CREDIT CARD DEFAULT PREDICTION

Project Report Presentation

Prepared by: Puneet Soni,
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 predicts if a given credit card holder has a probability of defaulting in the following month, using their demographic data and behavioral data from the past 6 months.

Objective

- Development of a model for predicting if a given customer id has a probability to default in the following month or not.
- Benefits:
  - Detection of upcoming frauds.
  - Gives better insight of customer base.
  - Allows financial institutions to take necessary steps to minimize the lose from the possible defaults.

Data Description

- **ID**: ID of each client
- **LIMIT_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit
- **SEX**: Gender
  - 1=male,
  - 2=female
- **EDUCATION:**
  - 1=graduate school,
  - 2=university,
  - 3=high school,
  - 0, 4, 5, 6=others)
- **MARRIAGE**: Marital status
  - 1=married,
  - 2=single,
  - 3=divorce,
  - 0=others

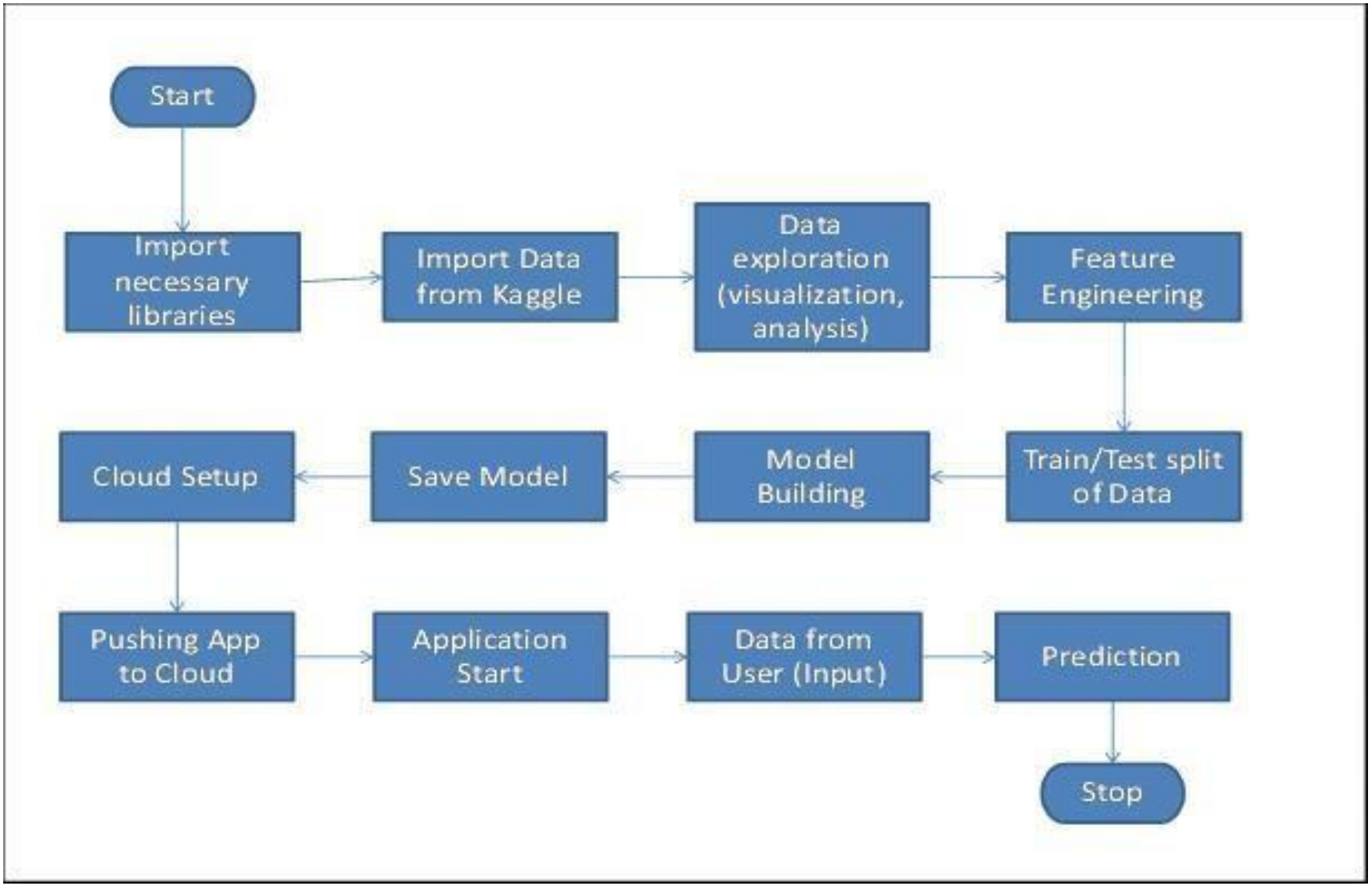
Data Description (cont..)

- **AGE**: Age in years
- **PAY_0**: Repayment status in September, 2005
  - -2: No consumption;
  - -1: Paid in full;
  - 0: The use of revolving credit;
  - 1 = payment delay for one month;
  - 2 = payment delay for two months; . . .;
  - 8 = payment delay for eight months;
  - 9 = payment delay for nine months and above.
- **PAY_2**: Repayment status in August, 2005 (scale same as above)
- **PAY_3**: Repayment status in July, 2005 (scale same as above)
- **PAY_4**: Repayment status in June, 2005 (scale same as above)
- **PAY_5**: Repayment status in May, 2005 (scale same as above)
- **PAY_6**: Repayment status in April, 2005 (scale same as above)

Data Description (cont..)

- **BILL_AMT1**: Amount of bill statement in September, 2005 (NT dollar)
- **BILL_AMT2**: Amount of bill statement in August, 2005 (NT dollar)
- **BILL_AMT3**: Amount of bill statement in July, 2005 (NT dollar)
- **BILL_AMT4**: Amount of bill statement in June, 2005 (NT dollar)
- **BILL_AMT5**: Amount of bill statement in May, 2005 (NT dollar)
- **BILL_AMT6**: Amount of bill statement in April, 2005 (NT dollar)
- **PAY_AMT1**: Amount of previous payment in September, 2005 (NT dollar)
- **PAY_AMT2**: Amount of previous payment in August, 2005 (NT dollar)
- **PAY_AMT3**: Amount of previous payment in July, 2005 (NT dollar)
- **PAY_AMT4**: Amount of previous payment in June, 2005 (NT dollar)
- **PAY_AMT5**: Amount of previous payment in May, 2005 (NT dollar)
- **PAY_AMT6**: Amount of previous payment in April, 2005 (NT dollar)
- **default.payment.next.month**: Default payment
  - 1=yes,
  - 0=no

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 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..)

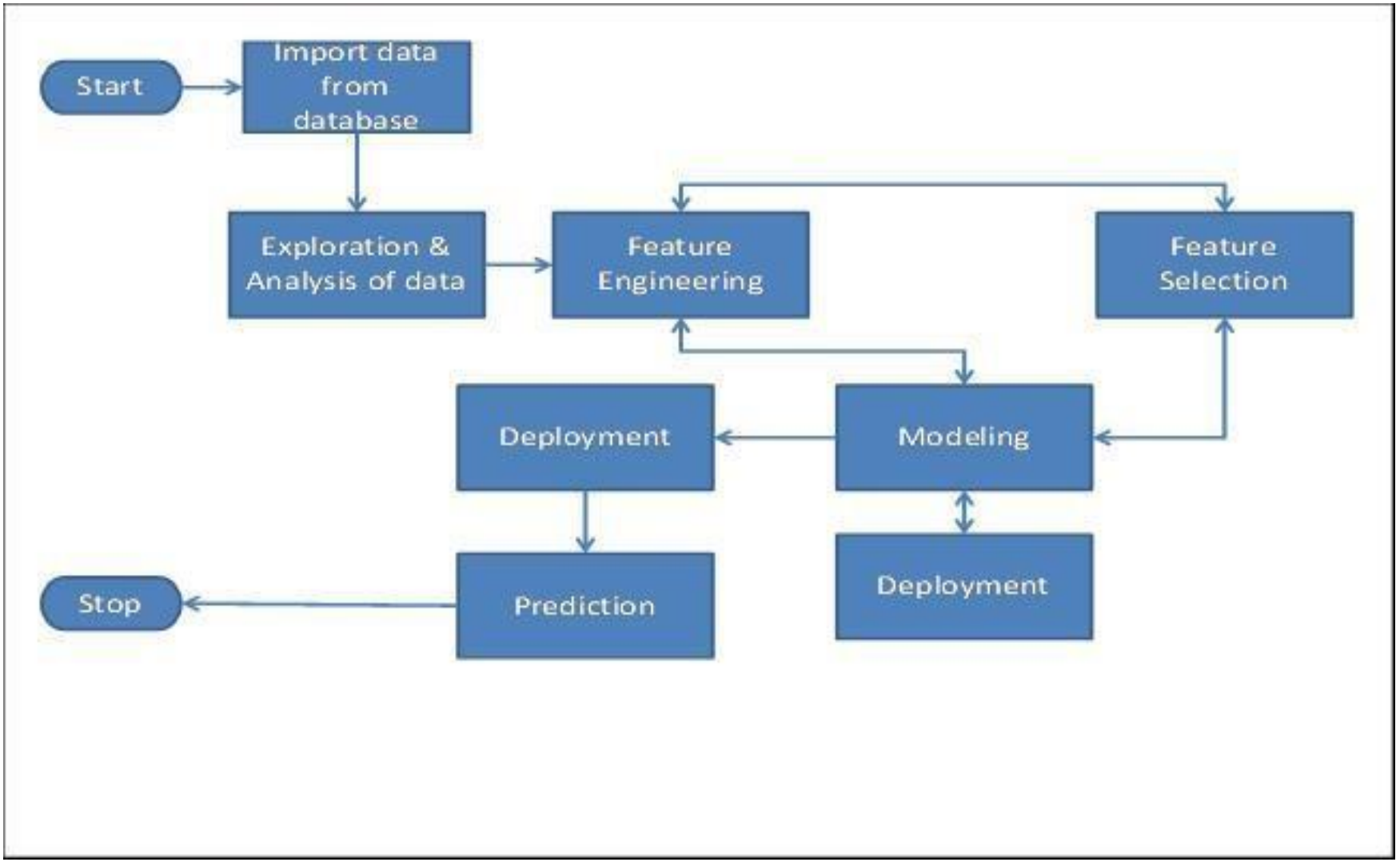
- **Save the model**
  Saved the model by converting into a pickle file.

- **Cloud Setup & Pushing the App to the Cloud**
  Selected AWS for deployment. Loaded the application files fromGithub to AWS.

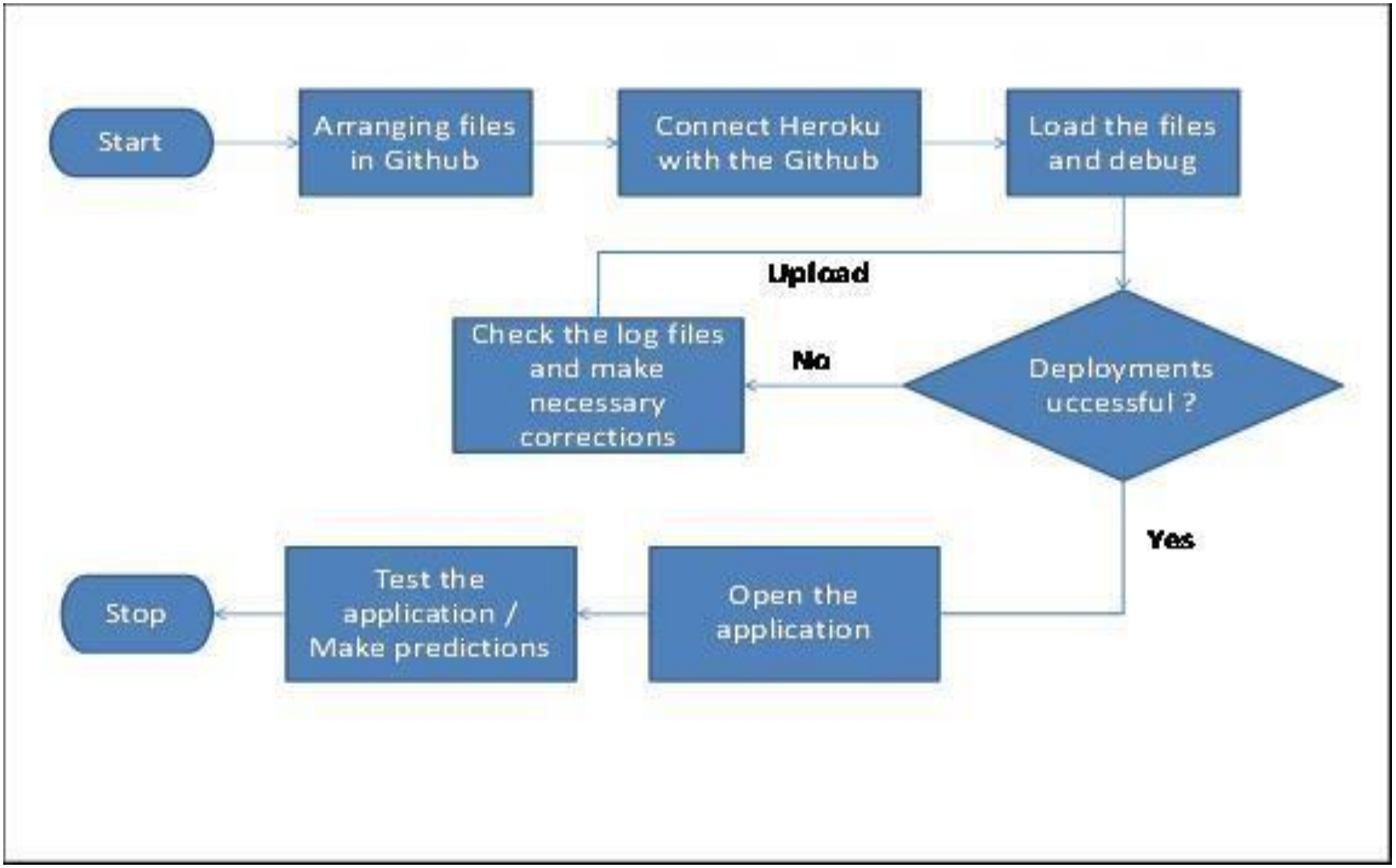
- **Application Start and Input Data by the User**
  Start the application and enter the inputs.

- **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 Kaggle. Link : https://www.kaggle.com/uciml/defaultof-credit-card-clients-dataset

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 7 to 10.

4) **How logs are managed?**

   We have a single log file for the entire project. However, we can create different log files for each stages in the project cycle, if needed.

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
   - Checking and changing Distribution of continuous values
   - Cleaning data and imputing if null values are present.
   - Encoding categorical variables
   - Scaling the data

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 Random Forest Classifier.
   - Hyper-parameter tuning, feature selection and new features were engineered 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 and new features required for the model were prepared.
- 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