Bank Credit Risk Classification

High Level Design

Gaurav Singh
12th July, 2023

Document Version Control

Date Issued	Version	Description	Author
12th July 2023	1.0	Initial HLD	Gaurav Singh

Contents

Abstract

Credit risk plays a major role in the banking industry business. Banks' main activities involve granting loan, credit card, investment, mortgage, and others. Credit card has been one of the most booming financial services by banks over the past years. However, with the growing number of credit card users, banks have been facing an escalating credit card default rate. As such data analytics can provide solutions to tackle the current phenomenon and management credit risks.
This project discusses the implementation of model which classifies a given profile as a good risk or a bad risk.

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. 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

1.3. 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.

2. General Description

2.1.Product Perspective

The Credit Card Risk Classification system is a machine learning-based classification model which will help us to classify a given customer profile into either Good Risk or Bad Risk class.

2.2. Problem statement

Financial threats are displaying a trend about the credit risk of banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats faced by banks is the risk prediction of credit clients. The goal is to predict the credit risk of an applicant based on their certain demographic and behavioral characteristics.

2.3. Proposed Solution

The solution proposed here is a web application, which predicts the credit risk for a customer based on the customer's demographic data and behavioral data.

2.4. Data Requirements

This dataset is taken from the UCI Machine Learning Repository (url: https://archive.ics.uci.edu/ml/datasets/South+German+Credit). It contains information on defaults, demographic factors, credit data etc. of customers.
There are 21 variables:

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

- **duration** : credit duration in months (quantitative)

- **credit_history** : history of compliance with previous or concurrent credit contracts (categorical)
    - **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

- **purpose** : purpose for which the credit is needed (categorical)
    - **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

- **amount** : credit amount in DM (quantitative; result of monotonic transformation; actual data and type of transformation unknown)

- **savings** : debtor's savings (categorical)
  - **1** : unknown/no savings account
  - **2** : ... < 100 DM
  - **3** : 100 <= ... < 500 DM
  - **4** : 500 <= ... < 1000 DM
  - **5** : ... >= 1000 DM

- **employment_duration** : duration of debtor's employment with current employer (ordinal; discretized quantitative)
  - **1** : unemployed
  - **2** : < 1 yr
  - **3** : 1 <= ... < 4 yrs
  - **4** : 4 <= ... < 7 yrs
  - **5** : >= 7 yrs

- **installment_rate** : credit installments as a percentage of debtor's disposable income (ordinal; discretized quantitative)
  - **1** : >= 35
  - **2** : 25 <= ... < 35
  - **3** : 20 <= ... < 25
  - **4** : < 20

- **personal_status_sex** : combined information on sex and marital status; categorical; sex cannot be recovered from the variable, because male singles and female non-singles are coded with the same code (2); female widows cannot be easily classified, because the code table does not list them in any of the female categories
  - **1** : male : divorced/separated
  - **2** : female : non-single or male : single
  - **3** : male : married/widowed
  - **4** : female : single

- **other_debtors :** Is there another debtor or a guarantor for the credit? (categorical)
  - **1** : none
  - **2** : co-applicant
  - **3** : guarantor

- **present_residence :** length of time (in years) the debtor lives in the present residence (ordinal; discretized quantitative)
  - **1** : < 1 yr
  - **2** : 1 <= ... < 4 yrs
  - **3** : 4 <= ... < 7 yrs
  - **4** : >= 7 yrs

- **property :** the debtor's most valuable property, i.e. the highest possible code is used. Code 2 is used, if codes 3 or 4 are not applicable and there is a car or any other relevant property that does not fall under variable savings. (ordinal)
  - **1** : unknown / no property
  - **2** : car or other
  - **3** : building soc. savings agr./life insurance
  - **4** : real estate
- **age :**age in years (quantitative)

- **other_installment_plans :**installment plans from providers other than the credit-giving bank (categorical)
  - **1** : bank
  - **2** : stores
  - **3** : none

- **housing :**type of housing the debtor lives in (categorical)
  - **1** : for free
  - **2** : rent
  - **3** : own

- **number_credits :**number of credits including the current one the debtor has (or had) at this bank (ordinal, discretized quantitative); contrary to Fahrmeir and HamerleÃ¢â‚¬â„¢s (1984) statement, the original data values are not available.
  - **1** : 1
  - **2** : 2-3
  - **3** : 4-5
  - **4** : >= 6

- **job :**quality of debtor's job (ordinal)
  - **1** : unemployed/unskilled - non-resident
  - **2** : unskilled - resident
  - **3** : skilled employee/official
  - **4** : manager/self-empl./highly qualif. employee

- **people_liable :**number of persons who financially depend on the debtor (i.e., are entitled to maintenance) (binary, discretized quantitative)
  - **1** : 3 or more
  - **2** : 0 to 2

- **telephone :**Is there a telephone landline registered on the debtor's name? (binary; remember that the data are from the 1970s)
    - **1 :** No
    - **2 :** Yes

- **foreign_worker :**Is the debtor a foreign worker? (binary)
    - **1** : yes
    - **2** : no

- **credit_risk :**Has the credit contract been complied with (good) or not (bad) ? (binary)
    - **0** : bad
    - **1** : good

2.5. Tools used

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

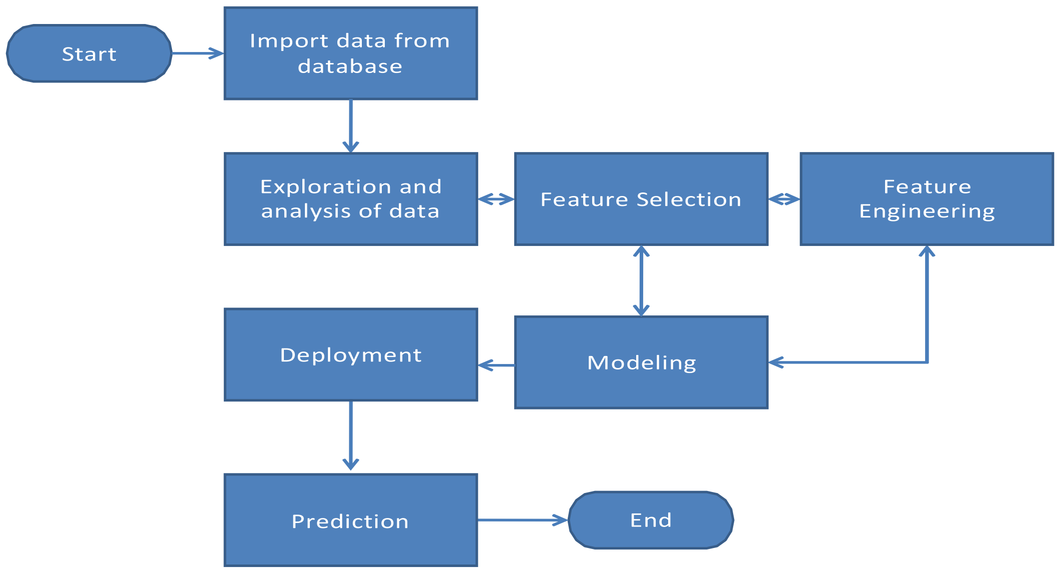
- Jupyter Notebook and VScode are used as ExperimentIDE.
- For visualization of the plots, Matplotlib and Seaborn are used.
- Model is deployed locally on the windows system.
- Front end development is done using HTML/CSS.
- Python Flask Module is used for backend RestApi development.
- Cassandra Nosql is used as database.
- GitHub is used as code version control system and DVC is used for controlling data version and for model drift analysis.
- Github Actions is used as ci/cd pipeline.

3. Design Details

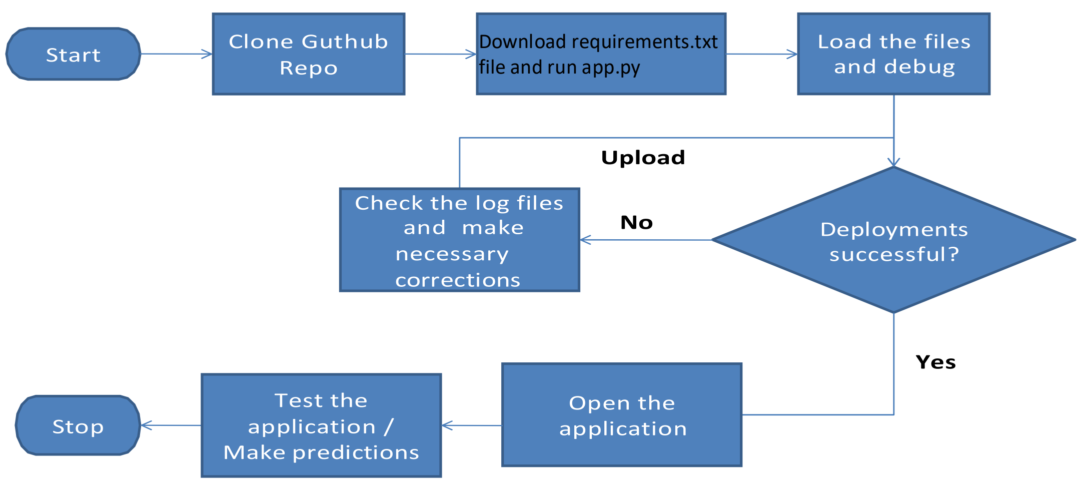
3.1. Process Flow

For identifying the class of each profile, we will use a machine learning model.
Below is the process flow diagram is as shown below.

3.2. Proposed methodology



3.3. Deployment Process



3.4. Event Log
The event logs are stored in log files.

3.5. Performance
The Bank Credit Risk Classification app is used to classify whether a given profile could be approved for credits, would they pose a risk for the bank/financial organization etc. Characteristics like age, job, savings, credit details etc are examined and based on the analysis, the applicants are classified into Good Risk and Bad Risk categories.

3.6. Reusability
The code written and the components used should have the ability to be reused with no problems.

3.7. 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.

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

3.9. Deployment
The model can be deployed locally as well as in any cloud services such as Microsoft Azure, AWS, Google, Heroku etc.

4. Conclusion

This application will classify profiles/applicants for credit into Good Risk and Bad Risk categories, and can help financial institutions in taking necessary actions to prevent further loss.