HIGH LEVEL DESIGN (HLD)

Credit Card Default Prediction

PRATHMESH JAGTAP
INeuron

Contents

Document Version Control

Date Issued	Version	Description	Author
30.11.2022	V1.0	Initial HLD – V1.0	PRATHMESH JAGTAP
10.01.2023	V1.1	HLD -V1.1	PRATHMESH JAGTAP

Abstract

Financial threats are displaying a trend about the credit risk of commercial banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats faces by commercial banks is the risk prediction of credit clients. The goal is to predict the probability of credit default based on credit card owner's characteristics and payment history.

For this, I have analyzed the “Default of Credit Card Clients” dataset released under the public license of Creative Commons to predict the default condition of credit card holder. Five classification models which are Logistic Regression, K-Nearest Neighbor, Naive Bayes Classifier, Support Vector Machines (LinearSVC), Decision Trees, Random Forest Classifier, Ensemble have been used to compare and contrast the performance of these algorithms.

The dataset was training dataset which is used for model training and that trained model helped to come up with some predictions. Then the predicted observations were compared with actual data to test and verify the accuracy of model. Later the accuracies and matrices of all these models were compared. After this comparison it was found that Random Forest and Gaussian Naïve Bayes algorithms performed better than the remaining models.

Lastly, Gaussian Naïve Bayes is best suited in this case because it gives best evaluation score compared to other models.

1.0 Introduction

1.1 Why this High-Level Design Document?

The Purpose of this High-Level document is to add necessary details to current project details to 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

- 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 databases 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*
DATABASE	Collection of all the information
IDE	Integrated Development Environment
AP	Application Programming Interface
KPI	Key Performance Indicator
VS CODE	Visual Studio Code
EDA	Exploratory Data Analysis

2.0 General Description

2.1 Product Perspective

The Credit Card Defaults Prediction system is a machine learning based predictive system which will help us to predict the defaults of the credit card holders. It is helps to find defaults situation for individuals and also for more and more holders.

2.2 Product Statement

To develop an API interface to predict the probability of credit default based on credit card owner's characteristics and payment history. Also analyzing the financial threads based on payment history, owner's behavior. To create API interface to predict the default, and to detect whether the status of owner affects the defaults of the credit card.

2.3 Proposed Solution

The solution proposed here is an estimating defaults of credit card based on owner's characteristics and payment history data; this can be implemented to perform above mentioned use cases. In first case, analyzing which are the most important features and how education and marital status affects the payment history and defaults situation of the credit card. In second case, if model detects that the owner defaults then bank takes action against them. And in the last use case, we will be making an interface to predict the defaults of credit card holder.

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 / Google Colab

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 followed as per decided.

2.6 Tool Used

Python programming language and framework such as NumPy, Pandas, Plotly, matplotlib, Flask, Scikit-learn, HTML/CSS, JavaScript 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 most commonly used package for scientific computing in python.
- **Plotly** and **matplotlib** is an open-source data visualization library used to create interactive and quality charts/graphs.
- **Scikit-learn** is used for a machine learning.
- **VS CODE** visual studio code is used as an IDE.
- **GitHub** is used as version control tool.
- **HTML** (**H**yper **T**ext **M**arkup **L**anguage) is the standard markup language for creating web pages.
- **CSS** (**C**ascading **S**tyle **S**heet) describe how HTML elements are to be displayed on screen, paper, or in other media. It can control the layout of multiple web pages all at once.
- **JavaScript** is used by programmers across the world to create dynamic and interactive web content like applications and browsers. JavaScript is so popular that it's the most used programming language in the world, used as a client-side programming language.
- **AWE (Amazon Web Service)** is used for deployment of the model.

2.7 Constraints

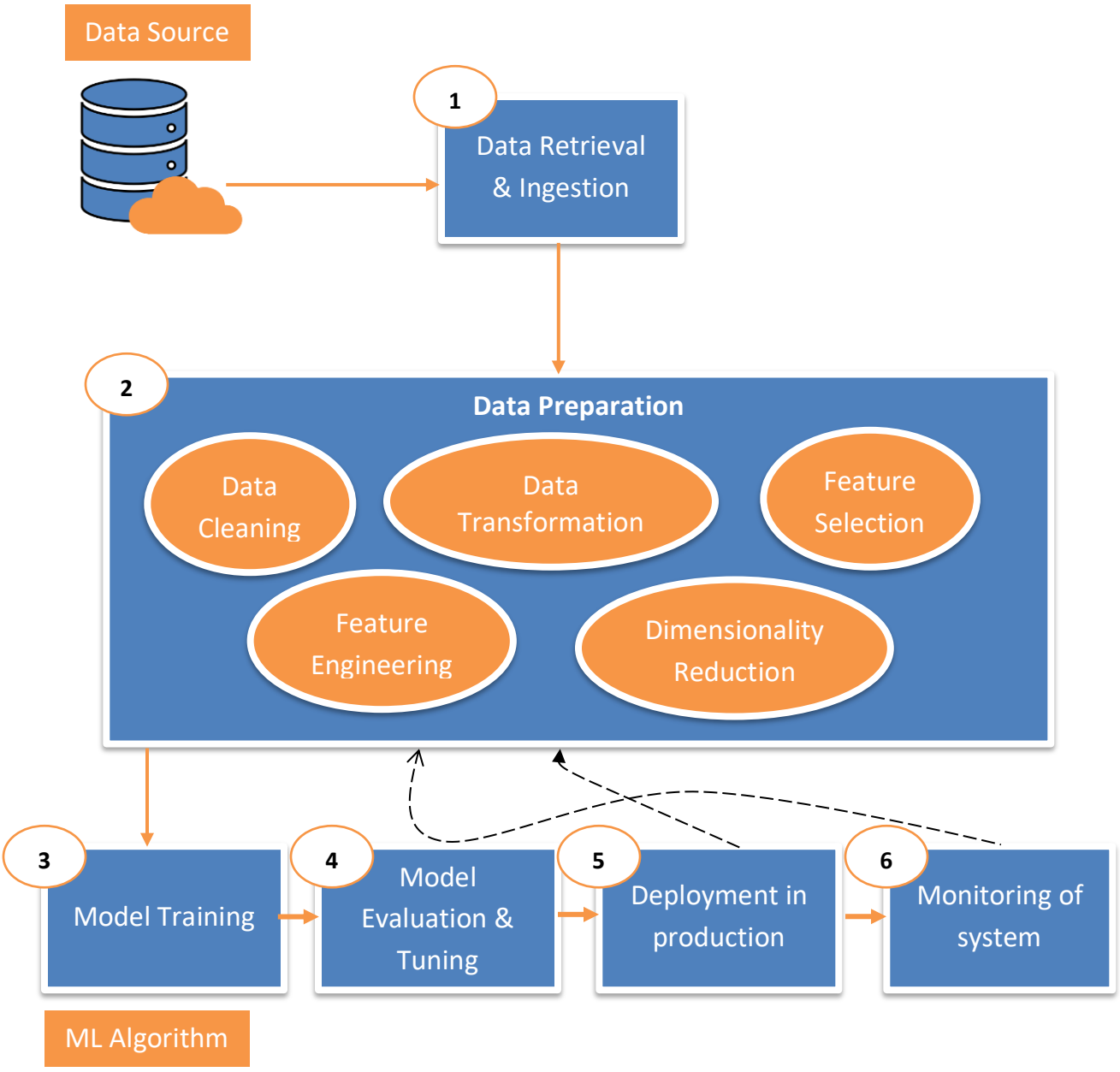
This model must be user friendly, as automated as possible and users should not be required to know any of the workings.

2.8 Assumptions

The main objective of the project is to develop an API to predict the defaults of the credit card holder on the basis of their characteristics and payment history. Machine learning based classification model is used for predicting and analyzing above mentioned problem statement on the input dataset.

3.0 Design Details

3.1 Process Flow



3.2 Event Log

The system should log every event or process so that the user will know what processes is running internally.

Initial Step-By-step Description:

1. The System identifies at what step logging required.
2. The System should be able to log each and every system flow.
3. Developer can choose logging method.
4. System should not hang even after using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

3.3 Error Handling

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

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 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 Resource Utilization

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

4.4 Deployment

5.0 Dashboards

A dashboard is a data visualization and analysis tool that displays on one screen the status of key performance indicators (KPIs) and other important business metrics.

As a high-level reporting mechanism, dashboards provide fast 'big picture' answer to critical business questions and assist and benefit decision making in several ways:

- Communicating how status and education is varies with payments.
- Visualizing relationship of gender with defaulters in easy-to-understand way.

6.0 Conclusion

The system is completely problem solving for commercial banks to analyze credit risk and predict the credit card defaulters by which they can take actions for recovery and also verify before giving loan or any other financial service to the following customers. Here, f1 evaluation score plays a key role in the prediction-based system. From the result we can see that Gaussian NB turned out to be best working model for this problem in terms of the accuracy.