CREDIT CARD DEFAULT PREDICTION

DETAIL PROJECT REPORT

CARD INSIDER

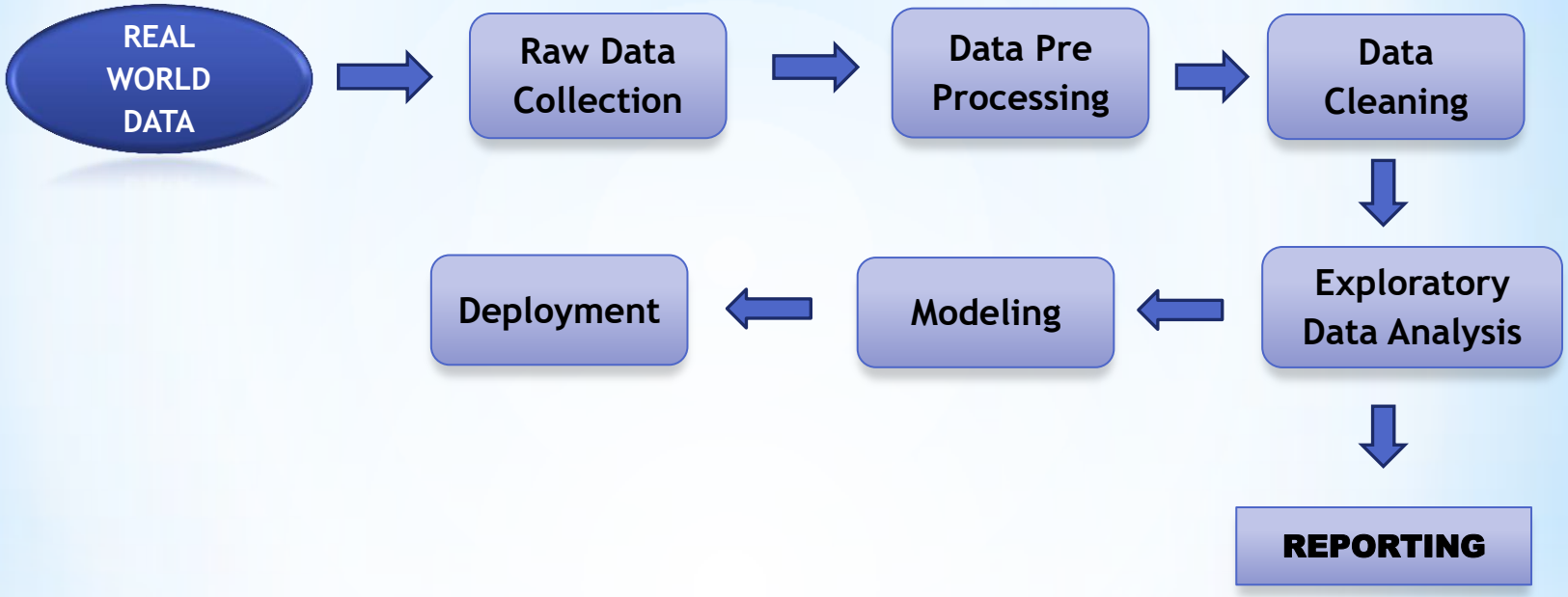
OBJECTIVE:

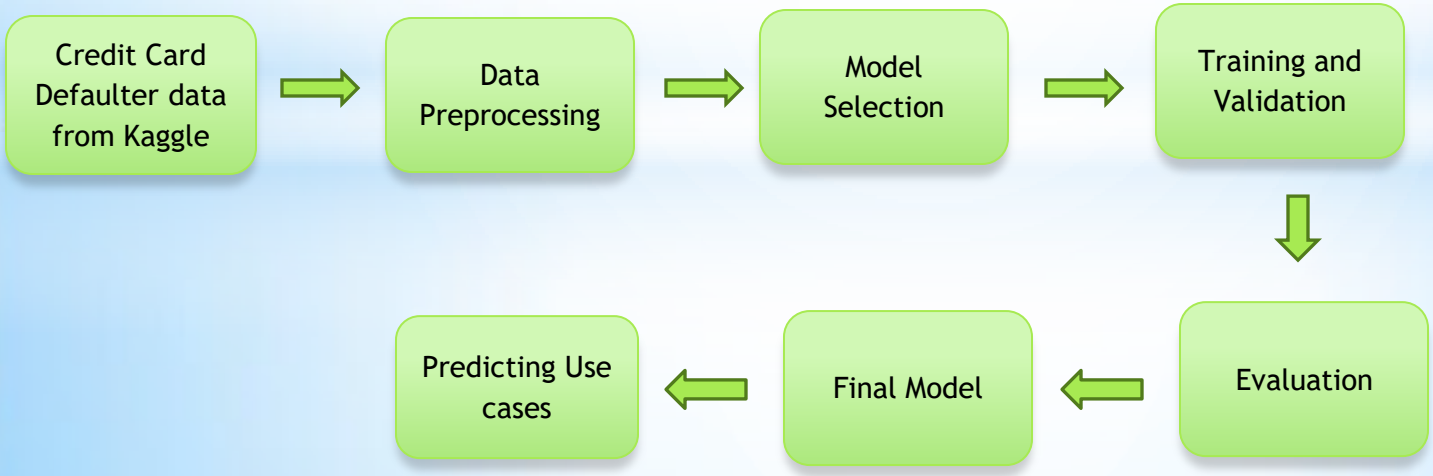
The goal of this project is to predict the probability of the Credit card Defaulter, based on the information of default payments, demographic factors, credit data, history of payment and bill statement and many others of the credit card clients .

Problem Statement

To achieve the goal, we used a data set that is formed by taking into consideration some of the information of 30000 credit card holder in Taiwan in which the dataset contains the transaction details 30000 card holders from April 2005 to September 2005.The problem is based on the given information about each individual we have to calculate that whether that individual with Defaulter or not a defaulter for next month

ARCHITECTURE





Dataset Information :

There are 25 variables:

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, 4=others, 5=unknown, 6=unknown)

MARRIAGE: Marital status (1=married, 2=single, 3=others)

AGE: Age in years

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)

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)

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)

MODEL TRAINING

Data Preprocessing

Our dataset consists of 30000 rows with 25 columns among default.next.month.payment is the target variable, among those only 24 columns are find to be necessary next process. When we observe that our target column is imbalance by using SMOTE we balance the data where it will not affect the accuracy of the model along with standardization is also done for better accuracy.

MODEL BUILDING

By doing all the preprocessing of the data our data is ready for model building, we applied Logistic Regression, Random Forest, Decision Tree, XGBM among all these the Random Forest gives the better accuracy and also f1 score of 81.2% to increase the f1 score we added one more column by taking average of all bill amount, by adding this column the f1 score has increased to 83.2% which is quite good score for predicting the model.

MODEL DEPLOYMENT AND PREDICTION

- Using Syder IDE with python as frontend and streamlit as backend we deploy our model in local system and push the code to Heroku cloud platform which provide paas as a service
- Data has been taken from the user
- Data preprocessing and validation techniques are applied
- Predicting the result

CONCLUSION

The Classification Algorithm is used is used to identify whether the customer is defaulter or not by providing the details of 6 months transaction.

Q & A

❑ What is the source of the data?

The Dataset was taken from iNeuron Provided Project Description Document.

https://drive.google.com/file/d/1AGRq2hG8zUbM_8LCo48cbcYy6W-2ujZC/view?pli=1

❑ What are the type of data?

Provided data are of Categorical and Numerical in which the categorical columns are already encoded

❑ What's is the complete followed in this project

Refer slide 4 which provide the complete guide of the project flow

❑ What techniques using for data preprocessing?

- ➢ Checking the null and duplicate values
- ➢ whether our data is balanced or not
- ➢ For balancing the data we applied SMOTE method
- ➢ Finally get the data in the standardization form

❑ How training was done or what models were used?

After the data preprocessing we try to apply Logistic Regression, Decision Tree, Random Forest, XGBM among all those Random Forest gives the better accuracy and f1 score so we finalize the Random Forest . Here we build the model for default parameter, when we try to apply the Hyper parameter tuning the model performance is same as that of default models, so we are not done any hyper parameter tuning.

❑ How was the prediction done?

getting the inputs from the user to predict the whether the customer will be default for next month or not

❑ What are different stages of deployment?

when the model is we try to deploy in local environment, using streamlit and try to check prediction accuracy , later push the code to the cloud called Heroku for public platform