iNeuron

# PROJECT LLD

| | |
|---|---|
| **Project Title** | Credit Card Default Prediction |
| **Technologies** | Machine Learning Technology |
| **Domain** | Banking |
| **Project Difficulties level** | Intermediate |
| **Submitted By** | Reeshma Ram Prasad |

## 1. INTRODUCTION

Banks provide loans and credit cards to their customers, allowing them to make purchases and pay later. However, an increasing number of credit card users are defaulting on their payments, which poses problems for banks in terms of profitability and trust from investors and stakeholders. One solution to this problem is to identify potential credit card defaulters ahead of time and implement measures to mitigate the risk of default.
This can be achieved by using machine learning algorithms to identify potential defaulters before they default. By analysing the financial history and behaviour of credit card users, banks can develop predictive models that can identify customers who are at high risk of defaulting on their payments. Once potential defaulters are identified, banks can take steps to mitigate the risk of default, such as by requiring these customers to provide additional collateral or by imposing stricter limits on their credit card usage. By taking these measures, banks can protect their profitability and maintain the trust of their investors and stakeholders.

## 2. PROBLEM STATEMENT

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 faces 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.

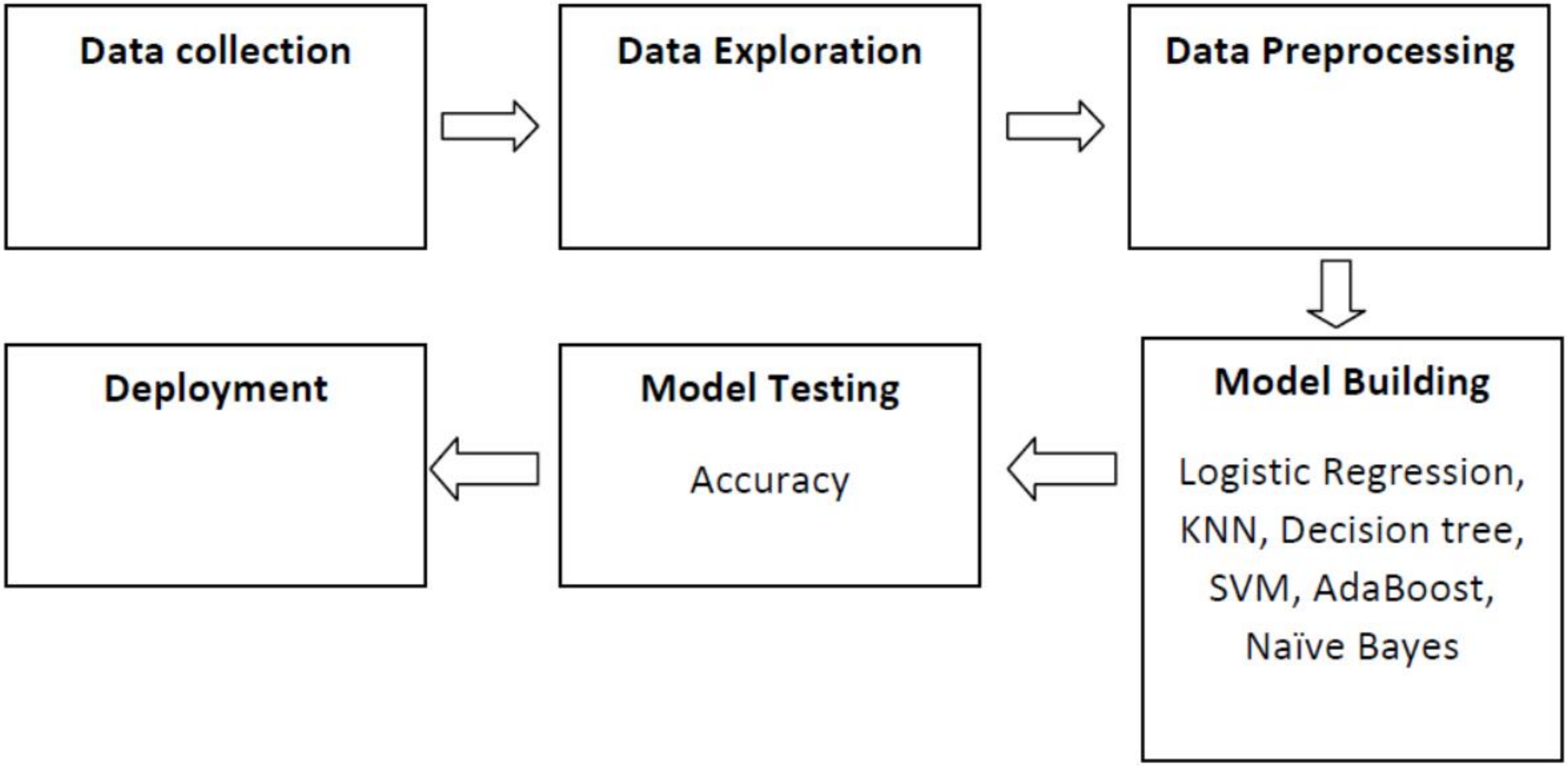
## 3. DATASET COLUMNS DESCRIPTION

Column 1 - ID: ID of each client
Column 2 - LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary = credit)
Column 3 - SEX: Gender (1=male, 2=female)
Column 4 - EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)

Column 5 - MARRIAGE: Marital status (1=married, 2=single, 3=others)
Column 6 - AGE: Age in years
Column 7 - PAY_0: Repayment status in September, 2005 (-1=pay duly, 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)

Column 8 - PAY_2: Repayment status in August, 2005 (scale same as above)
Column 9 - PAY_3: Repayment status in July, 2005 (scale same as above)
Column 10 - PAY_4: Repayment status in June, 2005 (scale same as above)
Column 11 - PAY_5: Repayment status in May, 2005 (scale same as above)
Column 12 - PAY_6: Repayment status in April, 2005 (scale same as above)
Column 13 - BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
Column 14 - BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
Column 15 - BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
Column 16 - BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
Column 17 - BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
Column 18 - BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
Column 19 - PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
Column 20 - PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
Column 21 - PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
Column 22 - PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
Column 23 - PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
Column 24 - PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
Column 25 - default.payment.next.month: Default payment (1=yes, 0=no)

## 2. Architecture


Data collection
Data Exploration
Data Preprocessing
Model Building
Logistic Regression, KNN, Decision tree, SVM, AdaBoost, Naïve Bayes
Model Testing
Accuracy
Deployment

## 3.1 Architecture Description

### 3.1.1 Data Description

The dataset used for the project the UCI Credit Card dataset which is in a .csv format consisting of 30000 rows of data relating to different credit card clients in Taiwan and 25 variables like demographic factors, history of payment, credit data and bill statements from April 2005 to September 2005.

### 3.1.2 Import Data

Data is stored and imported to Python in CSV format which is then used for data pre-processing and model training and testing

### 3.1.3 Data Pre-processing

The dataset was fairly clean as it did not contain any null values and categorical variables Gender, Education etc were already encoded. Before building the models, the data was normalized using Standard scaler so that variables in the dataset are within a certain range. Feature scaling helps to keep the comparison between variables on common grounds. There was an imbalance in the dependent variable namely, the possibility of defaulting and this sorted by using the SMOTE method were samples in the minority class were synthetically oversampled to the level of the majority class to account for a fair classification.

### 3.1.4 Splitting the data

The dataset was split into train-test sets in 80:20 with the independent variables being the different features and the dependent variable being the possibility the client will default.

### 3.1.5 Model Building

After cleaning and splitting the data, the machine learning models were built to understand which model best predicts the defaulter given a set of features. The algorithms built for the task were Logistic Regression, KNN, Decision tree, AdaBoost, Naïve Bayes and SVM which were then trained on the training dataset.

### 3.1.6 Model Testing

The models were then tested on the testing dataset and the model that predicted the dependent variable with the highest accuracy was treated as the best model.

### 3.1.7 Deployment

The model will be deployed as an API using FastAPI where the user can input the relevant values to find if the customer is a credit card defaulter or not