HIGH LEVEL DESIGN (HLD)

Predict Credit Risk Using South German Bank Data

Revision Number : 1.0

Last Date of Revision : 10/12/2021

Document Version Control

Date Issued	Version	Description	Author
10/12/2021	1.0	1. Document Created	Laiju P Joy

Content

Abstract

This dataset classifies people described by a set of attributes as good or bad credit risks. The data comes in two formats one all numeric & one comes with a cost matrix. The analysis of credit risk depends on the feature that is given in the dataset. There are 20 features available in dataset and one target feature credit risk is present. Total no. of records is 1000 and there is no duplicate value or missing value is present in the dataset. Out of 1000 records 700 records are good risk and 300 records are bad credit risk. The given classification in which the good credit risk is denoted by 1 and bad credit risk is denoted by 0. Two dataset are provided the original dataset, in the form provided by Prof. Hofmann, contains categorical/symbolic attributes and is in the file "German Data". For algorithms that need numerical attributes, Strathclyde University produced the file "German Data-Numeric". This file has been edited and several indicator variables added to make it suitable for algorithms which cannot cope with categorical variables. Several attributes that are ordered categorical (such as attribute 17) have been coded as integer. This was the form used by Stat Log.

Data Set Characteristics:	Multivariate	**Number of Instances:**	1000
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	20
Associated Tasks:	Classification	**Missing Values?**	N/A

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 modules interact 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 or mildly-technical terms which should be understandable to the administrators of the system.

1.3 Definitions

Below are given the short form some term that are used in the document.

Term	Definitions
Database	Collection of all the information monitored by this system
CR	Credit Risk
CD	Credit Data
IDE	Integrated Development Environment
AWS	Amazon Web Service

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

1.5 Uses

- This document is designed to help in operational requirement and can be used as a reference manual for how the modules interact.
- HLD briefly describes about the platforms/products/services/processes, flow of traffic that it depends on and includes any important changes that need to be made to them.
- HLD is the input for creating the LLD (Low Level Design) since the key communication items are displayed in HLD which are then converted to detailed communication in LLD, showing connectivity and physical level

1.6 Application

1. Website Development
2. Application Development
3. Data Science Project

2. General Description

2.1 Product Perspective

The prediction of credit risk using the South German Bank data application use for to check the whether the person credit risk is good or bad. The credit risk is useful to check that the person is having good credit risk or bad for the loan. In dataset each row represents the each person different conditions.

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

Credit risk is the possibility of a loss resulting from a borrower's failure to repay a loan or meet contractual obligations. Traditionally, it refers to the risk that a lender may not receive the owed principal and interest, which results in an interruption of cash flows and increased costs for collection. Excess cash flows may be written to provide additional cover for credit risk. When a lender faces heightened credit risk, it can be mitigated via a higher coupon rate, which provides for greater cash flows. Although it's impossible to know exactly who will default on obligations, properly assessing and managing credit risk can lessen the severity of a loss. Interest payments from the borrower or issuer of a debt obligation are a lender's or investor's reward for assuming credit risk. So from the above problem statement we can consider the some of the input parameter that are required to the calculation of credit risk.

2.4 Technical Requirements

In this Project the requirements to gather data from the client for the predictions of credit risk. For that we are using the different technologies. Below are some requirements for this project.

- Model should be exposed through API or User Interface, so that anyone can test model.
- Model should be deployed on cloud (Azure, AWS, and GCP) or Heroku for the public used.
- Cassandra database should be integrated in this project for storing the client data and then from that we can take it for the further purpose.

2.5 Data Requirements

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

Sr. No.	Column name	Variable name	Description	Data Content
1	Laufkont	Status	Status of the debtor's checking account with the bank	A11 : ... < 0 DM A12 : 0 <= ... < 200 DM A13 : ... >= 200 DM / salary assignments for at least 1 year A14 : no checking account
2	Laufzeit	Duration	Credit duration in months	Duration in month
3	Moral	Credit_history	History of compliance with previous or concurrent credit contracts	A30 : no credits taken/ all credits paid back duly A31 : all credits at this bank paid back duly A32 : existing credits paid back duly till now A33 : delay in paying off in the past A34 : critical account/ other credits existing (not at this bank)
4	Verw	Purpose	Purpose for which the credit is needed	A40 : car (new) A41 : car (used) A42 : furniture/equipment A43 : radio/television A44 : domestic appliances A45 : repairs A46 : education A47 : (vacation - does not exist?) A48 : retraining A49 : business A410 : others

5	Hoehe	Amount	Credit amount	Credit amount
6	Sparkont	Savings	Debtor's savings	A61: ... < 100 DM A62: 100 <= ... < 500 DM A63: 500 <= ... < 1000 DM A64: >= 1000 DM A65 : unknown/ no savings account
7	Beszeit	Employment_duration	Duration of debtor's employment with current employer	A71: unemployed A72: ... < 1 year A73: 1 <= ... < 4 years A74: 4 <= ... < 7 years A75: >= 7 years
8	Rate	Installement _rate	Credit installments as a percentage of debtor's disposable income	Instalment rate in percentage of disposable income
9	Famges	Personal_status_sex	Combined information on sex and marital status	Personal status and sex A91 : male : divorced/separated A92 : female : divorced/separated/married A93 : male : single A94 : male : married/widowed A95 : female : single
10	Buerge	Other_debtors	Is there another debtor or a guarantor for the credit?	Other debtors / guarantors A101 : none A102 : co-applicant A103 : guarantor
11	Wohnzeit	Present_residence	Length of time (in years) the debtor lives in the present residence	Present residence since
12	Verm	Property	The debtor's most valuable property	A121 : real estate A122 : if not A121 : building society savings agreement/ life insurance A123 : if not A121/A122 : car or other, not in attribute 6 A124 : unknown / no property
13	Alter	Age	Age in years	Age in years
14	Weitkred	Other_installment_plans	Installments plans from providers other than the credit-giving bank	A141 : bank A142 : stores A143 : none
15	Wohn	Housing	Type of housing the debtor lives in	A151 : rent A152 : own A153 : for free
16	Bishkred	Number_credit	Number of credits including the current one the debtor has (or had) at this bank	Number of existing credits at this bank

17	Beruf	Job	Quality of debtor's job	A171 : unemployed/ unskilled - non-resident A172 : unskilled - resident A173 : skilled employee / official A174 : management/ self-employed/ highly qualified employee/ officer
18	Pers	People_liable	Number of persons who financially depend on the debtor	Number of people being liable to provide maintenance for
19	Telef	Telephone	Is there a telephone landline registered on the debtor's name?	A191 : none A192 : yes, registered under the customer's name
20	Gastarb	Foregin_worker	Is the debtor a foreign worker?	A201 : yes A202 : no
21	Kredit	Credit_risk	Has the credit contract been complied with (good) or not (bad)?	(1 = Good, 2 = Bad)

.

2.6 Tools Used

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

- Cassandra is used to insert, delete, retrieve and update the database.
- For visualization of the plots, Matplotlib, Seaborn and Plotly are used.
- AWS & Heroku are used for deployment of the model.
- Front end development is done using HTML/CSS
- Python flask is used for backend development.
- Jupyter notebook is used for EDA purpose.
- GitHub is used as version control system.
- VS Code Is Used as IDE.
- PyCharm is used as IDE.

Figure: - the entire library used in project

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

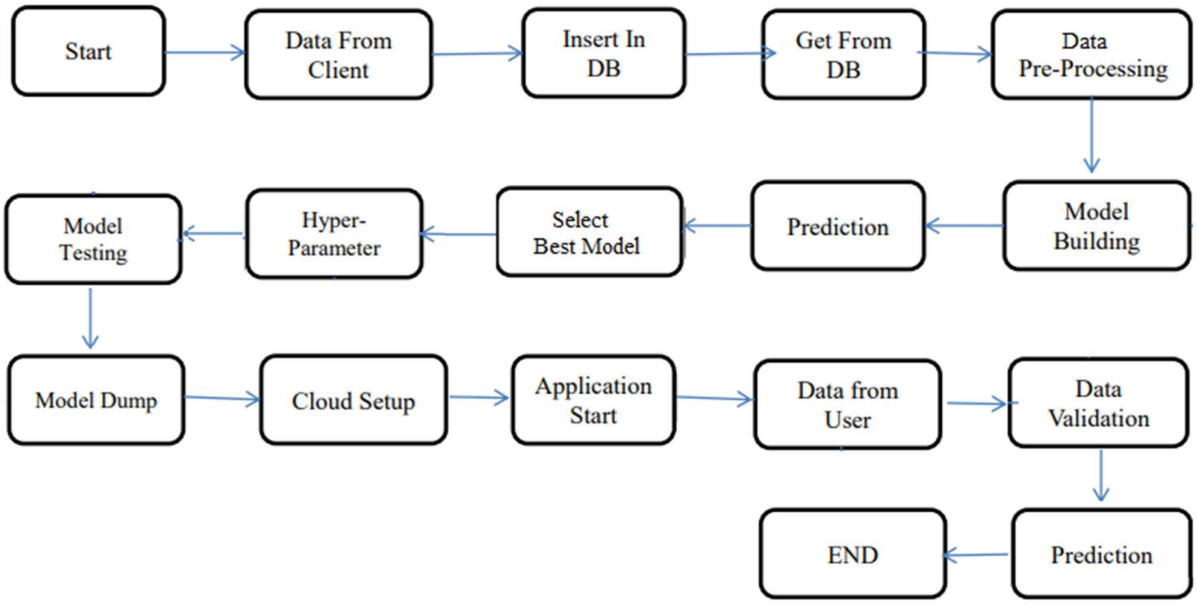

Figure: - The entire Project Flow

3.2 Deployment Flow

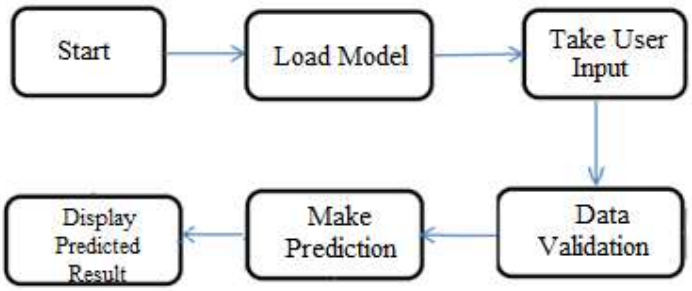

Figure: - The Deployment Flow

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 database.
- By logging we are monitor every step which may create problem or every step which is important in file system.
- We have designed logging in such a way that system should not hang even after so many logging's, so that we can easily debug issues which may arises 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.

4. 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 complete process of Machine Learning. Here are summary of complete process:

- First we cleaned our dataset properly by removing all null value and duplicate value present in dataset and if not present leave the dataset as it is.
- Then we have to handle the outlier form the dataset so that it can't effect on the accuracy. So in our dataset there is no outlier present.
- After that drop few non important coloumns in dataset, that maybe not import for this prediction and its more imbalanced data.
- Converting the categorical columns having ordinal values to Label Encoding
- Converting the categorical columns having non-ordinal values to One Hot Encoding
- Then we split the whole data set into train-test split with test size will be 30% of the whole dataset for the model training approach.
- After performing above step we ready for model training. In this step, we trained our dataset on different classification algorithm such as Logistic Regression, Random Forest Classifier, Support Vector Machine, K- Nearest Neighbor and Naïve Bayes Classifier, and test the models with test data for check F1 score and accuracy.
- After that we select best model from the highest F1 score and accuracy.
- After that we applied hyper-parameter tuning on selected model which have the highest F1 score and accuracy. The reason behind is that we can use best feature so that we get maximum accuracy and F1 score.
- Then we train selected model with best parameters available from the hyper-parameter tuning, and test the model with test data for check F1 score and accuracy.
- Then we saved the model in pickle file format for model deployment.
- After that our model is ready to deploy. We deployed our model in AWS cloud storage and Heroku platform.

4.1 Reusability

We have done programing of this project in such way that it should be reusable. So that anyone can add and contribute without facing problem.

4.2 Application Compatibility

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

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

We have deployed the project on AWS cloud & Heroku platform. In aws we use elastic beanstalk service for deployment

Figure: - Deployment Platform

4.5 User Interface

We have created an UI interface for user by using HTML and CSS. In which we created two pages, In that 1st one is our main page and 2nd one is our result page.

1. Home Page

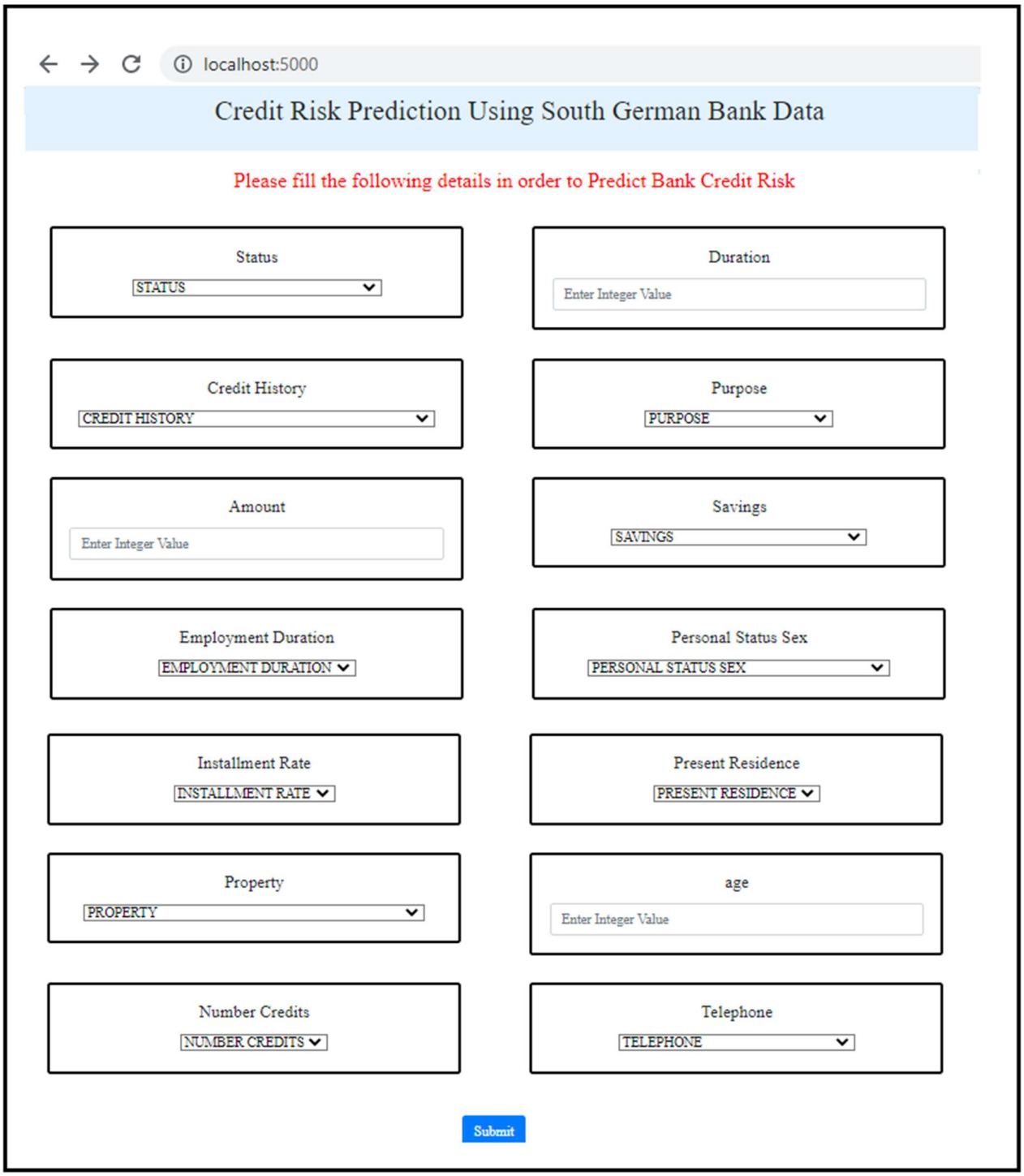

Figure: - Front Page

2. Result Page

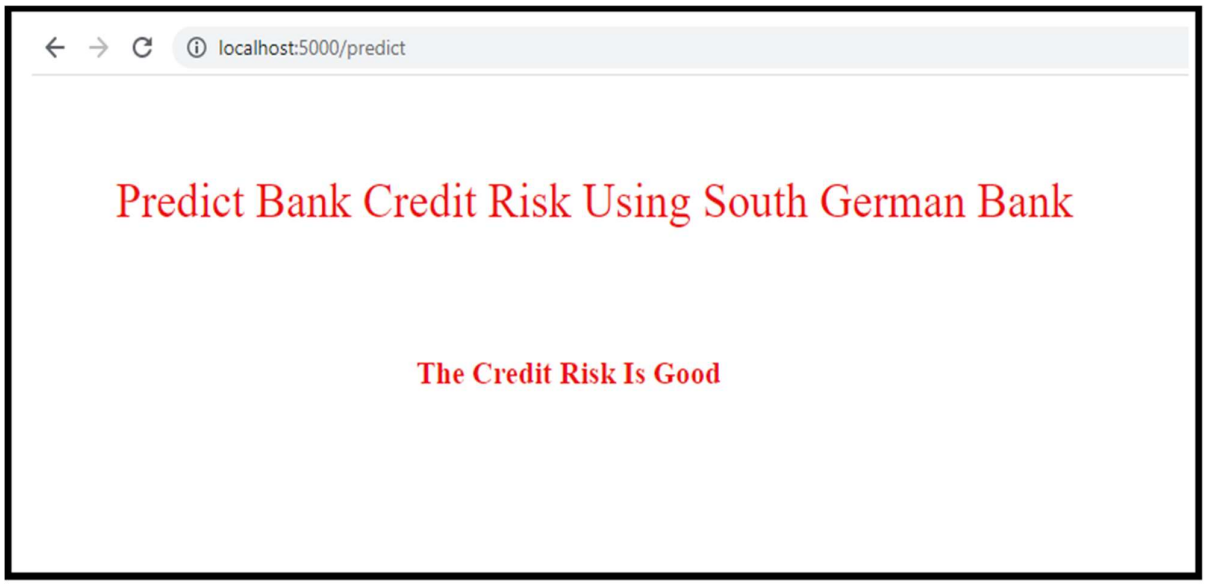

Figure: - Result Page

5. Conclusion

The Prediction of Credit Risk Using South German Bank Data application will find out the Credit Risk is good or bad. Then from that we decide that the given person eligible for bank loan or not.