**Credit Card Defaulter Prediction**

**IRSHAD**

**Data science trainees,**

**Alma Better, Bangalore**

**Abstract:**

In this research, a technique for `Credit Card Fraud Detection' is developed. As fraudsters are increasing day by day. And fallacious transactions are done by the credit card and there are various types of fraud. So to solve this problem combination of technique is used like Genetic Algorithm, Behavior Based Technique and Hidden Markov Model. By this transaction is tested individually and whatever suits the best is further proceeded. And the foremost goal is to detect fraud by filtering the above techniques to get better result.

**1.Problem Statement**

### This project is aimed at predicting the case of customers default payments in Taiwan. 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 can use the [K-S chart](https://www.listendata.com/2019/07/KS-Statistics-Python.html) to evaluate which customers will default on their credit card payments

Credit card frauds are increasing heavily because of fraud financial loss is increasing drastically. Every year due to fraud Billions of amounts lost. To analyze the fraud there is lack of research. Many machine learning algorithms are implemented to detect real world credit card fraud. ANN and hybrid algorithms are applied

**2. Introduction**

Recently, the state vigorously promotes the economic construction of large- and medium-sized cities, which not only improves people’s living standards but also changes people’s consumption concept and consumption mode. People are more and more inclined to spend ahead of time and mortgage their “credit” to the bank to enjoy certain things in advance. However, when consuming, people often lack rational thinking and overestimate their ability to repay loans to banks in time

Generally speaking, compared with the credit card customers who have not paid their loans overdue, there are fewer overdue repayments [2, 3]. This variable feature of overdue and overdue loan repayment is called “two classifications” in machine learning prediction. In the prediction of “two classifications,” a few categories are called positive examples (default), and most categories are called counterexamples (non default). However, most of the credit card loan data are unbalanced.

We can see that the problem of category imbalance is mainly solved from the following two perspectives: the first perspective is to balance the data by changing the number of samples. This method can also be divided into three aspects. On the one hand, it is to improve the oversampling method. On the other hand, it is based on the principle of under sampling to change the data distribution. On the third hand, it is the method of combining oversampling and under sampling

## **3.Data Description**

### 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. We 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.

**4.Payment Type**

**PAY0: 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)**

PAY2: Repayment status in August, 2005

PAY3: Repayment status in July, 2005

PAY4: Repayment status in June, 2005

PAY5: Repayment status in May, 2005

PAY6: Repayment status in April, 2005

**5. Steps involved:**

* **Exploratory Data Analysis**

After loading the dataset we performed this method by comparing our target variable that is DEFAULT PAYMENT MONTH with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

* **Null values Treatment**

Our dataset no null values present if the null values are present which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result.

* **Feature Selection**

In these steps we used algorithms like Extra Tree classifier to check the results of each feature i.e which feature is more important compared to our model and which is of less importance.

* **Standardization of features**

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

**Balance the dataset**

By the using oversampling we

balanced over dataset

* **Fitting different models**

For modelling we tried various classification algorithms like:

1. **Logistic Regression**
2. **k nearest neighbour**
3. **Random Forest Classifier**
4. **Naive bayes**
5. **Decision tree**
6. **Hyper parameter tuning on random forest algorithm**

* **Tuning the hyper parameters for better accuracy**

Tuning the hyper parameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of tree based models

like Random Forest Classifier.

* **SHAP Values for features**

We have applied SHAP value plots on the Random Forest model to determine the features that were most important while model building and the features that didn’t put much weight on the performance of our model.

**6.1. Algorithms:**

1. **Logistic Regression:**

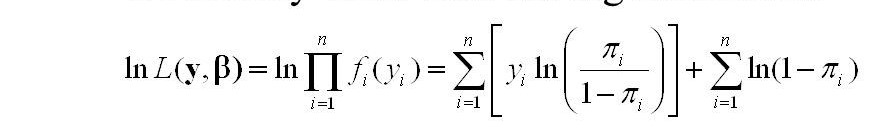
Logistic Regression is actually a classification algorithm that was given the name regression due to the fact that the mathematical formulation is very similar to linear regression.

The function used in Logistic Regression is sigmoid function or the logistic function given by:

f(x)= 1/1+e ^(-x)



The optimization algorithm used is: Maximum Log Likelihood. We mostly take log likelihood in Logistic:



.

**2.K nearest Neighbor**

The K-Nearest Neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

## **How does K-NN work?**

The K-NN 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 neighbor is maximum.
* **Step-6:** Our model is ready.

****

****

**3.Random Forest Classifier:**

Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicted out of all.



**4.Naive bayes-**

The Naive Bayes classification algorithm is **a probabilistic classifier**. It is based on probability models that incorporate strong independence assumptions. The independence assumptions often do not have an impact on reality. Therefore they are considered as naive.

## **Bayes' Theorem:**

* Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
* The formula for Bayes' theorem is given as:

Naïve Bayes Classifier Algorithm

**P(A|B) is Posterior probability**: Probability of hypothesis A on the observed event B.

**P(B|A) is Likelihood probability**: Probability of the evidence given that the probability of a hypothesis is true.

**1.Decision tree**

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

****

**Strengths and Weakness of Decision Tree approach**   
The strengths of decision tree methods are: 

* Decision trees are able to generate understandable rules.
* Decision trees perform classification without requiring much computation.
* Decision trees are able to handle both continuous and categorical variables.
* Decision trees provide a clear indication of which fields are most important for prediction or classification.

**7.2. Model performance:**

Model can be evaluated by various metrics such as:

1. **Confusion Matrix**-

The confusion matrix is a table that summarizes how successful the classification modelis at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.

1. **Precision/Recall**-

Precision is the ratio of correct positive predictions to the overall number of positive predictions : TP/TP+FP

Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: TP/FN+TP

1. **Accuracy**-

Accuracy is given by the number of correctly classified examples divided by the total number

of classified examples. In terms of the confusion matrix, it is given by: TP+TN/TP+TN+FP+FN

1. **Area under ROC Curve(AUC)**-

ROC curves use a combination of the true positive rate (the proportion of positive examples predicted correctly, defined exactly as recall) and false positive rate (the proportion of negative examples predicted incorrectly) to build up a summary picture of the classification performance.

**7.3. Hyper parameter tuning:**

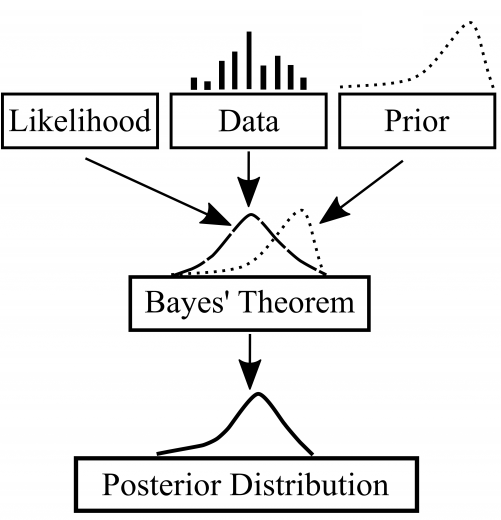
Hyper parameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyper parameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyper parameters grid that can be adjusted according to the business problem. Hyper parameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV, Randomized Search CV and Bayesian Optimization for hyper parameter tuning. This also results in cross validation and in our case we divided the dataset into different folds. The best performance improvement among the three was by Bayesian Optimization.

1. **Grid Search CV-**Grid Search combines a selection of hyper parameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.
2. **Randomized Search CV-** In Random Search, the hyper parameters are chosen at random within a range of values that it can assume. The advantage of this method is that there is a greater chance of finding regions of the cost minimization space with more suitable hyper parameters, since the choice for each iteration is random. The disadvantage of this method is that the combination of hyper parameters is beyond the scientist’s control

# **Bayesian Optimization-** Bayesian Hyper parameter optimization is a very efficient and interesting way to find good hyper parameters. In this approach, in naive interpretation way is to use a support model to find the best hyper parameters. A hyper parameter optimization process based on a probabilistic model, often Gaussian Process, will be used to find data from data observed in the later distribution of the performance of the given models or set of tested hyper parameters.

1. As it is a Bayesian process at each iteration, the distribution of the model’s performance in relation to the hyper parameters used is evaluated and a new probability distribution is generated. With this distribution it is possible to make a more appropriate choice of the set of values that we will use so that our algorithm learns in the best possible way.



**8. Conclusion:**

Using a Random Forest classifier, we can predict with ~78.2%. accuracy, whether a customer is likely to default next month.

The strongest predictors of default are the PAY\_X (ie the repayment status in previous months), the LIMIT\_BAL & the PAY\_AMTX (amount paid in previous months).

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