HIGH LEVEL DESIGN DOCUMENT

BANK CREDIT RISK PREDICTION

DOCUMENT VERSION CONTROL

DATE ISSUED	VERSION	DESCRIPTION	AUTHOR
	V1.0	HLD-V1.0	ADHIRAJ SINGH SHEKHAWAT

ABSTRACT

The purpose of this project is to predict credit risk of bank South German credit data and machine learning techniques. The credit risk prediction is a critical task for banks as it helps to evaluate the creditworthiness of borrowers and mitigate the risk of default. The project aims to build a model that can accurately predict whether a loan applicant will default or not, based on a set of features such as 'status', 'duration', 'credit history, 'purpose', 'amount', 'savings', 'employment duration', 'installment rate', 'personal status sex', 'other debtors', 'present residence', 'property', 'age', 'other installment plans', 'housing', 'number credits', 'job', 'people liable', 'telephone', 'foreign worker', 'credit risk'.

In this project, we will explore various machine learning algorithms such as Xgboost, Random Forest, Bagging and Boosting to predict credit risk. The dataset used in this project contains historical loan data of bank loan applicants. The dataset includes both good and bad loans, making it suitable for training and testing the models. We will perform exploratory data analysis to gain insights into the data and preprocess the data to prepare it for the machine learning models.

We will evaluate the performance of the models using various evaluation metrics such as accuracy, precision, recall, and F1-score. We will also use ROC curve and AUC to measure the performance of the models. The best-performing model will be selected based on the evaluation metrics.

The results of this project can be used by banks and financial institutions to make informed decisions about whether to approve or reject a loan application. Accurate credit risk prediction can help banks to minimize their losses from loan defaults and improve their profitability.

1.0 Introduction

1.1 Why this High-Level Design Document?

The purpose of this High-Level document is to add necessary detailsto current project description to represent a suitable model for coding. This document is used as a reference manual for how the model interact at a high-level.

The HLD will be useful in

- Presents all 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 feature and the architecture of the project.

1.2 Scope

The HLD document presents the structure of the system, such as the database architecture, application architecture, and technology architecture. The HLD uses non-technical to middle-technical terms which should be understandable to the administrators of the system.

1.3 Definitions

TERM	DESCRIPTION
DATBASE	Collection of information
IDE	Integrated Development Environment
EDA	Exploratory Data Analysis
API	Application programming interface
KPI	Key Performance Indicator
VS	Visual Studio
AWS	Amazon web services
GCP	Google Cloud Platform
ML	Machine Leaning

2.0 General Description

2.1 Product Perspective

From a product perspective, predicting credit risk of bank project using machine learning can be seen as a valuable tool for financial institutions to better assess creditworthiness of loan applicants. The ability to accurately predict credit risk can help banks and other lending institutions minimize financial losses, reduce default rates, and improve overall profitability.

By using machine learning algorithms to analyze large amounts of historical data, the model can identify patterns and trends that may be indicative of high-risk loan applicants. This can help lenders make more informed decisions about who to approve for loans, and at what interest rates.

The product could potentially be marketed to a wide range of financial institutions, from small community banks to large multinational corporations. It could also be used to improve existing loan approval processes or integrated into new lending platforms.

One potential challenge from a product perspective is ensuring that the machine learning model is accurate and reliable. This will require ongoing data analysis and refinement of the model to ensure that it remains up-to-date and effective. It will also be important to provide clear and transparent explanations of how the model works and how it arrives at its predictions to build trust among users and avoid potential legal or ethical concerns. Overall, predicting credit risk using machine learning can be seen as a valuable product for financial institutions, offering a more accurate and efficient way to assess creditworthiness and reduce financial risks.

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, you have to build a model to predict whether the person, described by the attributes of the dataset, is a good (1) or a bad (0) credit risk

2.3 Proposed Solution

The solution of the this problem statement if 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 depends on our problem statement.

- Comma separated values (CSV) file.
- Input file feature/field names and its sequence should be followedas 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 fordata analysis and machine learning tasks.
- NumPy is most commonly used package for scientific computing in Python.
- Matplotlib and Seaborn are an open-source data visualization library used to createinteractive and quality charts/graphs.
- Scikit-learn is used for a machine learning.

- Flask is used to build an API.
- VS Code and Pycharm are used as IDE (Integrated Development Environment)
- GitHub is used as version control system.
- Front end development is done using HTML/CSS.
- Heroku, GCP, Azure, AWS are used for deployment of the model.

2. 7 Constraints

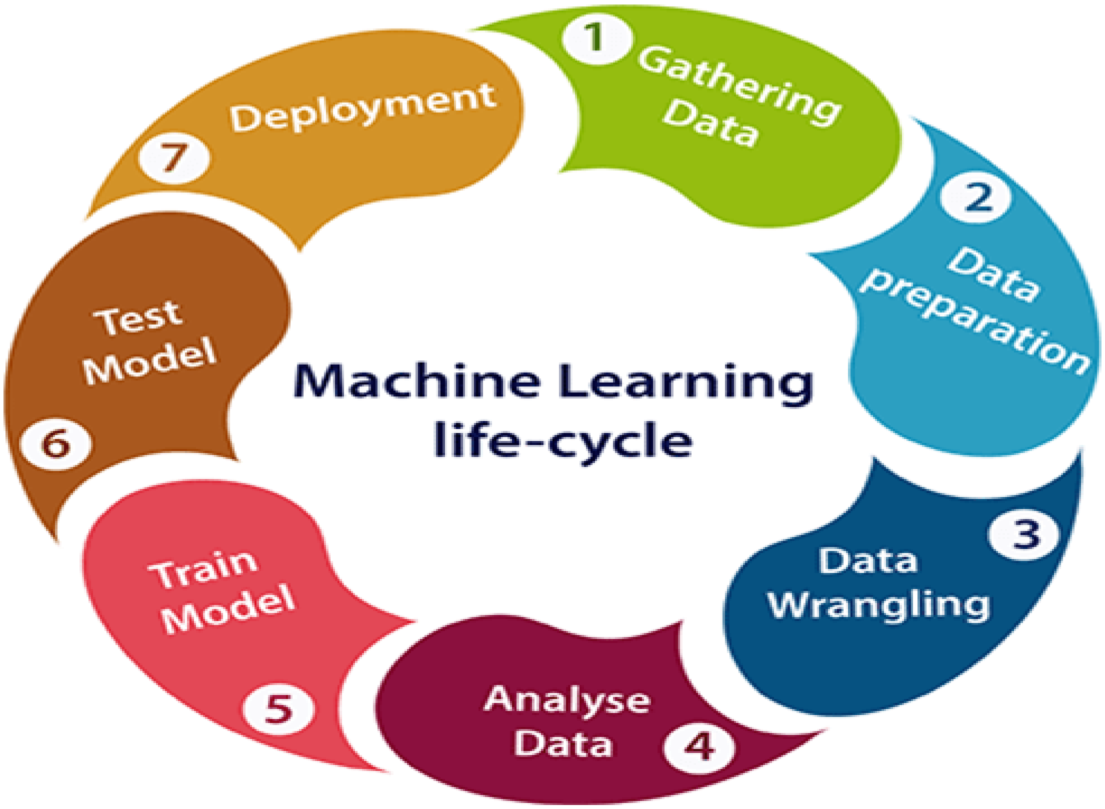
This model must be user friendly, as automated as possible and users should not be required to know any of the workings.

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. Machinelearning based classification models are used for predicting above mentioned cases on the input data.

3.0 Design Details

3.1Process Flow



3.2 Event Log

The system should log every event so that the user will know whatprocess is running internally.

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.
- Developer can choose logging method. You can choose database logging.

System should not hang out even after using so many loggings.

4.0 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 componentsare completely reusable.

4.2 Application Compatibility

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

4.3 Deployment

6.0 Conclusion

In conclusion, the use of machine learning algorithms to predict credit risk in banks has shown promising results. By analyzing various factors such as credit history, income, employment status, and others, these models can accurately predict the likelihood of a borrower defaulting on a loan.

This project has demonstrated the effectiveness of using XGBoost and Random Forest algorithms to predict credit risk with high accuracy. The feature importance analysis has also provided valuable insights into the most important factors affecting credit risk.

The implementation of this project can help banks and financial institutions in making informed decisions about lending and risk management. It can also aid in reducing the risk of default and improving the overall financial stability of the institution.

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