meta Data:**

The project uses structured financial and corporate governance data to build and evaluate machine learning models for predicting bankruptcy. The data primarily includes quantitative metrics from company financial statements, along with qualitative governance indicators.

**Dataset Source:** https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction.

**Data Collection:**

* **Source**: Downloaded from Kaggle (*evaluation of financial risk for loan approval processes*).
* **Format**: CSV file containing financial attributes.

**Document Control:**

* + **Source: Downloaded from Kaggle (evaluation of financial risk for loan approval processes).**
  + **Format: CSV file containing financial attributes.**

**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-s-project-2025.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.