Credit Card Defaulters Prediction using Machine Learning models

Yaswanth Reddy Soma   Sai Hemanth Reddy Kesamreddy   Sahithi Vakulabharanam

University of Michigan - Dearborn

IMSE 586

Introduction

Credit cards are the most popular method of payment these days, which offer free pre-approved credit usage from the financial institutions. Thus, the repayment is a crucial factor which needs to be analyzed. The main aim of this project is to predict the credit card defaulters based on several factors which contains Credit limit ,education etc. by using various machine learning algorithms and to find the best machine learning model with highest prediction accuracy.

Dataset

The link for the data set can be accessed through this link: `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`

The default of credit card clients dataset contains payment data of the customers from an important bank in Taiwan.Dataset has a total of 23 explanatory variables and the default payment as the response variable which is binary (Yes=1, No=0).

An overview of the default payment of the customers in the dataset is given below.

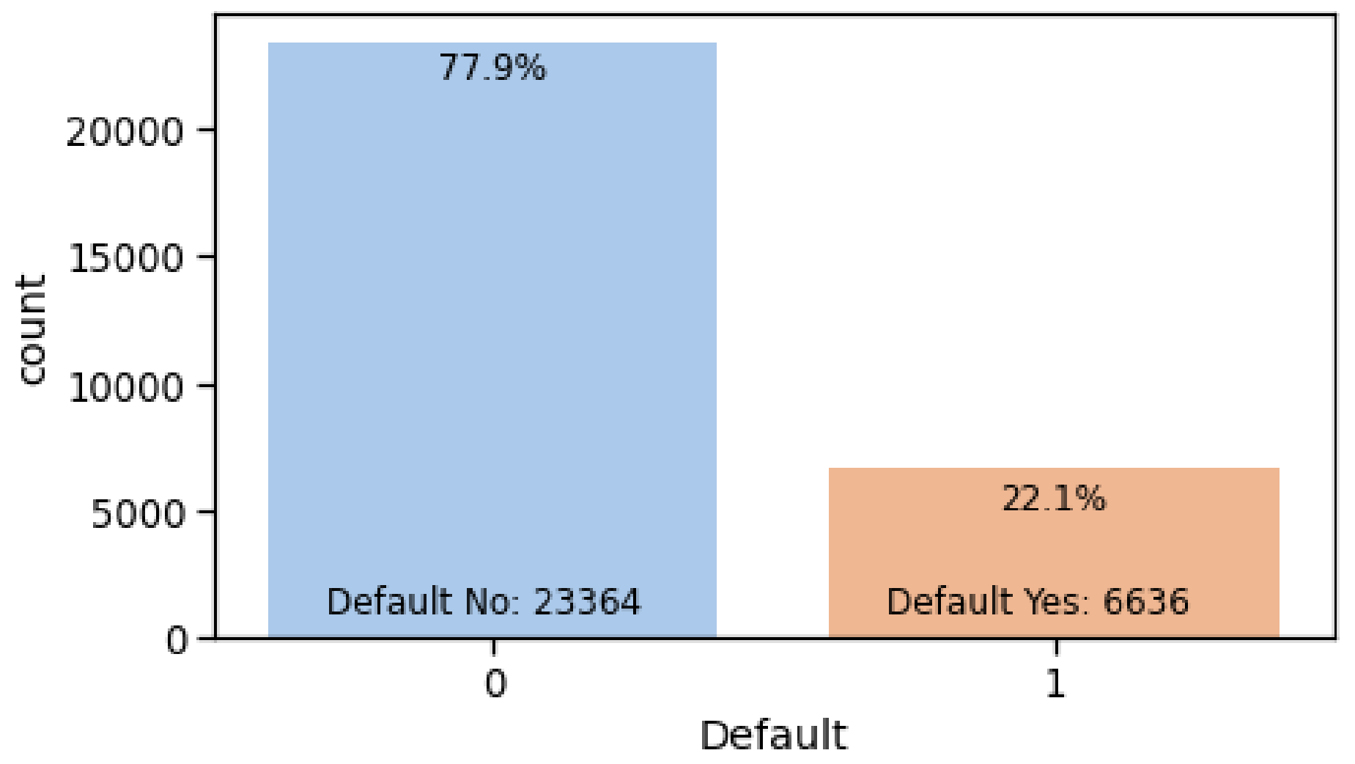

Figure 1. Bar graph showing the default count

Methodology

1. Data Preprocessing and Analysis: We first collect the data and import it into a dataframe to check for the irregularities, We then Pre process it for further study.
2. Feature engineering: Repeated categories in the dataframe columns are handled and make sure that data is ready for Explanatory data analysis.
3. Explanatory data analysis: Relations between different variables in the data are obtained for a clear understanding of the trends in imported dataset.
4. Standardization: All the various columns are in different scales and units, so standardization is done for unbiased prediction by eliminating the mean and scaling to unit variance.
5. Model Building and Evaluation: First, we split data in the ratio of 80:20 where 80 percent of the data is used for training the model and remaining 20 percent of the data is used for testing the model. Different analysis techniques are used to find the best accuracy prediction method for the data.

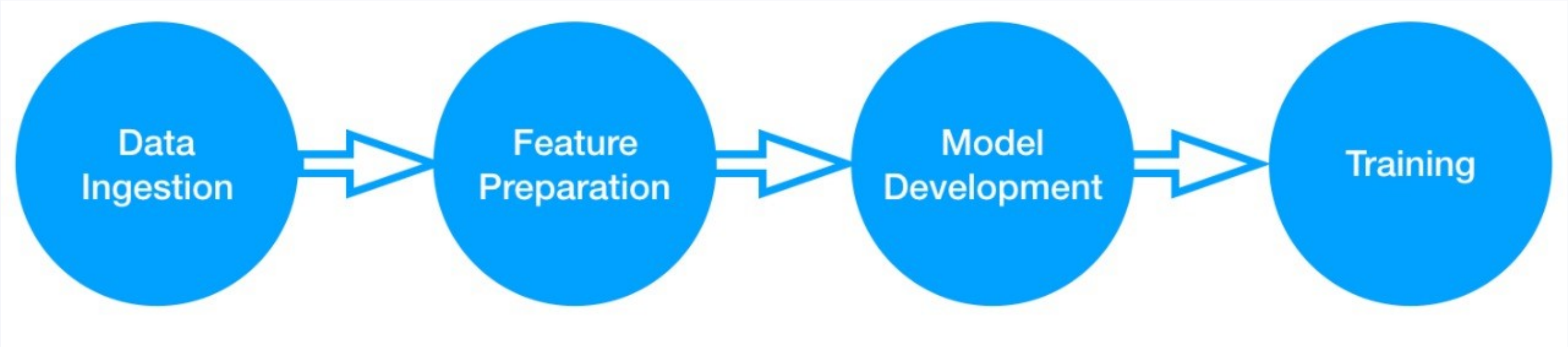

Figure 2. Flowchart expressing the process

Explanatory Data Analysis

In order to summarize the main characteristic and analyze the data, Explanatory data analysis is done and a good illustration of different relations is drawn.

A countplot is done for all the categorical variables, where the number of default customers for all those variables are visualized.

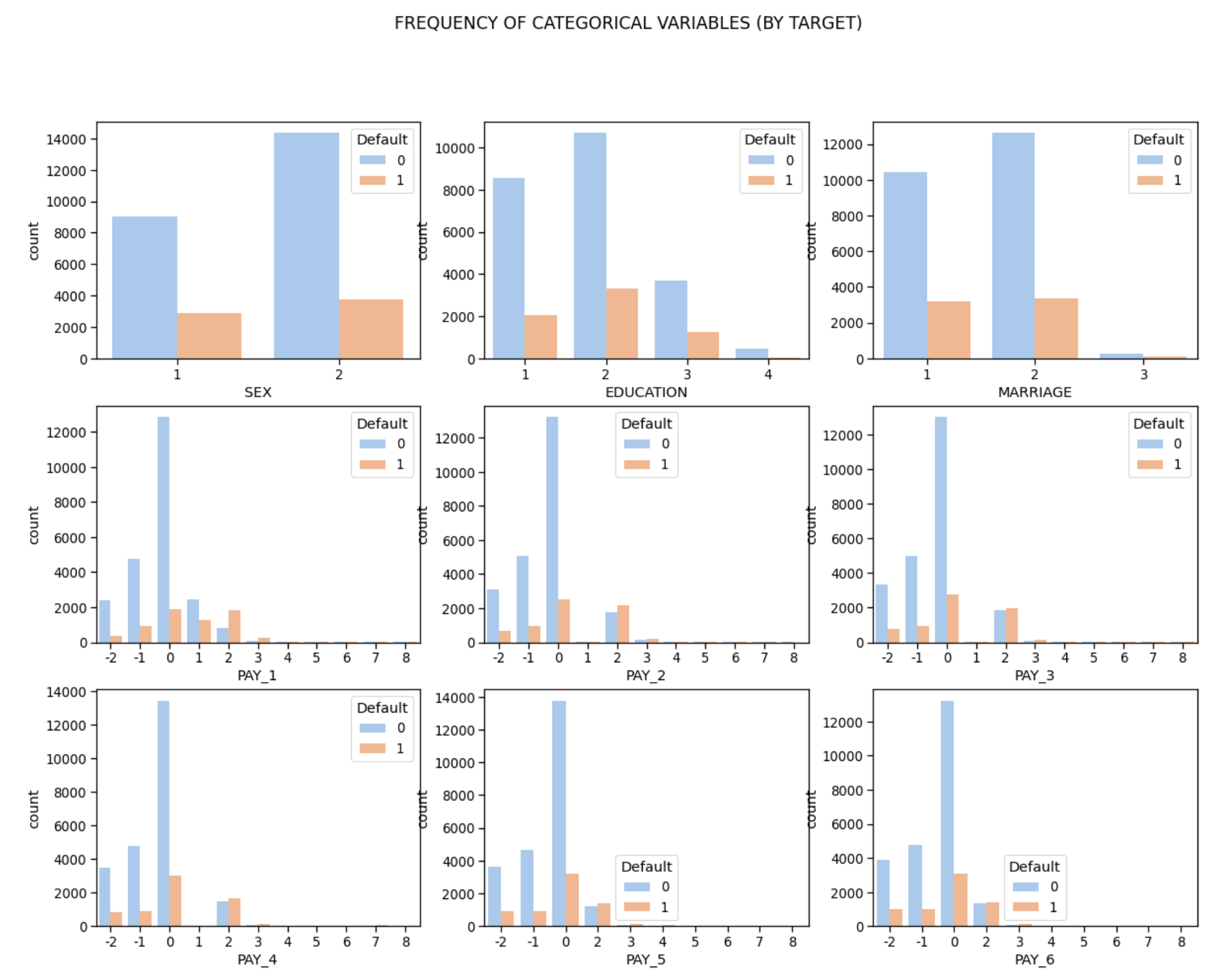

Figure 3. Bar graph showing trends in categorical variables

To further analyze the payment pattern of the customers a scatter plot is developed for the bill amount and the payment of the bill for the past 6 months.

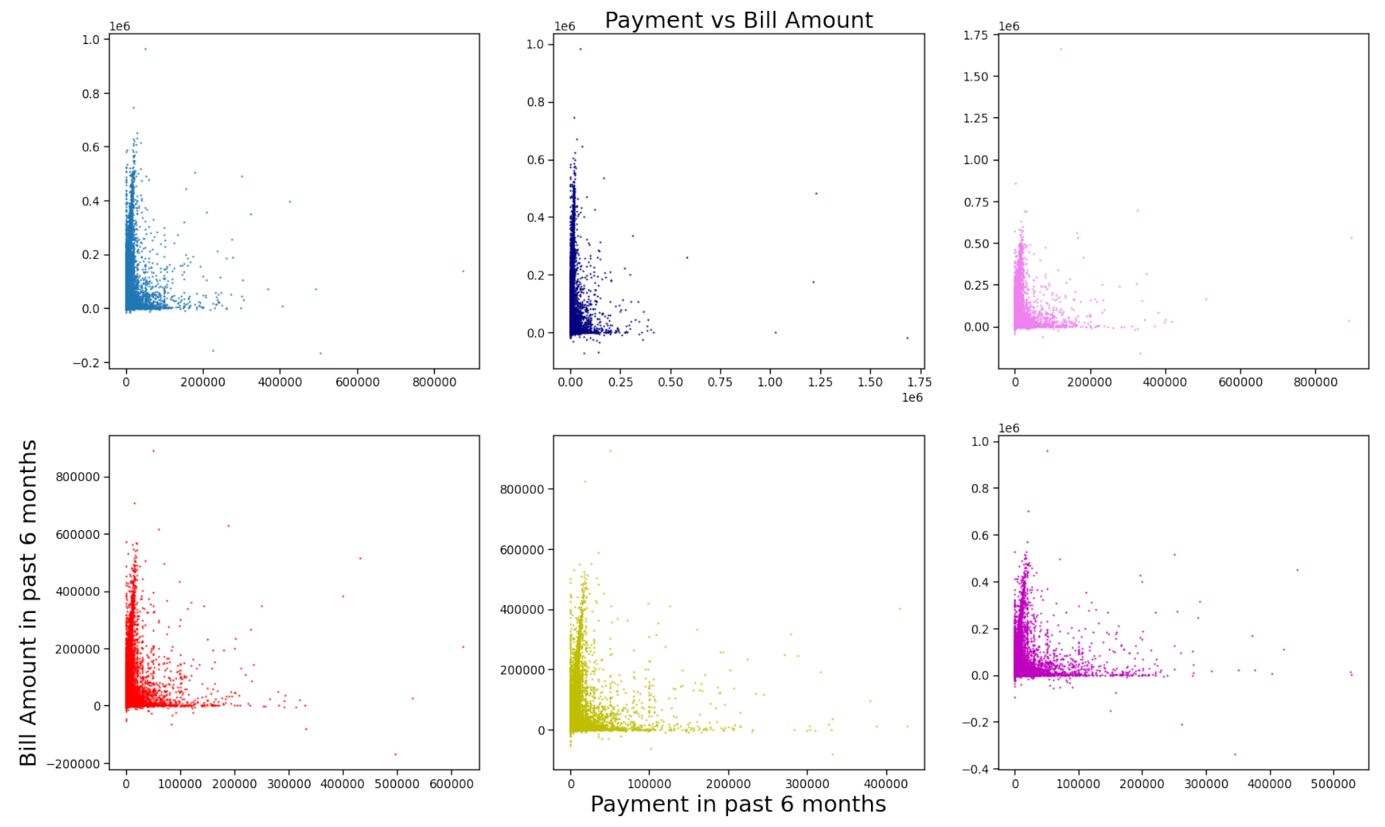

Figure 4. Scatter plot for Payment vs the bill amount

- For Sex we found that Males have lesser chance for default than females and when it comes to marriage Single persons has the least chance of default.
- In Education University student has high chances of default.
- When the delay in repayment increases the chances of getting default increases significantly.
- A heat map is derived for the various columns of the data where we found the correlation between them. The Pay columns and the bill-amount has a positive correlation among them

Feature Selection

The top 10 features of the model are selected for the best prediction of the result using the Recursive feature elimination method.

These features include 'SEX', 'EDUCATION', 'MARRIAGE', 'AGE', 'PAY1', 'PAY2', 'PAY3','PAY4', 'PAY5', 'PAY6'.

Machine Learning Models

Machine learning pipelines are used to build the models and to perform the hyper parameter tuning using GridSearchCV, which helps in finding the best parameters which gives us the highest frequency. We have used various parameters for each model along with two different standardization methods StandardScaler and MinMaxScaler.

Logistic Regression

Here, we use C value as a hyperparameter, where e gave multiple values for C between 0 to 100 to get the best possible C value.The best parameters after tuning are MinMaxScaler with a C value of 10.[1]

Decision Tree Classifier

Here,Max depth is used as a hyperparameter. We load multiple values for max depth between 1 to 10 to get the best possible max depth value.The best parameters after tuning are StandardScaler with a Max depth value of 3.

K-Nearest Neighbor Classifier

Here,n-neighbors and metric are used as hyperparameters, where we gave multiple values for n-neighbors between 1 to 22 and for metric we gave Euclidean and Manhattan .The best parameters after tuning are StandardScaler with a n-neighbors value of 17 and metric as euclidean.

Random Forest Classifier

N-estimators and max depth are used as hyperparameters, range of n-estimators is between 10 to 1000 and for max depth it is between 3 to 15.The best parameters are max depth value of 10 and n estimators value is 500 and no standardization(passthrough).

XGBoost Classifier

N-estimators and max depth are used as hyperparameters, range of n-estimators is between 10 to 1000 and for max depth it is between 3 to 15.The best parameters are max depth value of 3 and n estimators value is 500 and and StandardScaler.

Results

Model	Accuracy	Precision	Recall	f1-score
Logistic Regression	0.81	0.72	0.24	0.36
Decision Tree Classifier	0.82	0.67	0.38	0.49
K-Nearest Neighbors Classifier	0.82	0.65	0.37	0.47
Random Forest Classifier	0.82	0.68	0.37	0.48
XGBoost Classifier	0.82	0.68	0.37	0.48

Table 1. Performance metrics of various Machine learning models

Since the data is highly imbalanced, accuracy is not a good metric to evaluate the model .So, we should consider F1-score as the model evaluation metric.Decision Tree Classifier is the best performing model with an accuracy of 82 percent and highest f1-score of 0.49 among all the models.

References

[1] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480, 2009.