Credit Card Default Prediction

Rushabh Tikale

Data Science Enthusiast

Almabetter, Banglore.

Content

TITLE PAGE

1. Abstract 0-1
2. Problem Statement 1-3

3. Introduction 3-4

4. What is credit card default? 6-7

5. Steps Involved 7-9

6. Types of model used 9-13

9. Evaluation Metrics 13-14

10. Conclusion 14-15

1.Abstract:

The credit card default is one of the major issues increasing day by day, to solve this we develop a ML model .ML model help us to catch the credit card default customer and try to reduce the fraudster customer.

The data show us the behavior of customer as like payment history for different months, how much the bill is come and how much they paid.

We try to catch the fraudster and the bank does not repeat the mistake.

If the customer does not paying their installment month to month so it has the probability that customer will default for their transaction and loose their claim from credit card.

In short **Credit Card Default Prediction** is to build a model to identify whether the**credit card** applicant will default or not based on his repayment history and other important factors. In order to achieve this, we need to develop a supervised machine learning model using classification algorithms.

*Keywords: Data cleaning, Data analysis, Train & test model, Evaluation metrics, Model selection.*

2.Problem Statement

The Model is aimed at predicting the case of customers default payments. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. We evaluate which customers will default on their credit card payments**.**

The model should have a good accuracy as the default customer will not leave from catching.

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:

* ID - The unique ID of the customer.
* LIMIT\_BAL - Your credit limit is the maximum amount you can borrow using a credit card or line of credit.
* SEX – The sex of the customer i.e Male, Female and others.
* EDUCATION – Customer classified according to education category as like graduate school, university, high school and others.
* MARRIAGE – The marital status as like married, single and others.
* AGE – The different ages of customer.
* PAY\_0 – The payment status for September 2005.
* PAY\_2 - The payment status for August 2005.
* PAY\_3 – The payment status for July 2005.
* PAY\_4 – The payment status in June, 2005.
* PAY\_5 – The payment status in May, 2005.
* PAY\_6 – The payment status in April, 2005.
* BILL\_AMT1 – Amount of bill statement in September, 2005.
* BILL\_AMT2 – Amount of bill statement in August, 2005.
* BILL\_AMT3 – Amount of bill statement in July, 2005.
* BILL\_AMT4 – Amount of bill statement in June, 2005.
* BILL\_AMT5 – Amount of bill statement in May, 2005.
* BILL\_AMT6 – Amount of bill statement in April, 2005.
* PAY\_AMT1 – The total payable amount of the bill for the month of September, 2005.
* PAY\_AMT2 – The total payable amount for the month of August, 2005.
* PAY\_AMT3 – The total payable amount for the month of July, 2005.
* PAY\_AMT4 – The total payable for the month of June, 2005.
* PAY\_AMT5 – The total payable for the month of May, 2005.
* PAY\_AMT6 – The total payable for the month of April, 2005.
* DEFAULT – The binary number for the default customer, 1 – Yes and 0 – No.

3.Introduction:

**Credit Card Default Prediction** is to build a model to identify whether the**credit card** applicant will default or not based on his repayment history and other important factors. In order to achieve this, we need to develop a supervised machine learning model using classification algorithms.

4.What is default credit card?

The customer who are not able to pay their due payment monthly is called as default customer. The limit is provided for each credit card and the transaction above the limit can trigger into penalty rate. Your credit limit might not stay the same the entire time you have the credit card. If you use your credit card wisely and make your monthly payments on time, you can be approved for periodic credit limit increases, sometimes without requesting them.

5.Steps involved:

* **Fill Missing value**  
   After loading the dataset we performed this method by replacing

nan values with zero. This procedure giving us a approach to tackle with the null value and making a foundation strong.

* **Data cleaning**

In this process we convert all string contained data to numeric data. As instead of this we will unable to do the EDA on the data of play store.

* **Exploratory data analysis**

In this procedure we simultaneously work with each features from our dataset and clearly visualize each and every point of aspects. This gives a graphically representation of entire dataset.

* **Extraction of un-useful data**

In entire data some features has un-useful for the data analysis. So it’s better to extract this kind of features from our dataset.

* **Remove Outlier**   
  This procedure helps us to move out with unrelatable part from the   
  dataset. It gives the very clear data and save the time from playing with unrelatable part.
* **Correlation of data**

This procedure helps us to know the relation between the features and useful for time saving in analysis duplicates type of data.

* **Spliting Data to train and test**   
  The model have to fit for training and testing . We split the model as

70% for train and 30% test as it is already decided by the model for under better model act.

* **Import Baseline model**

As we discussed earlier we have to import model such as GB Classifier, KNN, Decision Tree, Random Forest. As it gives us a help to decide the better model for furthure prediction.

* **Use Evaluation Metrics**

The Precision, ROC AUC score, F1 Score are evaluation used to show the residual error between predicted and actual model.

* **Choose the best model**

The final predictor model is the last point for our model then after we are confirm to use the model.

6. Types of model use:

* GB Classifier
* KNN (K-Nearest Neighbours)
* Random Forest
* Decision Tree

1.GB Classifier:

It is used to minimise the loss, or the difference between actual and predicted value.

**How this work?**

Step 1: Import base model which is the average of the independent variable.

Step 2: compute residual error, i.e (actual – predicted)

Step 3: Passed the independent variable and residual error value into the decision tree.

By passing residual value to the base model we get overfit for the model, so to prevent it we use learning rate(alpha).

We simply used GB Classifier to reduce the residual value and prevent from overfiting by boosting the decision tree .

2. KNN (K- Nearest Neighbours)

### kNN classifier determines the class of a data point by majority voting principle. If k is set to 5, the classes of 5 closest points are checked. Prediction is done according to the majority class. Similarly, kNN regression takes the mean value of 5 closest points.

### How does kNN work?

### The kNN working can be explained on the basis of the below algorithm:

### Step-1: Select the number K of the neighbors

### Step-2: Calculate the Euclidean distance of K number of neighbors

### Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

### Step-4: Among these k neighbors, count the number of the data points in each category.

### Step-5: Assign the new data points to that category for which the number of the neighbour is maximum.

### Step-6: Our model is ready.

### 3. Random forest

### From the row and column we are taking the sample and passed to the decision tree model and estimate the majority vote.

### We are using more than one decision tree model, so it might be the chance that some sample from already trained that also can reconsist.

### We are not depending on a single decision tree, and we are reducing the high variance to low variance by taking the majority vote.

### Working For Random Forest:

### C:\Users\DELL\Pictures\Screenshots\Screenshot (36).png

4. Decision tree

**Decision tree**is a type of **supervised learning algorithm** that is

mostly used in classification problems. It works for both categorical

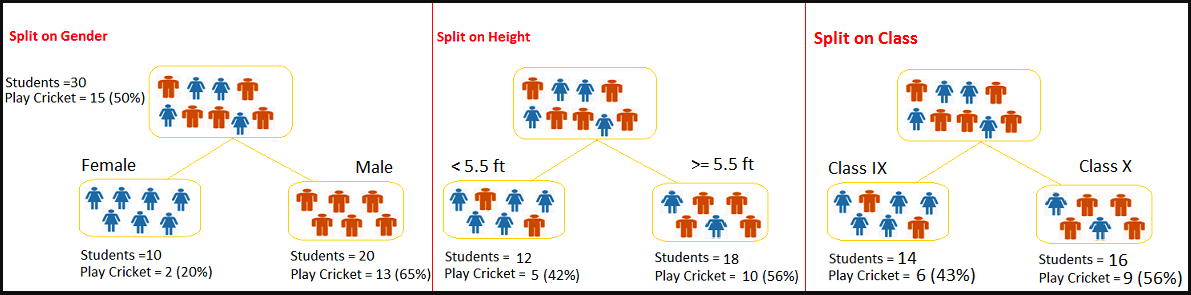
and continuous input and output variables.

* **Working Of Decision Tree Model**

Lets consider the example: There are total 30 students in the class and 15 of the student playing cricket in Leisure time, so we have to make a Decision Tree model for the student who playing cricket in leisure time.

We split the node according to the condition.

In below graph we split the nodes in three different division: split on Gender, Height and Class.



So from the end of the observation we select that data which is toataly robust free data.

7. Evaluation Metrics

* ROC-AUC CURVE

ROC required both TPR(True Positive Rate) and FPR(False Positive Rate). It is the plotting data of TPR and FPR.

AUC is the area under the curve for ROC.

More area under the curve better is the model accuracy.

We select the data as per the threshold value if suppose we need more true positive valur then we select the staring point from the plot.

* PRECISION score

Precision score deal with the positive prediction value.

Precision = Actual Positive / Predict Positive

= TP/TP+FP

* F1 – Score

The F-score, also called the F1-score, is a measure of a model’s accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into ‘positive’ or ‘negative’.

It is convenient to combine the performance of a precision and recall classifier.

The F-score is commonly used for evaluating information retrieval systems such as search engines, and also for many kinds of machine learning models.

* Confusion matrix

The confusion matrix deal with the outcomes of the predictor varaiable.

The outcomes such as True Positive, True Negative, False Positive, False Negative.

It is much better way to evaluate metrics.

* How to create confusion metrics
* Get the set of prediction, so they can be compared to the actual targets.
* Each row in matrix reperesents the actual class, while column represents the predicted class.
* First row is actual positive and second row is actual negative.
* First column predictive positive and second column is predictive negative.

Lets understand by example:

We will make the actual and the predicted variable.

Y-Actual – [0 1 0 1 1 0]

Y-Predicted –[1 1 0 1 0 1]

0 predicted 1 predicted

0 Actual 1 (TP) 2 (FP)

1 Actual 1 (FN) 2 (TN)

8.Conclusion

We train the different model like Decision Tree Classifier, Random Forest Classifier, KNN and Gradient Boosting classifier and predict the score simaltaneously.  
The Decision Tress and Gradient Boosting Classifier model are not giving the best prediction for model and on the another side, KNN and Random Forest giving the best prediction. There is almost same default score for both male and female customers.  
we analyse customer history of use of credit card from different months and there is less default customers who is paying their payment monthly and use as revolving card.  
Choosing Hyperparameter tuning is the case sensitive model so we have to take very much care of it.  
By visualising relation with dependent and independent variable we covered information of default customers by Education, Age, Marriage, Sex. Each feature giving a major information about total number of defaulters in this class.  
In **Random Forest classifier** we are getting the positive outcomes so from this our conclusion is that decision tree is the best predictive model.

Abstract:

CCo