CREDIT CARD DEFAULT PREDICTION

Raja Chowdhury
Data science trainee,
AlmaBetter, Bangalore

ABSTRACT:

Financial threats are displaying a trend about the credit risk of commercial banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats faces by commercial banks is the risk prediction of credit clients. Recent studies mostly focus on enhancing the classifier performance for credit card default prediction rather than an interpretable model. There are many people who default on their credit card payments. This can happen for a variety of reasons. Maybe they lost their job and could no longer afford the payments. Maybe they had an unexpected medical emergency that left them with a lot of debt. Whatever the reason, defaulting on a credit card can have serious consequences. If you default on your credit card, the first thing that will happen is that your account will be sent to collections. This means that a collection agency will try to get the money that you owe from you. They may call you, send you letters, or even come to your door. If you do not pay, they may take you to court. Defaulting on your credit card will also damage your credit score. This can make it harder for you to get a loan in the future. It can also make it more expensive to borrow money. Your interest rates will go up and you may not be able to get a credit card with a good interest rate. Defaulting on your credit card can be a stressful and difficult situation. If you are having trouble making your payments, contact your credit card company and try to work something out. You may be able to negotiate a lower interest rate or a payment plan. This model will help commercial banks, financial organizations, loan institutes, and other decision-makers to predict the loan defaulter earlier.

INTRODUCTION:

In recent years, the credit card issuers in Taiwan faced the cash and credit card debt crisis and the delinquency is expected to peak in the third quarter of 2006 (Chou,2006). In order to increase market share, card-issuing banks in Taiwan over-issued cash and credit cards to unqualified applicants. At the same time, most cardholders, irrespective of their repayment ability, overused credit card for consumption and accumulated heavy credit and cash–card debts. The crisis caused the blow to consumer finance confidence and it is a big challenge for both banks and cardholders._When it comes to personal finance, there is one thing that is certain: sooner or later, everyone will experience a financial setback. For some people, this will take the form of a job loss, a medical emergency, or some other unexpected expense. For others, it will be something more mundane, like overspending on a credit card. No matter the cause, the effect is the same: a person who is unable to make their credit card payments on time is said to have "defaulted." This can have serious consequences for one's credit score and financial future. Defaulting on a credit card is not something to be taken lightly. If you find yourself in this situation, it is important to take steps to correct the problem as soon as possible. The sooner you are able to get back on track with your payments, the less damage will be done to your credit score. There are many ways to avoid defaulting on your credit card. The most important thing is to be aware of your spending and to make sure that you only charge what you can afford to pay back. If you find yourself in a situation where you can no longer make your payments, contact your credit card company immediately and explain the situation. They may be able to work with you to create a payment plan that will help you get back on track. Defaulting on your credit card is not the end of the world, but it is a serious setback. If you find yourself in this situation, take steps to correct the problem as soon as possible. With a little effort, you can get your finances back on track and avoid further damage to your credit score.

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

DATA DESCRIPTION:

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

DATASET PREPARATION:

The credit card default prediction dataset is a data of Taiwan. The data is from the year 2005. Six months of bill payment and amount payment is given Below Table shows the data features.

Data-set description

Feature Name	Type
ID	Int64
LIMIT BAL	Int64
SEX	Int64
EDUCATION	Int64
MARRIAGE	Int64
AGE	Int64
PAY_0	Int64
PAY_2	Int64
PAY_3	Int64
PAY_4	Int64
PAY_5	Int64
PAY_6	Int64
BILL_AMT1	Int64
BILL_AMT2	Int64
BILL_AMT3	Int64
BILL_AMT4	Int64
BILL_AMT5	Int64
BILL_AMT6	Int64
PAY_AMT1	Int64
PAY_AMT1	Int64
PAY_AMT1	Int64
PAY_AMT1	Int64
PAY_AMT1	Int64
PAY_AMT1	Int64
Default payment next month	Int64

FEATURE BREAKDOWN:

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.

EXPLORATORY DATA ANALYSIS:

If we want to explain EDA in simple terms, it means trying to understand the given data much better, so that we can make some sense out of it. we using univariate frequency analysis was conducted to describe key characteristics of each feature including, minimum and maximum value, average, standard deviation and others. It was also used to produce a value distribution and identify missing values, and outliers.
EDA is a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures. In this chapter, we are going to discuss the steps involved in performing top notch exploratory data analysis

In statistics, A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing tasked in Python uses data visualization to draw meaningful patterns and insights

- **DATA ANALYSIS:**

This is one of the most crucial steps that deals with descriptive statistics and analysis of the data. The main tasks involve summarizing the data, finding the hidden correlation and relationships among the data, developing predictive models, evaluating the models, and calculating the accuracies. Some of the techniques used for data summarization are summary tables, graphs, descriptive statistics, inferential statistics, correlation statistics, searching, grouping, and mathematical models.

- **DATA SOURCING**

Data Sourcing is the process of finding and loading the data into our system. Broadly there are two ways in which we can find data.

1. Private Data
2. Public Data

Data collected from several sources must be stored in the correct format and transferred to the right information technology personnel within a company. As mentioned previously, data can be collected from several objects on several events using different types of sensors and storage tools.

- **DATA PREPROCESSING:**

A dataset may contain noise, missing values, and inconsistent data; thus, pre-processing of data is essential to improve the quality of data and time required in the data mining.

- **DATA CLEANING**

After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

Irregularities are of different types of data.

- Missing Values
- Incorrect Format
- Incorrect Headers
- Anomalies/Outliers

- **DATA TRANSFORMATION:**

Data transformation is the process of normalizing and aggregating the data to further improve the efficiency and accuracy of data mining.

- **DATA DEDUPLICATION:**

It is very likely that your dataset contains duplicate rows. Removing them is essential to enhance the quality of the dataset but we didn't find any.

- **MISSING VALUES:**

There is a representation of each service and product for each customer. Missing values may occur because not all customers have the same subscription. Some of them may have a number of service and others may have something different. In addition, there are some columns related to system configurations and these columns may have null values but in our orange telecom data set there are no null values present

If there are missing values in the Dataset before doing any statistical analysis, we need to handle those missing values.

There are mainly three types of missing values.

1. MCAR (Missing completely at random): These values do not depend on any other features.
2. MAR (Missing at random): These values may be dependent on some other features.

MNAR (Missing not at random): These missing values have some reason for why they are missing.

- **DROPPING MISSING VALUES:**

One of the ways to handle missing values is to simply remove them from our dataset. We have known that we can use the is null () and not null () functions from the panda's library to determine null values but we didn't find any null values.

- **HANDLING OUTLIERS:**

Outliers are data points that diverge from other observations for several reasons. During the EDA phase, one of our common tasks is to detect and filter these outliers. The main reason for this detection and filtering of outliers is that the presence of such outliers can cause serious issues in statistical analysis.

There are two types of outliers:

- **UNIVARIATE OUTLIERS:**

Univariate outliers are the data points whose values lie beyond the range of expected values based on one variable.

- **MULTIVARIATE OUTLIERS:**

While plotting data, some values of one variable may not lie beyond the expected range, but when you plot the data with some other variable, these values may lie far from the expected value.

- **MEASURES OF CENTRAL TENDENCY:**

The measure of central tendency tends to describe the average or mean value of datasets that is supposed to provide an optimal summarization of the entire set of measurements. This value is a number that is in some way central to the set. The most common measures for analysing the distribution frequency of data are the mean, median, and mode.

- **MEASURES OF DISPERSION**:

The second type of descriptive statistics is the measure of dispersion, also known as a measure of variability. If we are analysing the dataset closely, sometimes, the mean/average might not be the best representation of the data because it will vary when there are large variations between the data. In such a case, a measure of dispersion will represent the variability in a dataset much more accurately.

Multiple techniques provide the measures of dispersion in our dataset. Some commonly used methods are standard deviation (or variance), the minimum and maximum values of the variables, range, kurtosis, and skewness.

- **STANDARDIZING VALUES:**

To perform data analysis on a set of values, we have to make sure the values in the same column should be on the same scale. For example, if the data contains the values of the top speed of different companies' cars, then the whole column should be either in meters/sec scale or miles/sec scale.

- **UNIVARIATE ANALYSIS:**

  If we analyse data over a single variable/column from a dataset, it is known as Univariate Analysis. Univariate analysis looks at one feature at a time. When we analyse a feature independently, we are usually mostly interested in the distribution of its values and ignore other features in the dataset

  Univariate analysis is the simplest form of analysing data. It means that our data has only one type of variable and that we perform analysis over it. The main purpose of univariate analysis is to take data, summarize that data, and find patterns among the values. It doesn't deal with causes or relationships between the values. Several techniques that describe the patterns found in univariate data include central tendency (that is the mean, mode, and median) and dispersion (that is, the range, variance, maximum and minimum quartiles (including the interquartile range), and standard deviation).

- **BIVARIATE ANALYSIS:**

If we analyse data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis.

- **a) Numeric-Numeric Analysis:**

Analysing the two numeric variables from a dataset is known as numeric-numeric analysis. We can analyse it in three different ways.

  - Scatter Plot
  - Pair Plot
  - Correlation Matrix

- **b) Numeric - Categorical Analysis:**

Analysing the one numeric variable and one categorical variable from a dataset is known as numeric-categorical analysis. We analyse those mainly using mean, median, and box plots.

- **MULTIVARIATE ANALYSIS:**

Multivariate analysis is the analysis of three or more variables. This allows us to look at correlations (that is, how one variable changes with respect to another) and attempt to make predictions for future behaviour more accurately than with bivariate analysis.

One common way of plotting multivariate data is to make a matrix scatter plot, known as a pair plot. A matrix plot or pair plot shows each pair of variables plotted against each other. The pair plot allows us to see both the distribution of single variables and the relationships between two variables

- **CORRELATION AMONG VARIABLES:**

In words, the statistical technique that examines the relationship and explains whether, and how strongly, pairs of variables are related to one another is known as correlation. Correlation answers questions such as how one variable changes with respect to another. If it does change, then to what degree or strength? Additionally, if the relation between those variables is strong enough, then we can make predictions for future behaviour

- **GRAPHICAL REPRESENTATION OF THE RESULTS:**

This step involves presenting the dataset to the target audience in the form of graphs, summary tables, maps, and diagrams. This is also an essential step as the result analysed from the dataset should be interpretable by the business stakeholders, which is one of the major goals of EDA. Most of the graphical analysis techniques include Line chart, Bar chart, Scatter plot, Area plot, and stacked plot Pie chart, Table chart, Polar chart, Histogram, Lollipop chart etc.

ALGORITHMS:

1. LINEAR REGRESSION:

Linear regression is a supervised machine learning model majorly used in forecasting. Supervised machine learning models are those where we use the training data to build the model and then test the accuracy of the model using the loss function.

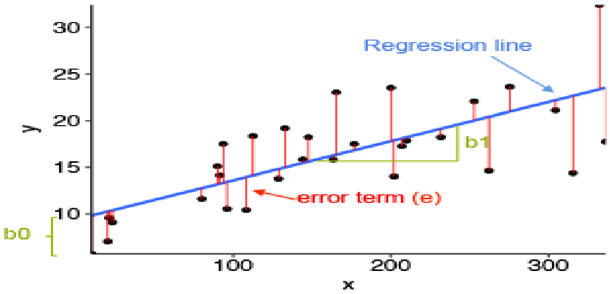
Linear regression is one of the most widely known time series forecasting techniques which is used for predictive modelling. As the name suggests, it assumes a linear relationship between a set of independent variables to that of the dependent variable (the variable of interest).

We're going to fit a line

$$y = \beta0 + \beta1x$$

to our data. Here, x is called the independent variable or predictor variable, and y is called the dependent variable or response variable. Before we talk about how to do the fit, let's take a closer look at the important quantities from the fit:

- $\beta1$ is the slope of the line: this is one of the most important quantities in any linear regression analysis
- $\beta0$ is the intercept of the line.



2. RIDGE REGRESSION:

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

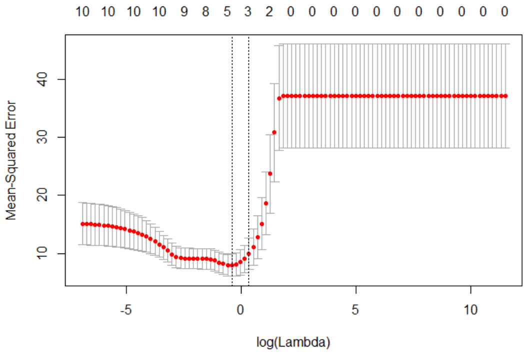
we have concluded that we would like to decrease the model complexity, that is the number of predictors. We could use the forward or backward selection for this, but that way we would not be able to tell anything about the removed variables' effect on the response. Removing predictors from the model can be seen as settings their coefficients to zero. Instead of forcing them to be exactly zero, let's penalize them if they are too far from zero, thus enforcing them to be small in a continuous way. This way, we decrease model complexity while keeping all variables in the model. This, basically, is what Ridge Regression does.

$$L_{ridge}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j=1}^{m}\hat{\beta}_j^2 = ||y - X\hat{\beta}||^2 + \lambda||$$

3. LASSO REGRESSION:

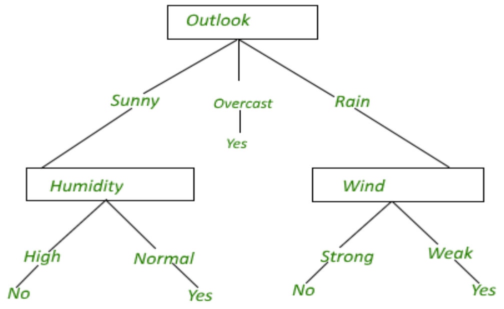
Lasso, or Least Absolute Shrinkage and Selection Operator, is quite similar conceptually to ridge regression. It also adds a penalty for non-zero coefficients, but unlike ridge regression which penalizes sum of squared coefficients (the so-called L2 penalty), lasso penalizes the sum of their absolute values (L1 penalty). As a result, for high values of λ, many coefficients are exactly zeroed under lasso, which is never the case in ridge regression. The only difference in ridge and lasso loss functions is in the penalty terms. Under lasso, the loss is defined as:

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j|.$$



4.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. A tree can be *"learned"* by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called *recursive partitioning*. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, and then moving down the tree branch corresponding to the value of the attribute as shown in the above figure. This process is then repeated for the subtree rooted at the new node.



5. RANDOM FOREST:

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 greatest number of times a label has been predicted out of all.

6. GRADIENT BOOSTING:

The term gradient boosting consists of two sub-terms, gradient and boosting. We already know that gradient boosting is a boosting technique. Let us see how the term 'gradient' is related here.

Gradient boosting re-defines boosting as a numerical optimisation problem where the objective is to minimise the loss function of the model by adding weak learners using gradient descent. Gradient descent is a first-order iterative optimisation algorithm for finding a local minimum of a differentiable function. As gradient boosting is based on minimising a loss function, different types of loss functions can be used resulting in a flexible technique that can be applied to regression, multi-class classification, etc

CONCLUSIONS:

Machine learning methods, in conjunction with the use of imbalanced methods, have been utilized in various domains. The objective of this paper is to train various supervised learning algorithms to predict the client's behaviour in paying off the credit card balance. In classification problems, an imbalanced dataset is also crucial to enhance the performance of the model, so different resampling techniques were also used to balance the dataset. We first investigated the datasets by using exploratory data analysis techniques, including data normalization. We started with the GBDT model, then compared the results with traditional machine learning-based models. The prediction accuracy rate of the GBDT model is higher than the traditional machine learning-based models. The GBDT method given the best accuracy of 88.7% while utilizing the K-means SMOTE resampling method on Taiwan client's credit dataset. The results obtained through Taiwan client's credit dataset have significantly better than other datasets employed in this
study.

In our dataset, we used many algorithms like Logistic Regression, Support vector classifier, decision tree classifier, XGBoost Classifier, Random Forest Classifier. Below is the best algorithm for the respective metrics.

1. Random Forest Classifier has the best value of accuracy score of 86%
2. Random Forest Classifier has the best value of precision score of 86%
3. Random Forest Classifier has the best value of recall score of 86%
4. Random Forest Classifier has the best value of f1 score of 86%
5. Random Forest Classifier has the best value of Roc_auc score of 86%

This proves Random Forest Classifier algorithm has perfectly fitted all the dataset.

REFERENCES:

- Data science for business: what you think about data mining
- https://book.akij.net/eBooks/2018/May/5aef50939a868/Data_Science_for_Bus.pdf
- Hands-On Exploratory Data Analysis with Python Perform EDA techniques to understand, summarize, and investigate your data by Suresh Kumar Mukhiya, Usman Ahmed (z-lib.org)
- https://bunker2.zlibcdn.com/dtoken/01c5fc197a94283bfb0c0943bd5b2d0c