**Abstract:**

This project presents and discusses data-driven predictive models for predicting the defaulters among the credit card users. Data used include details like limit balance, age, sex, amount of bill statement, repayment status and amount of previous payment. The paper discusses which variables are the strongest predictors of default, and to make predictions on which customers are likely to default.

**Problem Statement:**

This project is basically 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:**

**Credit Card -**

A credit card is a financial instrument issued by bank with a pre-set credit limit, helping you make cashless transactions. the credit card issuer determines the credit limit based on your credit score, credit history and your income.

**What is Credit Card default?**

Missing credit card payments once or twice does not count as a default. A payment default occurs when you fail to pay the Minimum Amount Due on the credit card for a few consecutive months. Usually, the default notice is sent by the card issuer after 6 consecutive missed payments. However, the final call rests with the bank.

**Data Description:**

• 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; . .

• X 11= 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.

**Exploratory Data Analysis:**

**Data Preparation:**

This is the stage of the project where one decides on the data that one will use for analysis. The criteria used to make this decision includes the relevance of the data to the data mining goals, the quality of the data, and technical constraints such as limits on data volume or data types. For data preparation, number of outliers were checked and observed to have more number of outliers in some features.

● **Null values :**

NO null value in the dataset

**● Encoding of categorical columns** We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

**LIMIT\_BAL:**

LIMIT\_BAL states the amount of given credit. This is the maximum amount a customer can spend with their credit card in a single month. The amount of balance limit is dependent on the bank’s own screening processes and other unknown factors.

**Age:**

This is the age of the customer which is stated in years. The mean and median of the age are 35 and 34 respectively. This is numerical column. There are no missing values in this column. There are some outliers in AGE column. The decline in number of customers starts from about 30 years among the nondefaulting group, while the number of customers of different ages stays much more constant from 25 to around 40 years. This indicates that likelihood of default among men grows with age. In box plot we can see many outliers which are considered here as extreme values.

**SEX:**

This variable can obtain a value of 1 for male and 2 for female and 3 for others. In this study, sex and gender are used interchangeably to intend the same thing.

**EDUCATION:**

The education level of a customer is represented as one of four values: 1 = Graduate school, 2 = University, 3 = High school, 4 = Other. For the purpose of analysing customer groups, this is assumed to indicate the highest level of education completed.

**Bill\_AMT:**

Amount of bill statement is recorded in this variable. It is represented in the data as 6 columns, one for each month. Data collected from 6 months, April to September.

**Fitting different models:**

For modelling we tried various classification algorithms like:

1. Logistic Regression

2. Decision tree

3. Random Forest Classifier

4. SVC

5. XGBoost classifier

6. Hyperparameter tuning

**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.

**2. 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.

**3. 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

**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.

We used Grid Search CV, Randomized Search CV and Bayesian Optimization for hyperparameter 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

**SMOTE Oversampling:**

In the initial model fitting, we start by using all models’ default parameters. To compensate for the rare classes in the imbalance dataset, we use SMOTE(Synthetic Minority OverSampling Technique) method to over sample the minority class and ensure the sampling is not biased. What this technique does under the hood is simply duplicating examples from the minority class in the training dataset prior to fitting a mode. After SMOTE sampling, the dataset has equal size of 0s and 1s.

In order to verify if SMOTE improves models’ performance, all 3 models are trained with SMOTE and without SMOTE. Below table shows the ROC\_ AUC scores on training data improved significantly with all models after over sampling with SMOTE. This proves SMOTE is an effective method in sampling imbalanced dataset.

**Conclusion:**

From the project we can conclude that the default rate is higher for males, increases as the education increases, and also increases as the age of a person increases. i.e clients whose age over 60 was higher than mid-age and young people. In all of these models, our recall revolves in the range of 78 to 86%.with the best fit model as random forest

We started with the logistic regression model, then compared the results with traditional machine learning-based models then K-means SMOTE resampling method

From all baseline model, Random forest classifier shows highest train accuracy, test accuracy and F1 score .

After cross validation and hyperparameter tunning, XG Boost shows highest test accuracy score of 83.03%.

Cross validation and hyperparameter tunning certainly reduces chances of overfitting and also increases performance of model.

**Reference:**

1. GeeksforGeeks
2. Analytics Vidhya