High Level Design (HLD)

Predict Credit Risk Using South German Bank Data

Revision Number : 1.0
Last Date of Revision : 31/08/2024

Document Version Control

Date Issued	Version	Description	Author
31/08/2024	0.0.1	Document Created	Kunal Shelke

Contents

Abstract

The "Predict Credit Risk Using South German Bank Data" project aims to develop a predictive model that assesses the credit risk of potential bank customers based on various financial and demographic factors. The project leverages machine learning algorithms to predict whether a customer poses a high or low credit risk. The model is trained on historical data, and the project includes data preprocessing, feature engineering, model training, evaluation, and deployment. The ultimate goal is to assist the bank in making informed lending decisions to minimize financial risk

1 Introduction

1.1. Why this High-Level Design Document?

This High-Level Design (HLD) document serves as a blueprint for the development, deployment, and management of the "Predict Credit Risk Using South German Bank Data" project. It outlines the architecture, data flow, components, and processes involved in building and deploying the machine learning model. This document ensures that all stakeholders have a clear understanding of the project's design and facilitates seamless collaboration and implementation.

The HLD will:

- **Present all of the design aspects and define them in detail**
- **Describe the performance requirements**
- **Include design features and the architecture of the project**
- **List and describe the non-functional attributes like:**
  - ➔ **Reliability**
  - ➔ **Maintainability**
  - ➔ **Portability**
  - ➔ **Reusability**
  - ➔ **Application compatibility**
  - ➔ **Resource utilization**
  - ➔ **Serviceability**

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

Include any specific terminology or acronyms used in the project.

Term	Definitions
DB	DataBase
AWS	Amazon Web Service
IDE	Integrated Development Environment
MD	Deutsche Mark (Past Currency of Germany)

2 General Description

2.1 Product Perspective

The "Predict Credit Risk Using South German Bank Data" project is designed which can be used to integrate into the bank's decision-making process by providing a reliable tool to predict the credit risk of potential borrowers. This predictive model will be a part of the bank's broader strategy to minimize default rates and optimize lending practices.

2.2 Problem Statement

Normally, most of the bank's wealth is obtained from providing credit loans so that a marketing bank must be able to reduce the risk of non-performing credit loans. The risk of providing loans can be minimized by studying patterns from existing lending data. One technique that you can use to solve this problem is to use data mining techniques. Data mining makes it possible to find hidden information from large data sets by way of classification. The goal of this project is to build a model to predict whether the person, described by the attributes of the dataset, is a safe(1) or an unsafe/not safe(0) credit risk.

2.3 Proposed Solution

This project proposes the development of a machine learning model trained on South German Bank Credit data to predict the credit risk of new loan applicants. The model will analyze various financial and demographic factors to classify applicants as high or low risk, thus aiding the bank in making better lending decisions.

2.4 Further Improvements

Integration with real-time data feeds for continuous model training and updates. Expansion to include additional data sources, such as social media profiles or transaction histories, for improved accuracy.Development of an explainable AI component to provide transparency in model predictions.

2.5 Data Requirements

Data Requirement completely depends on our problem. For training and testing the model, we are using the South German Credit Dataset. Here are the features in dataset

Here is the detailed description for each column in the dataset:

1. laufkont (status) : Status of the debtor's checking account with the bank.
1: No checking account
2: ... < 0 DM
3: 0 <= ... < 200 DM
4: ... >= 200 DM / salary for at least 1 year

2. laufzeit (duration) : Duration of the credit in months. Numerical value representing the credit duration.

3. moral (credit_history) : History of the debtor's credit at the bank.
0: Delay in paying off in the past
1: Critical account/other credits elsewhere
2: No credits taken/all credits paid back duly
3: Existing credits paid back duly till now
4: All credits at this bank paid back duly

4. verw (purpose) : Purpose for which the credit is being requested.
0: Others

1: Car (new)
2: Car (used)
3: Furniture/equipment
4: Radio/television
5: Domestic appliances
6: Repairs
7: Education
8: Vacation
9: Retraining
10: Business

5. hoehe (amount) : Amount of the credit requested. Numerical value representing the credit amount in DM.

6. sparkont (savings) : Savings account/bonds of the debtor.
1: Unknown/no savings account
2: ... < 100 DM
3: 100 <= ... < 500 DM
4: 500 <= ... < 1000 DM
5: ... >= 1000 DM

7. beszeit (employment_duration) : Length of time the debtor has been employed at their current job.
1: Unemployed
2: < 1 year
3: 1 <= ... < 4 years
4: 4 <= ... < 7 years
5: >= 7 years

8. rate (installment_rate) : Installment rate as a percentage of disposable income.
1: >= 35%
2: 25 <= ... < 35%
3: 20 <= ... < 25%
4: < 20%

9. famges (personal_status_sex) : Personal status and sex of the debtor.
1: Male: divorced/separated
2: Female: non-single or Male: single

3: Male: married/widowed
4: Female: single

10. buerge (other_debtors) : Other debtors or guarantors for the credit.
1: None
2: Co-applicant
3: Guarantor

11. wohnzeit (present_residence) : Length of time the debtor has lived at their current residence.
1: < 1 year
2: 1 <= ... < 4 years
3: 4 <= ... < 7 years
4: >= 7 years

12. verm (property) : Type of property owned by the debtor.
1: Unknown/no property
2: Car or other property
3: Building society savings agreement/life insurance
4: Real estate

13. alter (age) : Age of the debtor in years.
Numerical value representing the age of the debtor.

14. weitkred (other_installment_plans) : Other installment plans held by the debtor.
1: Bank
2: Stores
3: None

15. wohn (housing) : Type of housing the debtor lives in.
1: For free
2: Rent
3: Own

16. bishkred (number_credits) : Number of credits the debtor has in this bank.
1: 1
2: 2-3
3: 4-5

4: >= 6

17. beruf (job) : Job type of the debtor.
    1: Unemployed/unskilled -non-resident
    2: Unskilled -resident
    3: Skilled employee/official
    4: Manager/self-employed/highly qualified employee

18. pers (people_liable) : Number of people who are financially dependent on the debtor.
    1: 3 or more
    2: 0 to 2

19. telef (telephone) : Whether the debtor has a telephone registered under their name.
    1: No
    2: Yes (under customer name)

20. gastarb (foreign_worker) : Whether the debtor is a foreign worker.
    1: Yes
    2: No

21. kredit (credit_risk) : The credit risk assigned to the debtor.
    0: Bad
    1: Good

2.6 Tools Used

List the software, libraries, and platforms used in the project.

- VS code is used as an IDE.
- For visualization of the plots, Matplotlib and Seaborn are used.
- AWS is used for deployment of the model.
- Front end development is done using HTML/CSS
- Python is used for backend development.

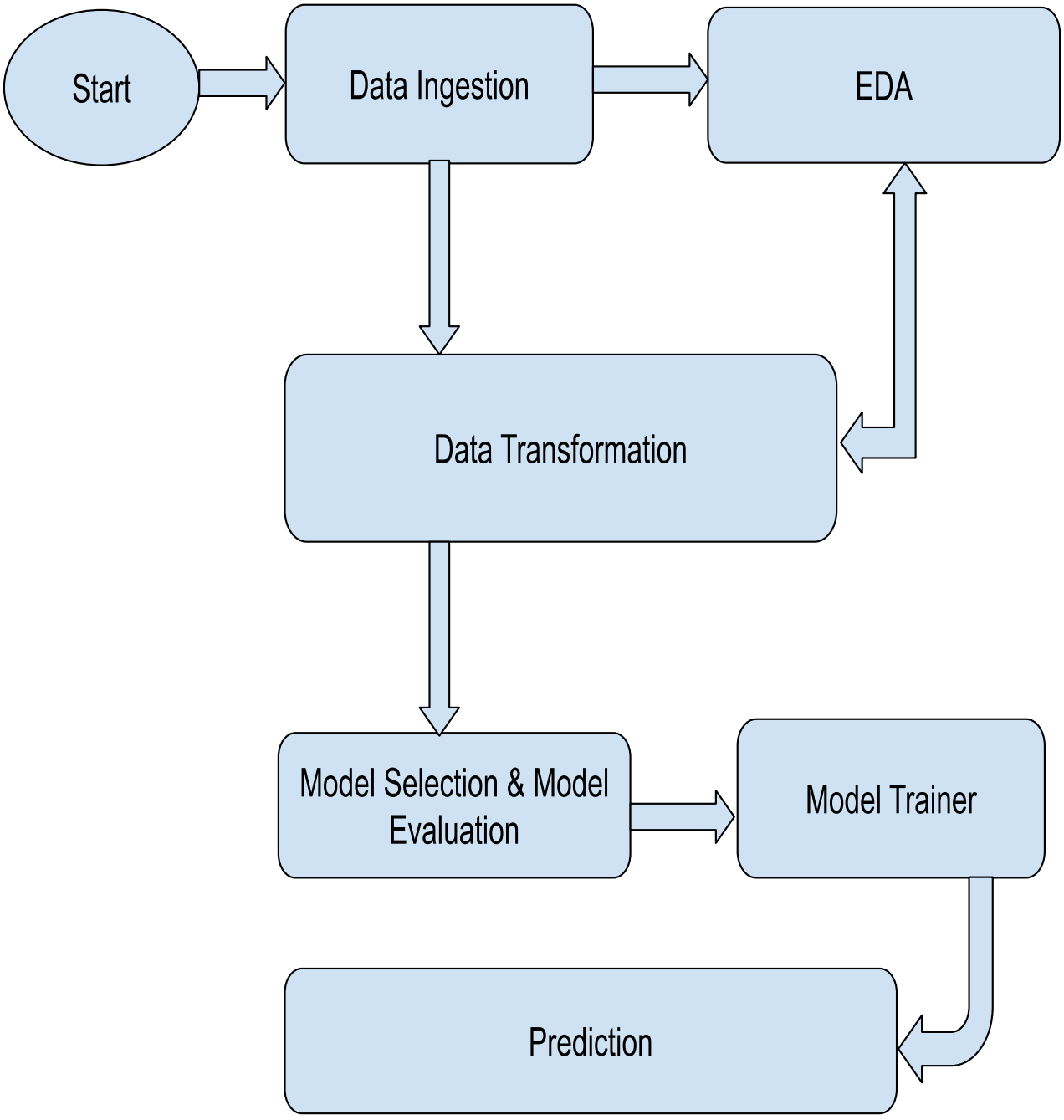
- DVC is use for Data Version Control.
- Model Tracking is done by MLflow/DagsHub.
- Docker is used for Containerisation .
- Jupyter notebook is used for EDA purpose.
- Flask is used for web frameworks.
- Cassandra is used as database .
- GitHub is used as a version control system.
- Github Actions is used as a ci/cd pipeline.

2.7 Constraint

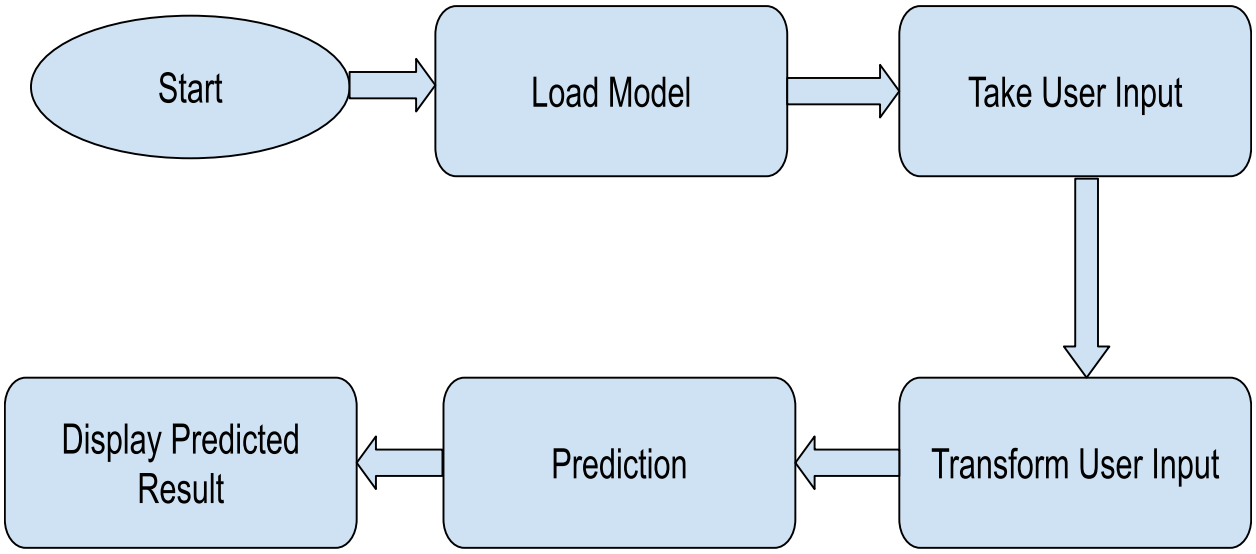
The prediction of credit risk using bank data system must be user friendly, errors free and users should not be required to know any of the back-end working

3 Design Details

3.1 Process Flow



3.2 Deployment Process



3.3 Event Log

The System should log every event so that the user will know what process is running internally. Internal Step-By-Step Description In this Project we defined logging for every function, class. By logging we can monitor every insertion, every flow of data in the database. By logging we are monitoring every step which may create problems or every step which is important in the file system. We have designed logging in such a way that the system should not hang even after so much logging, so that we can easily debug issues which may arise during process flow.

3.4 Error Handling

Should errors be encountered, an explanation will be displayed as to what went wrong? An error will be defined as anything that falls outside the normal and intended usage.

3.5 Performance

The use of South German bank credit data is used for the prediction of credit risk. So that it should be as accurate as possible. That's why before building this model we followed the complete process of Machine Learning.

3.6 Reusability

Modular Code: Components like data ingestion, transformation, and model training are modular and can be reused across different projects.

Dockerization: The use of Docker allows the entire setup to be easily replicated and reused in other environments.

3.7 Application Compatibility

Platform: Compatible with Linux, Windows, and macOS environments.

Browser: The web interface will be compatible with all modern web browsers.

3.8 Resource Utilization

When any task is performed, it will likely use all the processing power available until that function is finished.

- **CPU**: Multi-core processing to speed up data processing and model training.
- **Memory**: At least 8GB of RAM recommended for smooth operation during training.
- **Storage**: Approximately 2GB for data storage, model artifacts, and logs.

3.9 Deployment

I have done it on AWS , but the model can be deployed in any cloud services such as Microsoft Azure, Google, Heroku etc.

4 Conclusion

The "Predict Credit Risk Using South German Bank Data" project is designed to provide a robust, scalable, and accurate solution for assessing credit risk using machine learning. By deploying this solution, the bank can significantly improve its lending decisions, reduce default rates, and enhance overall financial performance. This High-Level Design document serves as a comprehensive guide for the development and deployment of the project, ensuring alignment with the bank's strategic goals.