Essentials of Data Science

Iteration 2: AI-Driven Financial Risk Management System

1. Project Kickoff

The AI-Driven Financial Risk Management System aims to leverage machine learning techniques to identify and predict potential financial risks such as credit defaults, loan irregularities, and market volatility. The goal of this project is to build a robust end-to-end system that integrates data management and predictive modeling components to provide actionable insights. By combining machine learning algorithms like Logistic Regression and Random Forest, the system will offer data-driven risk assessment and enhance decision-making accuracy.

The expected outcomes include the successful implementation of machine learning models capable of predicting credit risk with high accuracy that presents real-time risk indicators. The project also aims to evaluate and compare model performance using metrics such as precision, recall, F1-score, and ROC-AUC, ensuring both interpretability and reliability in predictions.

The scope of the project is clearly defined around three core components: data preprocessing, model development, and model validation. In the data preprocessing stage, we will clean, normalize, and prepare the dataset for modeling by handling missing values, encoding categorical variables, and performing feature scaling. The modeling stage will focus on training and tuning classification algorithms to predict financial risk and model validation will help us validate that our model can accurately predict the outcome. Real-time data ingestion, production-level deployment, and integration with live banking systems are excluded from this iteration but are identified as potential areas for future work once the system is validated.

The key deliverables of the project include a finalized dataset and preprocessing pipeline and trained and validated machine learning models. In the initial phase, we will define objectives, set up the dataset, and establish the project environment. The second phase will involve model training, evaluation, and refinement. The third phase will focus on optimizing the model and integrating all components, performing testing, and preparing comprehensive project documentation for submission.

In terms of team capabilities, we collectively possess strong skills in Python programming, data preprocessing, exploratory data analysis, and machine learning model development. These capabilities align closely with the objectives of the project.

For dataset selection, the team has chosen the publicly available Default of Credit Card

Clients Dataset from Kaggle. This dataset includes demographic and financial information from 30,000 clients, covering attributes such as credit limit, payment history, bill amounts, and default status. It is particularly suited for binary classification problems like credit default prediction and aligns perfectly with the goals of this project.

For this project, we would use the Default of Credit Card Clients Dataset available on Kaggle: https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

2. Team Discussions

The project team consists of three members, each bringing a distinct set of skills and expertise that collectively support the development of our project.

Afrah is responsible for exploratory data analysis (EDA) and initial data preparation. She will focus on understanding the dataset's structure, identifying missing values, detecting outliers, and performing statistical analysis to uncover patterns that influence financial risk. She will use Python, Pandas, and visualization tools such as Matplotlib and Seaborn to conduct thorough data profiling and feature analysis. Afrah's work will provide the foundation for model development by ensuring that the dataset is clean, balanced, and rich with meaningful features. Through EDA, she will also generate visual insights that guide variable selection and model input decisions.

Ishwarya will lead the model building and evaluation phase of the project. She will implement classification algorithms such as Logistic Regression and Random Forest using Scikit-learn. His responsibilities include training baseline models, analyzing performance metrics such as accuracy, precision, recall, and ROC-AUC, and validating the models through cross-validation.

Lakshmi will focus on hyperparameter tuning and model optimization to enhance performance and generalization. This involves applying techniques such as GridSearchCV and RandomizedSearchCV to fine-tune model parameters, comparing multiple configurations, and selecting the best-performing model. Additionally, she will monitor for overfitting and ensure that models maintain robust performance across validation datasets.

Based on the project's requirements and the team's existing skill set, we will use Python as the primary programming language due to its rich ecosystem of libraries for data science and ML. Development will take place primarily in VS Code and Jupyter Notebook, offering flexibility for both experimentation and production-ready code. Version control will be managed entirely through GitHub, allowing smooth collaboration, code reviews, and centralized storage of all project materials.

In conclusion, the team's collective strengths in data analysis and machine learning will provide a balanced foundation for achieving the project's objectives. Each member has been assigned a well-defined role aligned, ensuring that all critical components.

3. Skills and Tools Assessment

In terms of external resources, we plan to rely on open-source documentation, academic tutorials, and online learning materials to strengthen areas where experience is limited. Additionally, online communities such as Kaggle and Stack Overflow will be valuable for troubleshooting and optimizing model performance. If needed, the team will consult with teaching assistants for help.

The selected tools and libraries align closely with the project's scope and technical objectives. Python serves as the foundation of the system due to its versatility and strong ecosystem for data science and AI development. Pandas and NumPy will be used for data preprocessing and manipulation, while Scikit-learn will handle model development, training, and evaluation.

4. Initial Setup

We chose Python 3.10 as the primary programming language, owing to its versatility and strong ecosystem for machine learning and data analysis. We are using Jupyter Notebook for exploratory data analysis and model prototyping, and for structured code development and integration.

A virtual environment has been created to maintain consistent dependencies across all team members' systems. The environment includes essential Python libraries such as Pandas, NumPy, Scikit-learn, Matplotlib and Seaborn all of which are listed in a shared requirements.txt file for easy installation. This setup ensures that the project can be replicated on any system with minimal configuration issues.

To maintain consistency, the team has also established standard naming conventions for commits and files, promoting clarity and traceability throughout the project's lifecycle. Each update to the repository is accompanied by concise commit messages that describe the specific changes made. Regular synchronization sessions are held to merge updates and review progress collectively, ensuring that all members remain aligned with current versions of scripts and documentation.

In summary, the initial setup provides a stable and collaborative environment for us to work on our project. The use of Python, virtual environments, and GitHub ensures scalability, reproducibility, and transparency across all project activities.

5. Submission for This Iteration

This submission includes a three-page report summarizing the project scope, objectives, team members' contributions, tools, and programming languages used for the **AI-Driven Financial Risk Management System**.

Progress Tracking

Project progress has been documented in the Excel tracker file , which is also included in the GitHub repository. Additionally, the live project progress tracker is available on Google Sheets at:

`https://docs.google.com/spreadsheets/d/1PE0hJIzdOBSCcUmjf7VMsfMP8Fpli6ELRBo-nfbrJ2w/edit`

Verification

This submission satisfies all project requirements outlined for Iteration 2 and is ready for stakeholder review.

GitHub Repository

All project materials, including this PDF report, and the Excel progress tracker, are available in the following GitHub repository:

`https://github.com/iishwarya1415/DS5110-Project`