Credit Card Default Prediction

Team member- Vikas Kumar

Abstract- This project on ‘ Credit Card Default Prediction’ is using Machine learning. In this project, I had given the dataset for credit card default prediction. I have to predict whether the card holder will default the payment or not. I have 25 features here in total. And the feature “defaulter_payer” is my target feature. This is my classification project.

Keywords- Machine learning algorithm, Classification, Python.

Problem Statements:

1. Introduction
2. Data Wrangling
3. Data Visualization
4. Pairplot
5. Correlation between features of the dataset
6. Sample Data
7. Data cleaning
8. Feature Scaling
9. Applying train test split
10. Applying Machine Learning Algorithm for Classification Problem
    - Logistic regression
    - Decision tree
    - Random Forest
    - Stochastic Gradient Descent
    - K-Nearest Neighbor
    - Support Vector Machine
11. Grid Search CV on all Models
12. Conclusion

Introduction :

The credit card companies in Taiwan faced a cash and debt crisis in 2005, with a peak in delinquency anticipated for the third quarter of 2006. (Chou). Taiwan's card-issuing banks over-issued cash and credit cards to unauthorized applicants in an effort to gain market dominance. In addition, most cardholders, regardless of their capacity to pay back, abused their credit cards for consumption and racked up substantial cash and credit card debt. Consumer financial confidence was damaged by this crisis, which also provided major challenges for cardholders and banks.

1. Data Descriptions

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:

- X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6 - X11: History of past payment. I tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -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.
- X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
- X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 =

amount paid in April, 2005.

Steps involved in this Project

Step 1:

In the first step, I wrote the code for data wrangling. Data Wrangling is the process of gathering, collecting, and transforming Raw data into another format for better understanding, decision-making, accessing, and analysis in less time.Data Wrangling is also known as Data Munging.

Step 2:

In the second step, I wrote the python programming to find some of the results. I had different types of dataset in which there were different columns. So I extracted outcome insight from those data.. Then I wrote the programming to draw the graph.

Step 3:

In the third step, I divided the all dataset in sampling data. I took a 6000 dataset for sampling in this project. I took a total 20 percent of the dataset.

Step 4 :

In the fourth step, I cleaned our dataset for getting better insight from given data.

Step-5:

In this step, I wrote the program for feature scaling.
It refers to putting the values in the same range or same scale so that no variable is dominated by the other. It is mostly used in the categorical data where the categories are assigned simple integers such as 0,1,2… which might represent different categories.Here, I am using Z score normalization calculates the z-score of each value and replaces the value with the calculated Z-score.

Step-6:

In the sixth step, I splitted my whole dataset into training and test dataset.
I take 70 percent of the data for training and 30 percent for the test dataset.

Step-7:

In the last step, I applied different machine learning algorithms.

- I applied here a total of six machine learning algorithms. The names of those algorithms are given below-
  - Logistic regression
  - Decision tree
  - Random Forest
  - Stochastic Gradient Descent
  - K-Nearest Neighbor
  - Support Vector Machine

I also used GridsearchCV for hypertuning to get the better result.

Conclusions :

1)Using a Logistic Regression classifier, I can predict with 81.6% accuracy, whether a customer is likely to default next month.

2)Using a Decision Tree classifier, I can predict with 81.5% accuracy, whether a customer is likely to default next month.

3)Using a Random Forest classifier, I can predict with 81.33% accuracy, whether a customer is likely to default next month.

4)Using a Stochastic Gradient Descent classifier, I can predict with 81.7% accuracy, whether a customer is likely to default next month.

5)Using a K-Nearest Neighbor classifier, I can predict with 80.7% accuracy, whether a customer is likely to default next month.

6)Using a Support Vector Machine classifier, I can predict with % accuracy, whether a customer is likely to default next month.

- The strongest predictors of default are the PAY_X (i.e. the repayment status in previous months), the LIMIT_BAL & the PAY_AMTX (amount paid in previous months).
- I found that we are getting best results from Stochastic Gradient Descent and then Logistic regression.
- The credit limit is a good indicator of financial stability. Whatever

mechanism the bank is currently using works well and some of the features that go into choosing the credit line can be used directly in the model for default prediction.

Demographics:- I see that being Female, More educated, Single and between 30-40 years old means a customer is more likely to make payments on time.