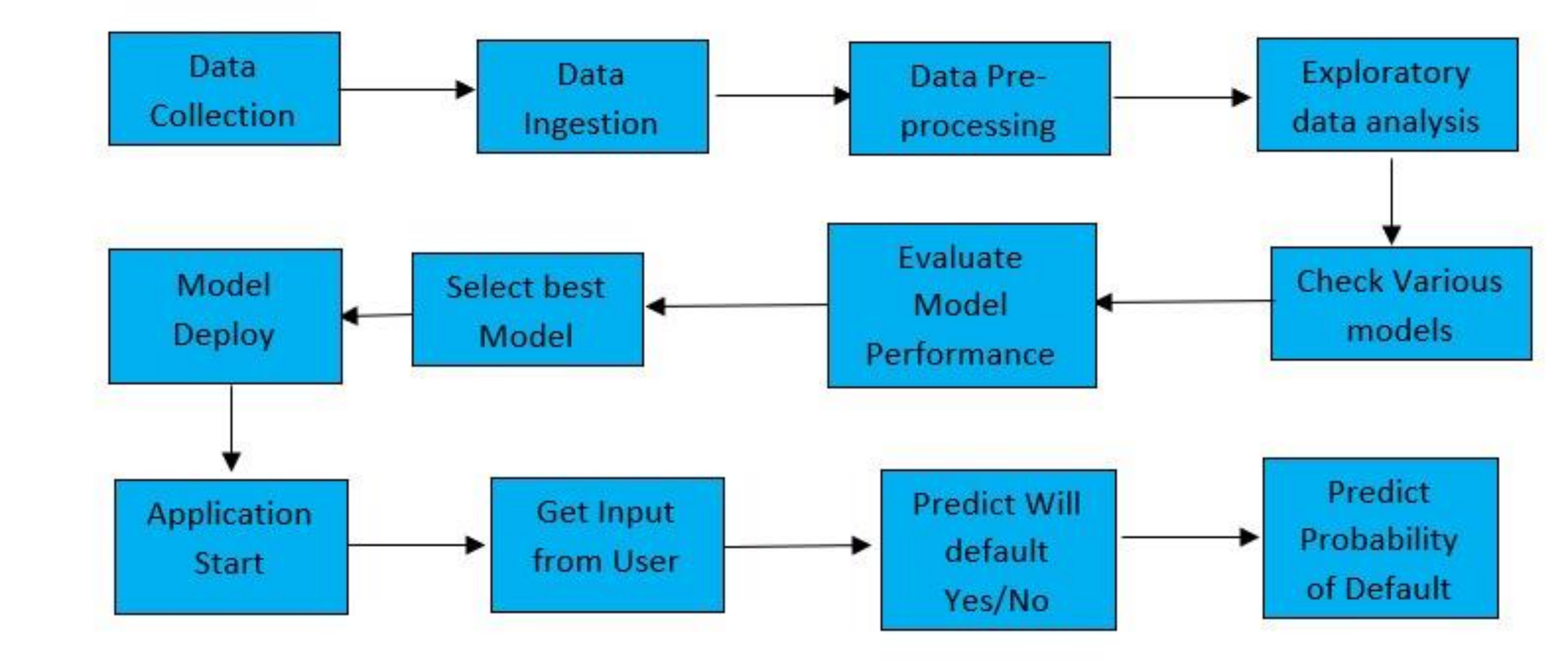
CREDIT CARD DEFAULT PREDICTION

Utkarsh Gaikwad

INTRODUCTION

- 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.

ARCHITECTURE



DATASET

- This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

- DATA SET LINK-
- https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

Variable Name	Measurement Unit	Description
LIMIT_BAL	NT Dollar	Amount of given credit
SEX	Integer	Gender (1=male, 2=female)
EDUCATION	Integer	1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown
MARRIAGE	Integer	Marital status (1=married, 2=single, 3=others)
AGE	Years	Age of the person in years
PAY_0-6	Integer	Repayment status for various months
BILL_AMT1-6	NT Dollar	Amount of billed statements for various months
PAY_AMT1-6	NT Dollar	Amount of Previous payments done
default.payment.next.month	Binary	Will Customer default? Yes/No

DATA PRE-PROCESSING

- Drop the column which is statistically not Important, here I dropped the ID column
- Convert variables: SEX, EDUCATION, MARRIAGE into object as they are categorical variables
- Separate Categorical and Continuous variable.
- For Continuous variables Scale the model with StandardScaler or MinMaxScaler if necessary for model. Scaling is not necessary for tree-based models.
- Perform One-Hot Encoding on Categorical variables
- Join Continuous and One Hot Encoded Variables
- Data Pre-processing is done.

MODEL BUILDING

- I have built various classification models like – Decision Tree, Random Forest, XGBoost, K-Nearest-Neighbours, MLP(Multi Layer Perceptron).
- Base model of each above model was created
- Hyperparameter tuning for each model was done using 4-Fold GridSearchCV
- Model with best accuracy score and which required less training time was selected
- Classification report of Models was also checked

MODEL EVALUATION

- Metrics used for Evaluation are : Accuracy Score and Classification Report
- A Classification report is used to measure the quality of predictions from a classification algorithm.
- How many predictions are True and how many are False
- More specifically, True Positives, False Positives, True negatives, and False Negatives are used to predict the metrics of a classification report.

BEST MODEL - XGBOOST

Training Accuracy	0.8236
Testing Accuracy	0.8230

Parameters	Value
n_estimators	500
Learning_rate	0.01
max_depth	4

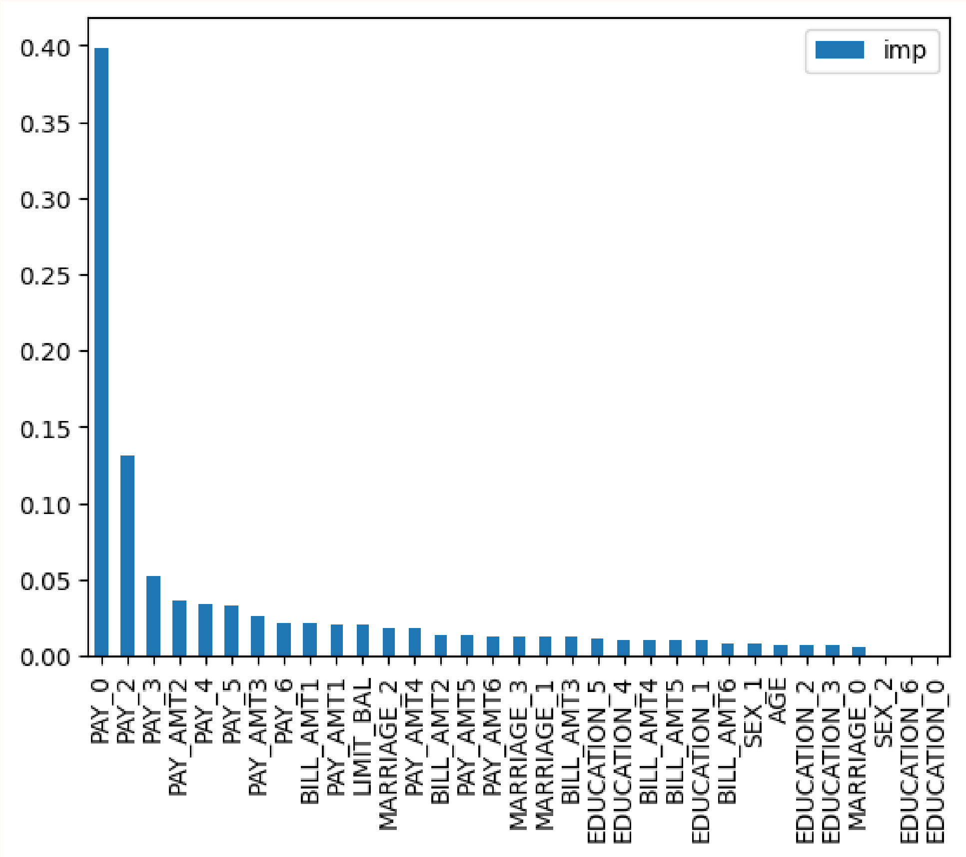
Classification Report of Testing Data

```
              precision    recall  f1-score   support

           0       0.84      0.96      0.89      4654
           1       0.71      0.36      0.48      1346

    accuracy                           0.82      6000
   macro avg       0.77      0.66      0.69      6000
weighted avg       0.81      0.82      0.80      6000
```

FEATURE IMPORTANCE



DEPLOYMENT

- Saved the XGBoost Classifier Pickle file.
- Used Dash library to create Front-end
- Used @app.callbacks decorator for User Interaction
- Created a Procfile for deployment using Gunicorn
- Committed All project to GitHub Repository
- Deployed the Web-Application to Heroku
- Web-App URL : https://creditcarddefaultutkarshg.herokuapp.com/
- Project Demo URL: https://youtu.be/ltktrNRSWEU