**CREDIT CARD DEFAULT PREDICTION**

**Sayali Kamalapurkar**

**Data Science Trainee,**

**AlmaBetter, Bengaluru**

**ABSTRACT:** A Taiwan-based credit card issuer wants to better predict the likelihood of default for its customers, as well as identify the key drivers that determine this likelihood. This would inform the issuer’s decisions on who to give a credit card to and what credit limit to provide. It would also help the issuer have a better understanding of their current and potential customers, which would inform their future strategy, including their planning of offering targeted credit products to their customers.

Credit card is a flexible tool by which one can use bank’s money for a short period of time. If one accepts a credit card, he agrees to pay his bills by the due date listed on your credit card statement. Otherwise, the credit card will be defaulted. When a customer is not able to pay back the loan by the due date and the bank is totally certain that they are not able to collect the payment, it will usually try to sell the loan. After that, if the bank recognizes that they are not able to sell it, they will write it off. This is called a charge-off. This results in significant financial losses to the bank on top of the damaged credit rating of the customer and thus it is an important problem to be tackled. In this project, I will build machine learning models which will predict individuals who will default their credit card payment next month.

The Evaluation Metric used is the recall score. After we train on some training data, we will evaluate the performance of the model on some test data. For this, we use the Confusion Matrix.

Text

Description automatically generated

Models used:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Support Vector Machine (SVM)

## 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.

**INTRODUCTION:** Banks all around the world would receive countless applications for loans every day. Some of them are good and will be repaid, but there is a still high risk that one creditor defaults his/her loans. How could we prevent this problem from happening? Or, in another word, how could we know in advance which creditors are trustworthy? Hence, we are analysing the credit card dataset to get better predictions of the defaulters in advance.

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:

Some of the features of the credit card dataset are as follows:

* 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; 5 = unknown; 6 = unknown)

* 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)

* X12-X17: Amount of bill statement (NT dollar)
* X18-X23: Amount of previous payment (NT dollar)

**STEPS INVOLVED:**

1. **Importing libraries :**

Imported Python libraries such as NumPy, Pandas for data manipulation and Matplotlib, Seaborn for data visualization. Sci-kit Learn and classification models from Sci-kit learn, also metrics and hyperparameter tuning libraries are imported.

1. **Data Collection :**

To proceed with the problem dealing first we will load our dataset that is given to us in .xls file into a dataframe. Mount the drive and load the excel file into a dataframe.

1. **Explore dataset :**

Found out different attributes of the dataset, like – shape, number of null values in the dataset, statistical information about the numerical columns and categorical columns. Renamed some of the columns. Replaced values by categories in categorical columns.

1. **EDA on features :**

Using visualization libraries in python, did some exploratory data analysis on different features with respect to target variable, default. Further, did univariate analysis on each column, then bivariate analysis where two columns are analysed together to find out relationships. Finally did multivariable analysis for understanding relationships among features and with target variable.

1. **Handling imbalanced data:**

As we know there are very few defaulters in general. But we have to get those defaulters predictions right for almost all time. If we apply classification models on such highly imbalanced data, then model favours the majority class due to its larger volume presence. Thus, there is a need to make data balanced by using either undersampling or oversampling techniques. When we have a tons of data to deal with then we prefer undersampling. But here we have just 30000 rows. Hence we have done oversampling with SMOTE (Synthetic Minority Oversampling Technique).

1. **Feature engineering :**

Feature engineering is the process of selecting, transforming, extracting, combining, and manipulating raw data to generate the desired variables for analysis or predictive modelling. It is a crucial step in developing Machine Learning models. We created some meaningful numerical features from the given columns, such as – paid\_total, total\_bill, Dues, etc.

1. **Data Cleaning and Outlier treatment :**

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. But, in this case we are not going to remove those outliers, as it can be a defaulter which is very high risk if we ignore such rows.

1. **One Hot Encoding:**

One Hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.

Here, we have performed One Hot encoding on ‘Education’, ‘Marriage’, ‘gender’ and columns having repay status for April to September.

1. **Train Test split :**

The train-test split is a**technique for evaluating the performance of a machine learning algorithm.** The procedure involves taking a dataset and dividing it into two subsets.

1. **Model Building :**

A classification model attempts to draw some conclusions from the observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. Outcomes are labels that can be applied to a dataset. Here, we have defaulter and non-defaulter.

* 1. **Logistic Regression Model** –

**A logistic regression estimates the probability of an event occurring, such as default or non-default, based on the given dataset of independent variables. Since the outcome is the probability, the dependent variable is bounded between 0 and 1.**

* 1. **Decision Tree Classifier Model** –

Decision tree is a supervised machine learning algorithm that uses a set of rules to make decisions, similarly to how humans make decisions. The intuition behind Decision trees is that we use the dataset features to create yes/no questions and continually split the dataset until we isolate all data points belonging to each class.

* 1. **Random Forest Classifier Model**–

Random Forest consists of multiple decision trees just as a forest has many trees. On top of that, it uses randomness to enhance its accuracy and combat overfitting, which can be a huge issue for such a sophisticated algorithm.

* 1. **Support Vector Machine** –

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

1. **Hyperparameter Tuning :**

Hyperparameter tuning, also called hyperparameter optimization, is the process of**finding the configuration of hyperparameters that results in the best performance**.

1. **Selection of Model :**

**We have considered evaluation metrics as follows –**

* **Accuracy – The proportion of the total number of predictions that are correct.**
* **Precision – The proportion of positive predictions that are actually correct.**
* **Recall – The proportion of positive observed values correctly predicted as such**

**Metric Selection:**

False Negatives that is a person who defaults, predicted as payer is worse. Hence we will look for better recall to have good predictions.

Thus, from all four models which we have built, we chose Random Forest which had recall score as the highest among others – 82%.

**CONCLUSIONS:**

**Following are few conclusions which we drew from the whole process.**

* According to RandomForest model, features like paid\_total, total\_bill, dues, cred\_lim, age, repay\_status\_sep\_2, repay\_status\_aug\_1 and marriage status as married are found to be most important features to predict future defaulters.
* If the credit card holder has paid minimum credit amount for past 6 months, also for every month if the dues of that customers are increasing, then it is obvious that the customer will have default for sure!
* While this age and limit\_bal are the other two important predictors as we discussed in the data preparation part. This also makes sense because as one gets older, one is more likely to accumulate more resource and cares more about his reputations, which makes credit default less likely. Also, if one and one’s family get more given credits, the person is more likely to live in a wealthier environment which also makes credit default less likely.
* Further, repay status is also playing a vital role in predicting future defaulters. If the creditor is not repaying for the past one or more months, he is more likely to default in upcoming months.

At the end of the day, having the ability to predict 82% (recall score) of potential defaults would save a-lot of money on credit card charge-offs. Obviously, real-world application is more nuanced, but this modeling process is a step in the right direction.