High Level Design

Credit Risk Predictor

Revision Number: 1.0
Last date of revision: 1/8/23

Pothuri Harish Varma

Document Version Control

Date	Version	Description	Author
01-08-2023	V1.0	HLD-V1.0	Pothuri Harish Varma

Contents

1. Introduction

1.1. Why this High Level Design Document?
1.2. Scope
1.3. Definitions
1.4. Overview

2. General Description

2.1. Product Perspective
2.2. Problem Statement
2.3. Proposed Solution
2.4. Technical Requirements
2.5. Data Requirements
2.6. Tools Used
2.7. Constraints
2.8. Assumptions

3. Design Details

3.1. Process Flow
3.2. Event Log

4. Performance

4.1. Reusability
4.2. Application Compatibility
4.3. Deployment

5. Conclusion

1. Introduction

1.1. Why this High Level Design Document?

The purpose of this High Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the model interacts at a high level.

1.2. Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

1.3. Definitions

- IDE – Integrated Development Environment
- EDA – Exploratory Data Analysis
- API – Application programming interface
- KPI – Key Performance Indicator
- VS – Visual Studio
- AWS – Amazon web services
- ML– Machine Learning
- RAM– Random Access Memory

1.4. Overview

The HLD will:

- present all of the design aspects and define them in detail
- describe the user interface being implemented
- describe the hardware and software interfaces
- describe the performance requirements
- include design features and the architecture of the project
- list and describe the non-functional attributes like:

2. General Description

2.1. Product Perspective

The product perspective of predicting credit risk using machine learning presents a valuable tool for financial institutions to assess loan applicants' creditworthiness more effectively. By accurately predicting credit risk, banks and lending institutions can minimize financial losses, reduce default rates, and improve overall profitability. The machine learning algorithms analyze historical data to identify patterns and trends, aiding lenders in making informed decisions about loan approvals and interest rates.

This product has the potential to cater to a wide range of financial institutions, from small community banks to large multinational corporations. It can be integrated into existing loan approval processes or new lending platforms to enhance efficiency.

However, a key challenge lies in ensuring the accuracy and reliability of the machine learning model. This requires continuous data analysis and model refinement to remain up-to-date and effective. Transparent explanations of the model's workings are essential to build trust among users and address any legal or ethical concerns.

The credit risk prediction product powered by machine learning offers financial institutions a more accurate and efficient way to assess creditworthiness and mitigate financial risks, leading to improved decision-making and overall financial stability.

2.2. Problem Statement

Financial institutions heavily rely on credit loans as a significant source of revenue. However, the risk of non-performing credit loans poses a substantial challenge, leading to potential financial losses. To address this issue, the marketing bank aims to minimize credit risk by studying patterns from existing lending data.

The goal of this project is to build a predictive model using data mining techniques that can accurately classify individuals as either good credit risks (labeled as 1) or bad

credit risks (labeled as 0) based on their attributes in the dataset. By leveraging data mining, hidden information can be extracted from large datasets, facilitating effective classification and pattern recognition.

2.3. Proposed Solution

The solution of the this problem statement is to perform EDA on the dataset to generate meaningful insights from the data and use this data to hyper tune with appropriate machine learning algorithms which will have the maximum accuracy in predicting the credit risk. Thus creating a user interface where a user can put in the various features of the data which will in return give the credit risk is present or not.

2.4. Technical Requirements

The solution can be a cloud-based or application hosted on an internal server or even be hosted on a local machine. For accessing this application below are the minimum requirements:

- Good internet connection.
- Web Browser.

For training model, the system requirements are as follows:

- +4 GB RAM preferred
- Operation System: Windows, Linux, Mac
- Visual Studio Code / Jupyter notebook/ Pycharm

2.5. Data Requirements

Data requirements completely depend on our problem statement.

- Comma separated values (CSV) file.
- Input file feature/field names and its sequence should be followed as per decided.

2.6. Tools Used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn and Flask are used to build the whole model.

- Pandas is an open-source Python package that is widely used for data analysis and machine learning tasks
- NumPy is the most commonly used package for scientific computing in Python.
- Matplotlib and Seaborn are an open-source data visualization library used to create interactive and quality charts/graphs.
- Scikit-learn is used for machine learning.
- Flask is used to build an API.
- VS Code is used as IDE (Integrated Development Environment)
- GitHub is used as a version control system.
- Front end development is done using HTML.
- AWS,Docker is used for deployment of the model.
- GitHub Actions is used to integrate and deploy code changes to a third-party cloud application

2.7. Constraints

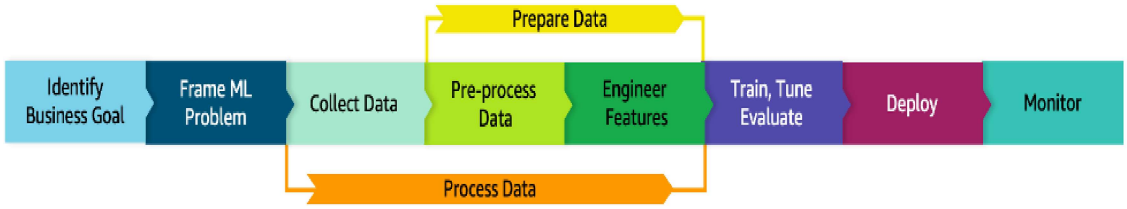
This model must be user friendly, as automated as possible and users should not be required to know any of the workings.

2.8. Assumptions

The main objective of the project is to develop an API to predict the bank credit risk using South German credit data. Machine Learning based classification models are used for predicting above mentioned cases on the input data.

3. Design Details

3.1. Process Flow



3.2. Event Log

The system should log every event so that the user will know what process is running internally.System should not hang out even after using so many loggings.

Initial Step-By-Step Description:

- The system identifies at what step logging required
- The system should be able to log each and every system flow.
- Developers can choose logging methods. You can choose database logging.

4. Performance

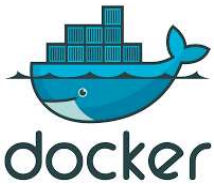
4.1. Reusability

The entire solution will be done in modular fashion and will be API oriented. So, in the case of the scaling the application, the components are completely reusable.

4.2. Application Compatibility

The interaction with the application is done through the designed user interface, which the end user can access through any web browser.

4.3. Deployment



5. Conclusion

In conclusion, the utilization of machine learning algorithms to predict credit risk in banks has yielded promising results. By analyzing a diverse array of factors, including credit history, income, employment status, and more, these models have demonstrated their ability to accurately forecast the likelihood of borrower default.

Throughout this project, we showcased the effectiveness of XGBoost and Random Forest algorithms in predicting credit risk with high precision. The feature importance analysis has also yielded valuable insights into the key factors influencing credit risk.

The implementation of this project offers immense benefits to banks and financial institutions. It equips them with data-driven decision-making tools for lending and risk management, enabling them to make informed choices while assessing loan applicants. Additionally, this approach contributes to reducing the risk of default and enhancing the overall financial stability of the institution.

However, it is essential to acknowledge that the accuracy of these models can be influenced by various factors, such as data quality, quantity, feature engineering, and model selection. Continuous monitoring and refinement of the models are crucial to ensure their ongoing effectiveness and reliability in predicting credit risk.

As the financial landscape evolves, integrating advanced machine learning techniques into credit risk assessment becomes increasingly crucial. By staying abreast of technological advancements and conducting regular evaluations, banks can continue leveraging the power of machine learning to make sound credit decisions and mitigate financial risks effectively. Ultimately, this empowers institutions to forge stronger relationships with customers and maintain a healthy lending ecosystem.