Credit Card Default Prediction

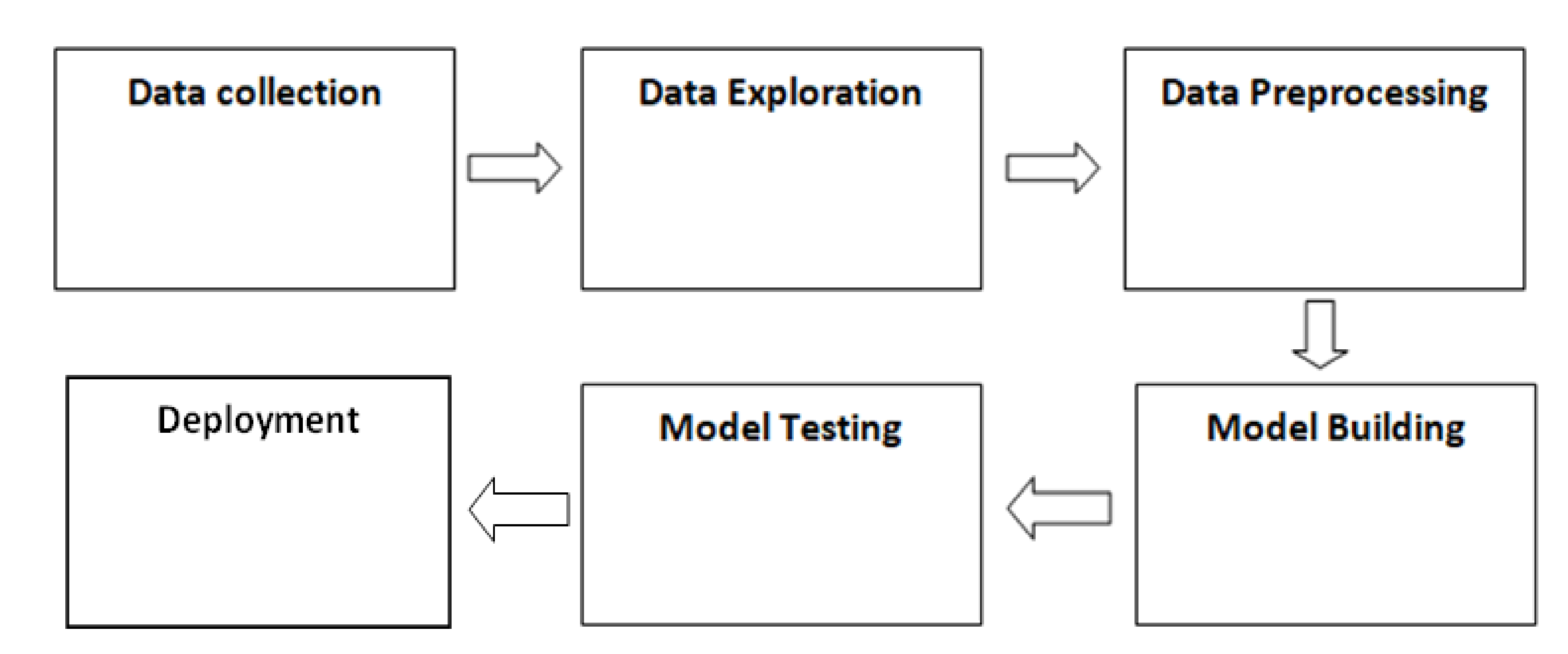
Objective:

To develop a predictive model for detecting credit card defaulters. The model will determine whether a customer is a defaulter or not.

Benefits:

- Detection of future defaulters.
- Gives better insight into customers behaviour.

Methodology



Model Training:

- Dataset

  The dataset for the task is downloaded from Kaggle in csv format for model training

- Data Preprocessing

  - Performing data exploration to get insight of data like understanding trends in the data etc.
  - Impute null values if they are present
  - Encode categorical values so that they can be understood by the models.
  - Scaling down values using Standard Scalar

Classification –

- After data preprocessing, the data is fed into each of the models for prediction.

- First the data into split into train-test sets in 80:20

- It was noticed that the dataset was very imbalanced and to make the classification fair, the data in the training dataset was balanced using SMOTE so that the minority class comes in level with the majority class.

- Then the training dataset was fed into different machine learning models like Logistic Regression, Random Forest, Decision tree, SVM and AdaBoost for training

- The models were then tested on the testing set to find the model with highest accuracy of prediction

Prediction

- It was found that Random Forest had the highest accuracy of 77% among the proposed machine learning models
- The best model was pickled to be used in further cases and for deployment of the model as an API to be used for prediction.

Q & A:

Q1) What’s the source of data?

The data for training can be obtained from Kaggle in csv format

Q 2) What was the type of data?

The dataset was a combination of numerical and Categorical variables.

Q 3) What’s the complete flow you followed in this Project?

The project started with data collection, then data exploration, data preprocessing, model building, model training and finally deployment

Q 4) What techniques were you using for data pre-processing?

- Visualizing relation between the dependent and independent variables
- Cleaning data and imputing if null values are present.
- Converting categorical data into numeric values.
- Scaling values using Standard Scalar.

Q 5) How training was done or what models were used?

- The dataset was split into train and test sets in 80:20
- SMOTE was performed to address the class imbalance issue
- Algorithms like Logistic regression, Random Forest, Decision tree, SVM and AdaBoost were used for classification and training

Q 6) How Prediction was done?

After training the models, the models were tested with the testing data where the actual prediction is done and this also outputs the model with the highest accuracy

- Q 7) What are the different stages of deployment?
  - The model was deployed as an API using FastAPI where the user can input values to the relevant variables to detect if the customer is a credit card defaulter or not