Credit Card Default Prediction

Kanika Raj
Data science trainees,
Alma Better, Bangalore

ABSTRACT:

A credit card is a payment card issued to users (cardholders) to enable the cardholder to pay a merchant for goods and services based on the cardholder's accrued debt (i.e., promise to the card issuer to pay them for the amounts plus the other agreed charges). The card issuer (usually a bank or credit union) creates a revolving account and grants a line of credit to the cardholder, from which the cardholder can borrow money for payment to a merchant or as a cash advance. There are two credit card groups: consumer credit cards and business credit cards. Most cards are plastic, but some are metal cards (stainless steel, gold, palladium, titanium), and a few gemstone-encrusted metal cards.

A regular credit card is different from a charge card, which requires the balance to be repaid in full each month or at the end of each statement cycle. In contrast, credit cards allow the consumers to build a continuing balance of debt, subject to interest being charged. A credit card differs from a charge card also in that a credit card typically involves a third-party entity that pays the seller and is reimbursed by the buyer, whereas a charge card simply defers payment by the buyer until a later date

INTRODUCTION:

Whenever you accept a credit card, you agree to certain terms and conditions including making your minimum payment by the due date listed on your credit card statement.

When you miss the minimum payment by 6 months or more in a row, your credit card will be in default. In such situations, your credit card issuer will first send several notices via email or SMS and call you asking to make the payment. If you do not make the payment after a stipulated period, they will close your account and report the default to the credit bureaus.

This period may vary from one credit card provider to another. This tends to impact your credit score and it will be difficult for you to get approved for loans in the future. Once you are listed as a credit card defaulter, you become a risk for any credit obligation.

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 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:

1) X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
2) X2: Gender (1 = male; 2 = female).
3) X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
4) X4: Marital status (1 = married; 2 = single; 3 = others).
5) X5: Age (year).
6) 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.
7) 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.
8) 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.

Data-set description

Feature Name	Type
ID	int64
LIMIT_BAL	int64
GENDER	int64
EDUCATION	int64
MARRIAGE	int64
AGE	int64
PAY_SEPT	int64
PAY_AUG	int64
PAY_JUL	int64
PAY_JUN	int64
PAY_MAY	int64
PAY_APR	int64
BILL_AMT_SEPT	int64
BILL_AMT_AUG	int64
BILL_AMT_JUL	int64
BILL_AMT_JUN	int64
BILL_AMT_MAY	int64
BILL_AMT_APR	int64
PAY_AMT_SEPT	int64
PAY_AMT_AUG	int64
PAY_AMT_JUL	int64
PAY_AMT_JUN	int64
PAY_AMT_MAY	int64
PAY_AMT_APR	int64
default_payment_next_month	int64

FEATURE BREAKDOWN:

We have records of 30000 customers. Below is the description of all features:

1) **ID**: ID of each client
2) **LIMIT_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
3) **SEX:** Gender (1 = male, 2 = female)
4) **EDUCATION**: (1 = graduate school, 2 = university, 3 = high school, 0,4,5,6 = others)
5) **MARRIAGE**: Marital status (0 = others, 1 = married, 2 = single, 3 = others)
6) **AGE**: Age in years

7) **PAY_0**: Repayment status in September, 2005 (scale same as above)
8) **PAY_2**: Repayment status in August, 2005 (scale same as above
9) **PAY_3**: Repayment status in July, 2005 (scale same as above)
10) **PAY_4**: Repayment status in June, 2005 (scale same as above)
11) **PAY_5**: Repayment status in May, 2005 (scale same as above)
12) **PAY_6**: Repayment status in April, 2005 (scale same as above)
13) **BILL_AMT1**: Amount of bill statement in September, 2005 (NT dollar)
14) **BILL_AMT2**: Amount of bill statement in August, 2005 (NT dollar)
15) **BILL_AMT3**: Amount of bill statement in July, 2005 (NT dollar)
16) **BILL_AMT4**: Amount of bill statement in June, 2005 (NT dollar)
17) **BILL_AMT5**: Amount of bill statement in May, 2005 (NT dollar)
18) **BILL_AMT6**: Amount of bill statement in April, 2005 (NT dollar)
19) **PAY_AMT1**: Amount of previous payment in September, 2005 (NT dollar)
20) **PAY_AMT2**: Amount of previous payment in August, 2005 (NT dollar)
21) **PAY_AMT3**: Amount of previous payment in July, 2005 (NT dollar)
22) **PAY_AMT4**: Amount of previous payment in June, 2005 (NT dollar)
23) **PAY_AMT5**: Amount of previous payment in May, 2005 (NT dollar)
24) **PAY_AMT6**: Amount of previous payment in April, 2005 (NT dollar)
25) **Default_payment_next_month**: Default payment (1=yes, 0=no)

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 multiple graphs 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.

EDA is a process of examining the available dataset to discover patterns, spot anomalies, one hot encoding, feature engineering 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 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 DUPLICATION:**

It is very likely that your dataset contains duplicate rows. Removing them is essential to enhance the quality of the dataset.

- **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 pandas.

- **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.

- **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 heatmap
- Boxplot

- **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.

- **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

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:

➢ Logistic Regression:

Logistic regression is a process of modelling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.

It is used in Machine Learning to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. This type of analysis can help you predict the likelihood of an event happening or a choice being made.

➢ Decision Tree Classifier:

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

The decision tree classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute.

➢ Random Forest Classifier:

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

➢ XGB Classifier:

XGBoost, which stands for (extreme Gradient Boosting), is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It is a popular supervised-learning algorithm used for regression and classification on large datasets. It uses sequentially-built shallow decision trees to provide accurate results and a highly-scalable training method that avoids overfitting.

XGBoost provides a highly efficient implementation of the stochastic gradient boosting algorithm and access to a suite of model hyperparameters designed to provide control over the model training process. The most important factor behind the success of XGBoost is its scalability in all scenarios.

REFERENCES:

1. **Kaggle**
2. **W3 school**
3. **Geeks for geeks**
4. **Almabetter class notes**
5. **Towards data science**