**Credit card default Prediction**

**Shubhangi Devidas Dharmik**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract**

This analysis used for the dataset that consists of 30,000 credit card usage records and 3 machine learning models - Logistic Regression, Random Forest and ADA-Boost. There might be other classification models that could work to give better performances. Predicting credit card default is an inherently challenging task and there is an inevitable trade-off between precision and recall.

From the analysis we understand that, creditors need to make decisions efficiently. The machine learning models in this analysis can be served as an aid to credit card companies, loan lenders, and banks make informed decisions on credit worthiness based on accessible to customer data.

***Keywords: Default payment next month, correlation, Null values, classification model.***

.

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 to evaluate which customers will default on their credit card payments

The dataset contain following columns :

* X1: Amount of the given credit for both the individual and his/her family 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 = the repayment status in September
* X7 = the repayment status in August,
* 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 = amount of bill statement in Sept,
* X13 = amount of bill statement in August,
* X17 = amount of bill statement in April,
* X18 = amount paid in September, 2005;
* X19 = amount paid in August, 2005;
* X23 = amount paid in April, 2005.

The main objective of this project is to conduct quantitative analysis on credit card default by using interpretable machine learning models with accessible customer data, instead of credit score or credit history, with the goal of assisting and speeding up the human decision making process.

1. **Introduction :**

Credit risk has traditionally the greatest risk among all the risks that the banking and credit card industries are facing, and it is usually requiring the most capital. This can be proven by industry business report and statistical data.

The purpose of this project is to conduct quantitative analysis on credit card default risk by using interpretable machine learning models with accessible customer data, instead of credit score or credit history, with the goal of assisting and speeding up the human decision making process.

1. **Factors Affecting :**

Following are the factors affecting to the default payment next month or our target variable:

1. **Gender** :

Males have more delayed payments than females in this dataset. Since there are more females than males in the dataset, we use percentage of default within each sex group shows 30% males have default payment while only 26% females have default payment. The difference is not significant. From gender we can find out the person is credible or not.

1. **Education** :

Customers with higher education have less delayed payment. Customers with lower education levels default more. Customers with high school and university educational level have higher default percentages than customers with grad school education. there is an education group others which appears to have the least default payment. From education we can find out the person is credible or not.

1. **Age** :

Middle-aged customers have the lowest default rate. The default probability increases for customers younger than 30 and older than 70. Customers aged between 30 and 50 have the lowest delayed payment rate, while younger groups which is from 20-30 and older groups 50-70 all have higher delayed payment rates. This aligns with social reality that customers aged 30-50 typically have the strongest earning power. We also notice the delayed rate drops slightly again in customers older than 70. This is understandable because elder customers consumption tends to decrease.

1. **Marital Status :**

Marital status is also in consideration which affect the default payment next month. Married people seems to default more often. So from their marital status we can find that the person is credible or not.

1. **Credit Limit** :

Higher credit limit is associated with lower default risk. Unsurprisingly, customers with higher credit limits have lower delayed payment rates.

1. **Inactive Customers** :

Customers who had no consumption in 6 months then default in the next month. We first detected 870 customers that were inactive for 6 months. Then we checked if these customers all had no default next month. To our surprise, 317 out of 870 inactive customers had default payment next month.

From the above factors we can calculate the person is credible or not.

1. **Steps involved :**

The following steps are involved in the project

1. **Exploratory Data Analysis** :

After loading and reading the dataset in notebook, we performed exploratory data analysis. The purpose of exploratory data analysis is to identify the variables that impact payment default next month and the correlations between them. We use graphical data exploratory analysis to check every categorical variable. We plot the graph and visualize data.

1. **Null values Treatment and Outliers :**

Dataset contains a no null values to disturb the accuracy.

1. **Numerical and categorical Features :** With the help of exploratory data analysis we analyzed the categorical as well as numerical features in the dataset.
2. **Correlation Analysis :**

We plot the heatmap to find the correlation between both dependent variable and independent variables.

1. **Train test Split :**

In train test split we take x as dependent variables and y take as independent variable then train the model.

1. **Models :**

We uses 3 modeling to train the data and for predicting the accuracy.

1. Logistic regression
2. Random forest
3. ADA-Boost
4. **Modeling Preparation :**

Since there are labeled data and the expected outcome is the probability of customer default, we define this as supervised machine learning for better model performance.

1. **Feature Selection**

There are 25 columns in this dataset and the target variable is the column is default payment next month. We drop the column ‘ID’ and target variable and save the rest 23 as predictor features. Those predictor variables include categorical variables such as sex, age, education marital status along with numerical variables, such as payment status, credit limit, bill amount, etc.

1. **Imbalance data :**

Imbalanced dataset will mislead machine learning algorithms and affect their performances so then we apply train test split to balance data.

1. **Split Training and Test Data:**

In the train test split we take two variables ie X and Y where X contain all the independent variables and Y contain dependent variable. Here the independent variable is default payment next month and dependent variables is affecting the default payment next month like age, gender, credit limit, education, bill payment etc.

For the model, we use the ratio for training and test data split by 80% for training, 20% for test to ensure consistency. After splitting the data, we set the test data aside and leave it for the very end, which is the final testing after hyper parameter tuning.

1. **Hyperparameters Tuning:**

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of tree based models like Random Forest Classifier and ADA-Boost classifier.

We utilize Scikit library’s built in functions such as cross-validation, randomized search and grid search to make the process easier. In Logistic Regression, by using hyperparameter reduces model complexity and prevents overfitting. With Random Forest, there are many hyperparameters available for tuning, but we use the default settings in sklearn and only focus on few. After creating a parameter grid, we use grid search to find the best parameters combinations.

1. **Performance Metrics:**

Since this is a classification problem with imbalanced classes, we use performance metrics i.e. accuracy precision and recall is a better choice. However, there is a known trade-off between precision and recall. We can raise recall to arbitrarily high, but the precision will decrease. We use below metrics to measure model performances.

**a. Confusion matrix**

**b. ROC\_AUC**

**c. Precision recall curve**

1. **Algorithms used:**

### Following algorithm is used to predict the value and for calculating the results. For finding out  the result of predictive accuracy of the estimated probability of default.

1. **Logistic Regression:**

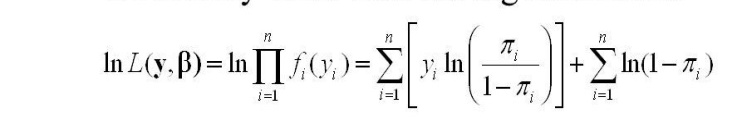
Logistic Regression is actually a classification algorithm that was given the name regression due to the fact that the mathematical formulation is very similar to linear regression.

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

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

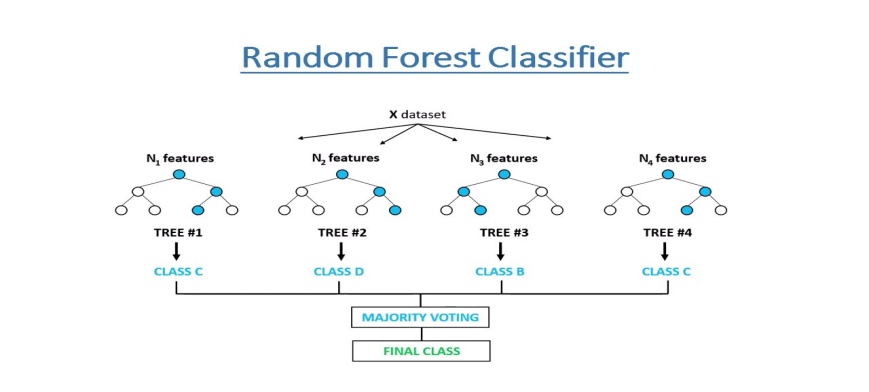


The optimization algorithm used is: Maximum Log Likelihood. We mostly take log likelihood in Logistic:

****

1. **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.

****

1. **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 model is 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 i.e. TP/TP+FP

Recall is the ratio of correct positive predictions to the overall number of positive examples in the set i.e. 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 i.e. the proportion of positive examples predicted correctly, defined exactly as recall and false positive rate i.e. the proportion of negative examples predicted incorrectly to build up a summary picture of the classification performance.

1. **Hyper parameter tuning :**

Hyperparameters 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 hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

1. **Grid Search CV:** Grid Search combines a selection of hyperparameters established and runs through all of them to evaluate the model’s performance. The 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 hyperparameters 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 hyperparameters, since the choice for each iteration is random. The disadvantage of this method is that the combination of hyperparameters is beyond the scientist’s control.

We used Grid Search CV, Randomized Search CV for hyperparameter tuning. This also results in cross validation and in our case we divided the dataset into different folds.

1. **Conclusions**
2. In Logistic Regression model has the highest recall but the lowest precision, if the firm expects high recall, then this model is the best candidate.
3. If the balance of recall and precision is the most important metric, then Random Forest is the ideal model.
4. Random Forest has slightly lower recall but much higher precision than Logistic Regression.
5. A comparison of the time required to train models vs. their predictive power.
6. Consider the applicants marital status. Married people seem to default more often.
7. Consider the age of the applicant. Younger people are at higher risk of defaulting.
8. We can say that random forest model gives us more accurate model instead of remaining ones.
9. Our best prediction accuracy was around 82-83%, our lowest measured prediction accuracy was about 79%.