Bank Credit Risk Classification

Project Report Presentation

Prepared by: Shruti Balan,
Data Science Intern,
Ineuron

AGENDA

- Introduction
- Objective
- Data Description
- Architecture
- Model Training and Evaluation Workflow
- Deployment
- Questions

Introduction

- 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 an model which classifies a customer/applicant profile into Good risk or Bad risk based on certain demographic and behavioral criteria.

Objective

- Development of a model for classifying a customer profile into Risk categories(Good or Bad).
- Benefits:
  - Predicting the level of risk any given customer might pose for the bank/financial institution if allowed with the requested credit.
  - Gives better insight of customer base.
  - Allows financial institutions to take necessary steps to minimize the lose.

Data Description

- **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

Data Description (cont..)

- **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)

Data Description (cont..)

- **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

Data Description (cont..)

- **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

Data Description (cont..)

- **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

Data Description (cont..)

- **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
  - **1** : 1
  - **2** : 2-3
  - **3** : 4-5
  - **4** : >= 6

Data Description (cont..)

- **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

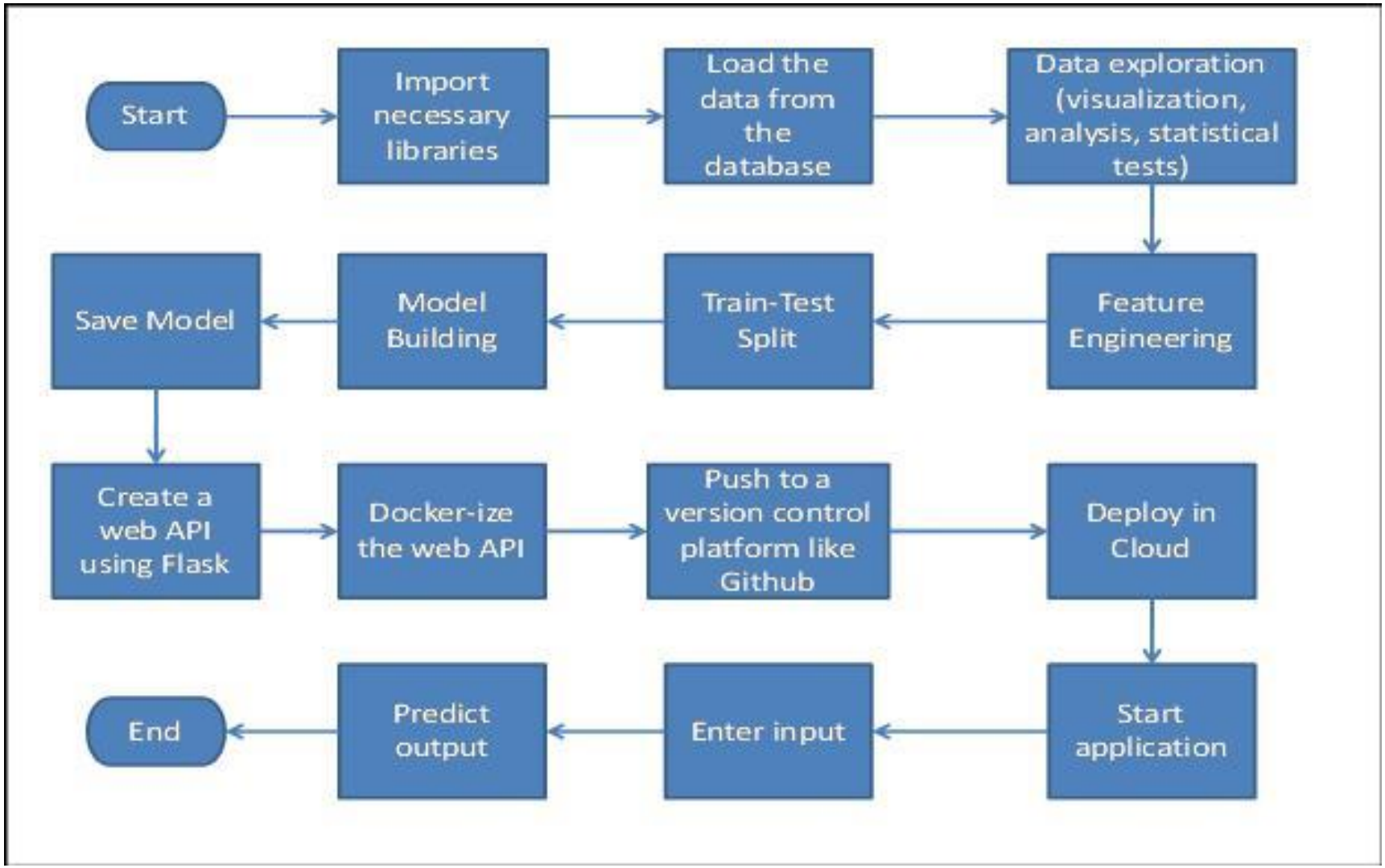
- **telephone :**Is there a telephone landline registered on the debtor's name? (binary; remember that the data are from the 1970s)
  - **1 :** No
  - **2 :** Yes

Data Description (cont..)

- **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

Architecture



Architecture (cont..)

- **Data Exploration**

  We divide the data into two types: numerical and categorical. We explore through each type one by one. Within each type, we explore, visualize and analyze each variable one by one, perform statistical tests and note down our observations. We also make some minor changes in the data like change column names for convenience in understanding.

- **Feature Engineering**
  - Encoded categorical variables.
  - Engineering new features

- **Train/Test Split**

  Split the data into 80% train set and 20% test set.

- **Model Building**
  - Built models and trained and tested the data on the models.
  - Compared the performance of each model and selected the best one.

Architecture (cont..)

- **Save the model**
  Saved the model by converting into a pickle file.

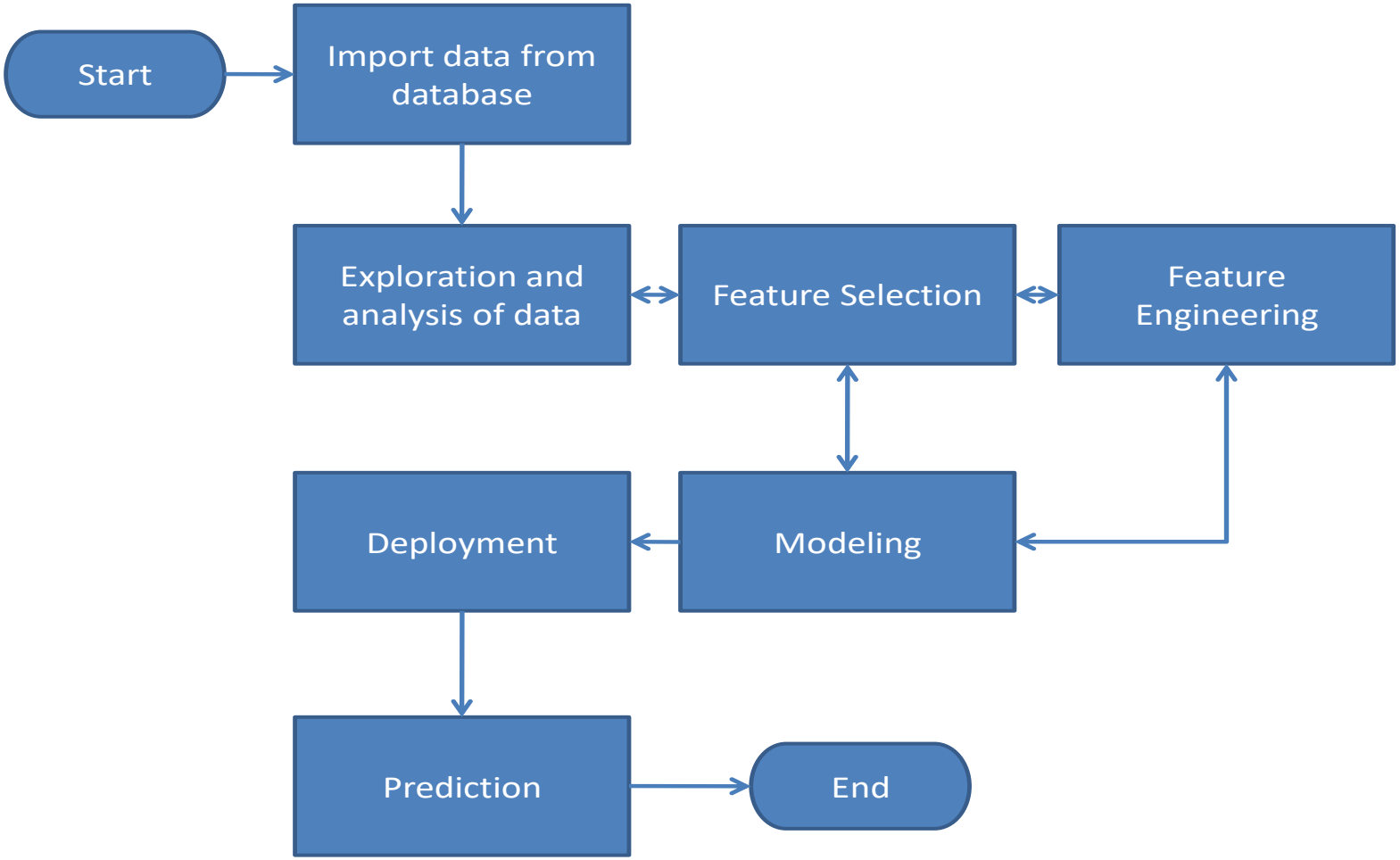
- **Create an API and Dockerize the entire application**

- **Cloud Setup & Pushing the App to the Cloud**
  Selected Heroku for deployment. Loaded the application files from Github to Heroku.

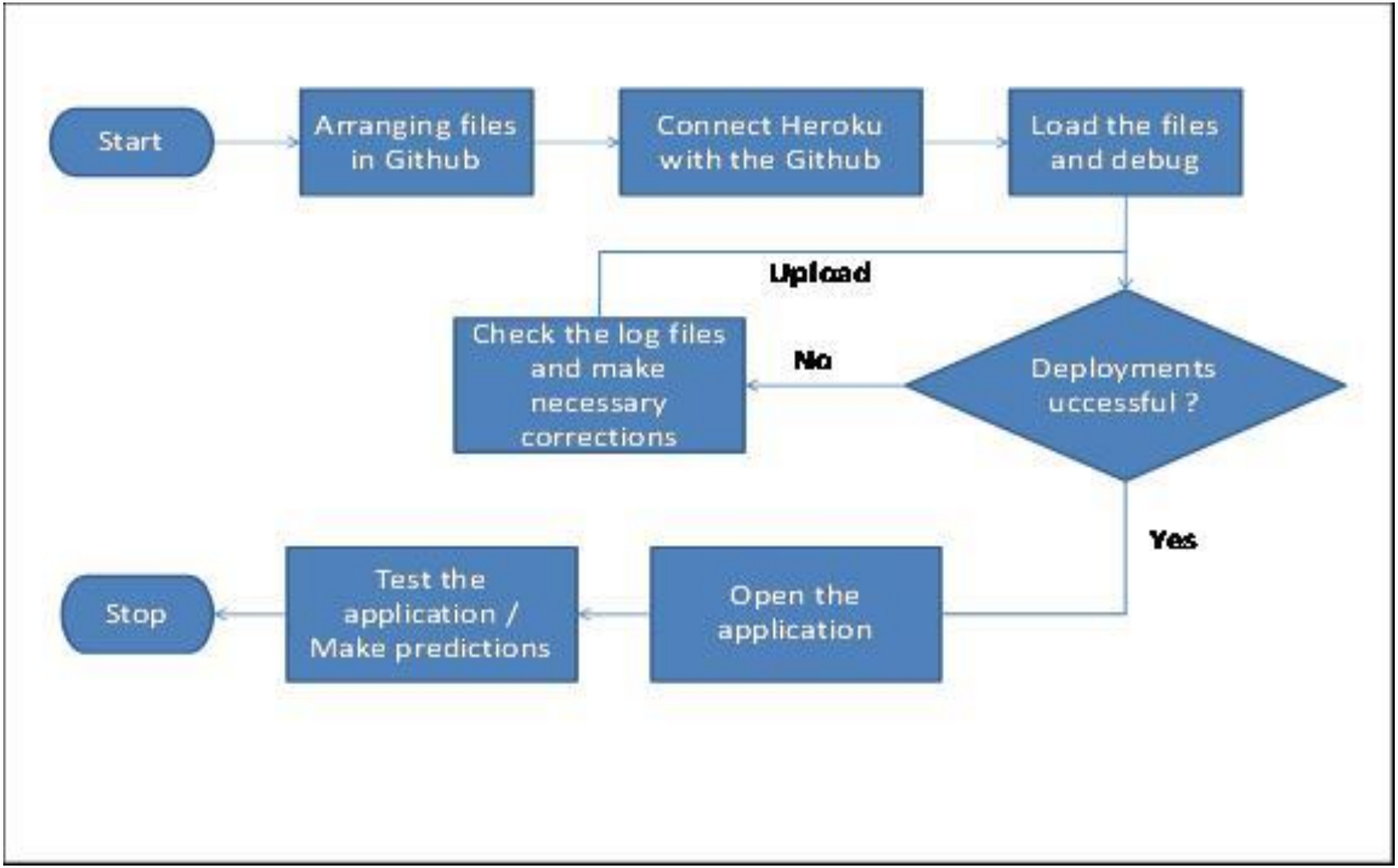
- **Application Start and Input Data by the User**
  Start the application and enter the inputs.

- **Prediction**
  After the inputs are submitted the application runs the model and makes predictions. The out is displayed as a message indicating whether the customer whose demographic and behavioral data are entered as inputs, is likely to default in the following month or not.

Model Training and Evaluation Workflow



Deployment


Start
Arranging files in Github
Connect Heroku with the Github
Load the files and debug
Upload
Check the log files and make necessary corrections
No
Deployments uccessful ?
Yes
Stop
Test the application / Make predictions
Open the application

FAQs

1) **What is the data source?**

   The data is obtained from UCI Machine Learning Repository.

   Link : https://archive.ics.uci.edu/ml/datasets/South+German+Credit

2) **What was the type of data?**

   The data contained both numerical and continuous type data.

3) **What was the complete flow that you followed in this project?**

   Please refer to slides 13 to 15.

4) **How logs are managed?**

   We have a separate log files for each stage of the project.

FAQs

5) **What techniques were you using for data pre-processing?**
   - Removing unwanted attributes
   - Visualizing relation of independent variables with each other and output variables
   - Cleaning data and imputing if null values are present.
   - Encoding categorical variables

6) **How training was done or what models were used?**
   - After loading the dataset, data pre-processing was done.
   - For this project, we opted to train the data using the XGBoost Classifier.
   - Hyper-parameter tuning, feature selection were performed during the various versions of modeling.
   - The best model was selected.

FAQs

7) How Prediction was done?

- The test files were provided.
- The test data also underwent preprocessing.
- Then the data was passed through the model and output was predicted.

8) What are the different stages of deployment?

- After training the model, we prepared all the necessary files required for deployment and uploaded in a document version control system called Github.
- We then connected to and deployed the model in, Heroku.

THANK YOU