

DOCUMENT VERSION CONTROL

| Date Issue | Version | Description | Author |
| --- | --- | --- | --- |
| 07/12/2023 | 1 | Initial HLD – V 1.0 | Abdul Jaweed  Lovely Patra |
| 17/12/2023 | 2 | Final HLD - V 2.0 | Abdul Jaweed  Lovely Patra |

CONTENT

* DOCUMENT VERSION
* ABSTRACT
* INTRODUCTION
* WHY THIS HIGH LEVEL DESIGN DOCUMENT
* SCOPE
* DEFINITION
* GENERAL DESCRIPTION
* PRODUCT PERSPECTIVE
* PROBLEM STATEMENT
* PROPOSED SOLUTION
* FURTHER IMPROVEMENTS
* TECHNICAL REQUIREMENTS
* TOOLS USED
* DATA REQUIREMENTS
* CONSTRAINTS
* IMPROVEMENTS
* DESIGN DETAILS
* PROCESS FLOW
* PRODUCT PERSPECTIVE
* PRODUCT STATEMENT
* PROPOSED SOLUTION
* EVENT LOG
* ERROR HANDLING
* PERFORMANCE
* REUSABILITY
* APPLICATION COMPATIBILITY
* RESOURCE UTILIZATION
* DEPLOYMENT
* KEY PERFORMANCE INDICATOR
* CONCLUSION

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 faced 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 holders. 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 a 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 the 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.

INTRODUCTION

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 modules interact at a high level.

THE HLD WILL -

* PRESENT ALL THE DESIGN ASPECTS AND DEFINE THEM IN DETAIL
* DESCRIBE THE USER INTERFACE BEING IMPLEMENTED
* DESCRIBE THE HARDWARE AND SOFTWARE INTERFACE
* DESCRIBE THE PERFORMANCE REQUIREMENT
* INCLUDE DEFINE FEATURE AND ARCHITECTURE OF THE PROJECT
* LIST AND DESCRIBE THE NON FUNCTIONAL ATTRIBUTES
* SECURITY
* RELIABILITY
* MAINTAINABILITY
* PORTABILITY
* REUSABILITY
* APPLICATION COMPATIBILITY
* RESOURCE UTILIZATION
* SERVICEABILITY

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.

TECHNICAL TERMS

| TERM | DESCRIPTION |
| --- | --- |
| DATABASE | Collection of Information Monitored by the System |
| IDE | Integrated Development Environment |
| AWS | Amazon Web Services |

GENERAL DESCRIPTION

PROBLEM STATEMENT

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 faced 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.

**Approach**: The classical machine learning tasks like Data Exploration, Data Cleaning, Feature Engineering, Model Building and Model Testing. Try out different machine learning algorithms that’s best fit for the above case.

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
* S3 buckets for cloud storage for the collected data.
* EC2 instances for running applications
* EC2 instances for tracking and monitoring experiments
* ECS cloud deployment.

TOOLS USED

1. AWS EC2 instance for MLFlow Server
2. MLFlow for Experiment tracking and monitoring metrics, parameters and models
3. Fast-API as Web Server Application
4. MongoDB for CRUD operations
5. Linux Service Management for MLFlow as a service
6. Nginx for MLFlow configuration.
7. GitHub for version control
8. Docker for containerization
9. AWS ECS and ECR for Container Deployment and Management

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.

CONSTRAINTS

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

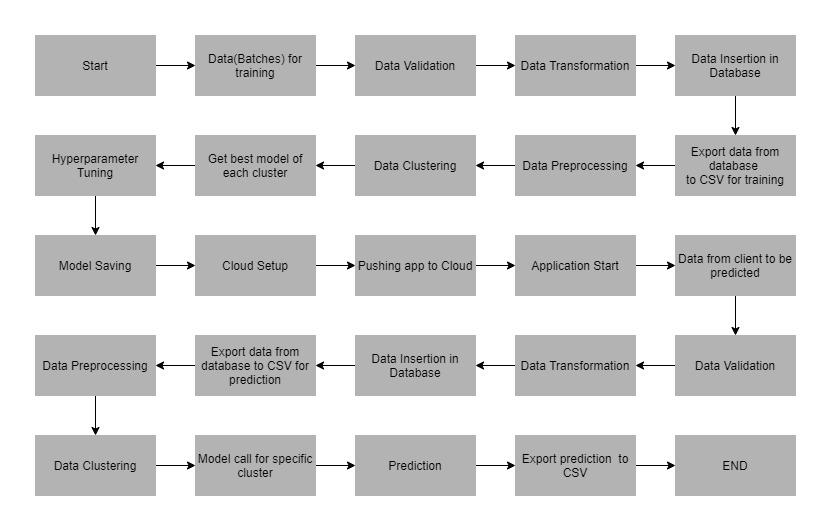
ASSUMPTIONS

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 statements on the input dataset.

DESIGN DETAIL

PROCESS FLOW

For predicting the failure of an engine component., we will use machine learning based models.

The process will look like 

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 helps to find default situations for individuals and also for more and more holders.

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 an API interface to predict the default, and to detect whether the status of the owner affects the defaults of the credit card.

PROPOSED SOLUTION

The solution proposed here is to estimate the defaults of credit cards based on owner’s characteristics and payment history data; this can be implemented to perform above mentioned use cases. In the 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 the second case, if the model detects that the owner defaults then the bank takes action against them. And in the last use case, we will be making an interface to predict the defaults of credit card holders.

EVENT LOG

The system is going to log everything so that the user gets to know which process is running internally.. Logs are stored in MongoDB

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.

PERFORMANCE

The credit card default prediction solution is used for predicting the credit card defaulter. Whenever there is a default prediction, it will inform concerned parties and authorities. Also model retraining is important to improve the performance.

REUSABILITY

The code written should have the ability to be reused with no problems.

APPLICATION COMPATIBILITY

The different components for this project will be using Python as an interface between them. Each component will have its own task to perform, and it is the job of the Python to ensure proper transfer of information.

RESOURCE UTILIZATION

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

KEY PERFORMANCE INDICATOR

The only indicator in our application will be the score of the trained model and system performance. This means that the application is able to predict the correct output for a given input.

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, the 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 the best working model for this problem in terms of the accuracy.