Credit Card Default Prediction

Ajit Kumar Toppo
Data Science Trainee,Almabetter

Abstract

With the change in time, more and more people are using credit cards. Not only people's living standards have improved but also people's consumption concept and consumption mode has drastically changed. People are more and more inclined to spend ahead of time and mortgage their “credit” to the bank to enjoy certain things in advance. However, when consuming, people often lack rational thinking and overestimate their ability to repay loans to banks in time. On the one hand, it increases the loan risk of banks; on the other hand, it increases the credit crisis of consumers themselves . With a large number of banks selling credit cards, the phenomenon of credit card default emerges one after another. It is very important for banks to effectively identify high-risk credit card default users.

Keywords- Pandas, Numpy, Seaborn, Machine learning, Classification, Scikit learn

Problem Statement

The project aims to predict the customers who are more likely to default on their credit card payment. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

About Dataset

There are 30,000 rows and 25 columns present in our dataset. Columns present in the data is shown below:

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

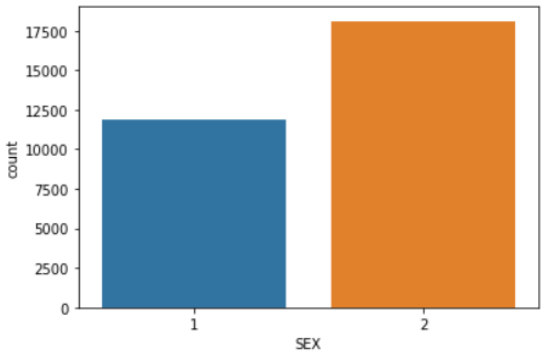
- **PAY_0:** Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
- **PAY_2:** Repayment status in August, 2005 (scale same as above)
- **PAY_3:** Repayment status in July, 2005 (scale same as above)
- **PAY_4:** Repayment status in June, 2005 (scale same as above)
- **PAY_5:** Repayment status in May, 2005 (scale same as above)
- **PAY_6:** Repayment status in April, 2005 (scale same as above)
- **BILL_AMT1:** Amount of bill statement in September, 2005 (NT dollar)
- **BILL_AMT2:** Amount of bill statement in August, 2005 (NT dollar)
- **BILL_AMT3:** Amount of bill statement in July, 2005 (NT dollar)
- **BILL_AMT4:** Amount of bill statement in June, 2005 (NT dollar)
- **BILL_AMT5:** Amount of bill statement in May, 2005 (NT dollar)
- **BILL_AMT6:** Amount of bill statement in April, 2005 (NT dollar)
- **PAY_AMT1:** Amount of previous payment in September, 2005 (NT dollar)
- **PAY_AMT2:** Amount of previous payment in August, 2005 (NT dollar)
- **PAY_AMT3:** Amount of previous payment in July, 2005 (NT dollar)
- **PAY_AMT4:** Amount of previous payment in June, 2005 (NT dollar)
- **PAY_AMT5:** Amount of previous payment in May, 2005 (NT dollar)
- **PAY_AMT6:** Amount of previous payment in April, 2005 (NT dollar)
- **Default payment next month:** Default payment (1=yes, 0=no)

Scale for PAY_0 to PAY_6 : (-2 = No consumption, -1 = paid in full, 0 = use of revolving credit (paid minimum only), 1 = payment delay for one month, 2 = payment delay

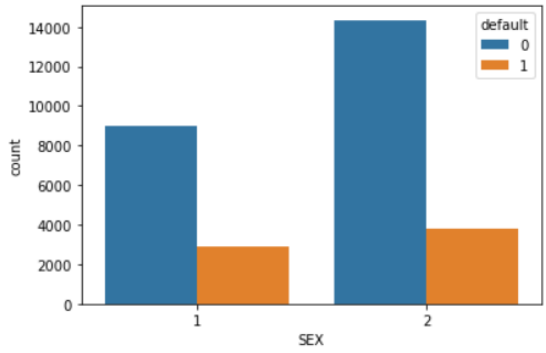
for two months, ... 8 = payment delay for eight months, 9 = payment delay for nine months and above).

Data Exploration

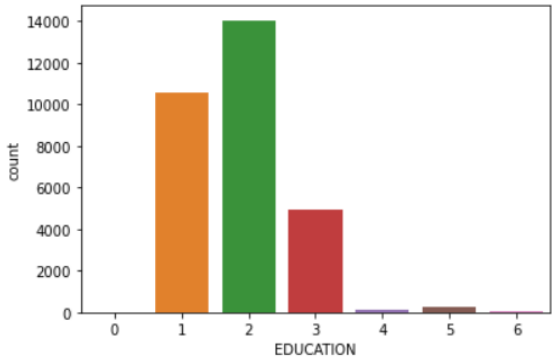
There are no null values present in the dataset. For convenience, I changed the column name from ‘Default payment next month’ to ‘default’. Plotted hist bins to get a sense of how the data is distributed within the variable.



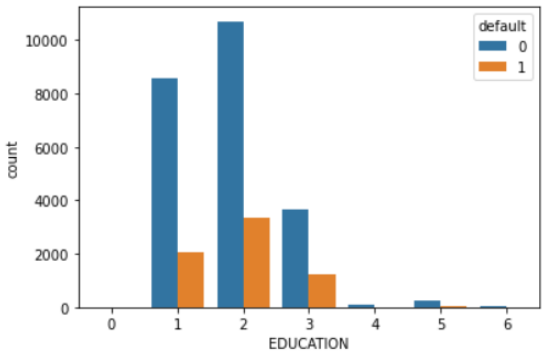
The graph above shows the male and female credit card holders.1 is for male and 2 is for female. Out of 30,000 records available to us of different credit card users the blue line represents 11,888 are males and the orange represents 18,112 are females.



The above graph (1-males and 2-females) shows the number of defaulters vs non-defaulters from male and female category.



From the above graph maximum records are available for people having education as (university-14030), (graduate school-10585),(high school-4917) .Where 1=graduate school, 2=university, 3=high school, 4=others,5 & 6= unknown.



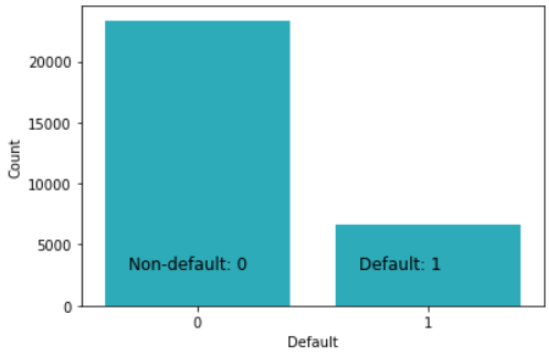
Above graph shows the relation between the educational background and being a defaulter. Credit card holders by majority age group are given below:

29	1605
27	1477
28	1409
30	1395
26	1256
31	1217
25	1186
34	1162
32	1158
33	1146

Majority of card holders are in the age range of 20's and 30's.

Data Preparation

Dropped the column ID because it won't be adding value in the classification model. Separated the feature and target class.



For the classification model in this dataset our target class is "default". This is an imbalanced dataset because the target column contains more number of observations for non-default cases and less number of observations for default cases as observed from the above graph. To handle this imbalance dataset problem I applied the SMOTE technique.

SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem.It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesises new minority instances between existing minority instances. It generates the virtual training records by linear interpolation for the minority class. After the oversampling process, the data is reconstructed and several classification models can be applied for the processed data.

One Hot Encoding

Most Machine Learning algorithms cannot work with categorical data and needs to be converted into numerical data. Sometimes in datasets, we encounter columns that contain categorical features (string values) for example parameter Gender will have categorical parameters like Male, Female. These labels have no specific order of preference and also since the data is string labels, machine learning models misinterpreted that there is some sort of hierarchy in them.One approach to solve this problem can be label encoding where we will assign a numerical value to these labels for example Male and Female mapped to 0 and 1. But this can add bias in our model as it will start giving higher preference to the Female parameter as 1>0 and ideally both labels are equally important in the dataset. To deal with this issue we will use One Hot Encoding technique.

StandardScaler

It standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation. StandardScaler does not meet the strict definition of scale I introduced earlier. StandardScaler results in a distribution with a standard deviation equal to 1. The variance is equal to 1 also, because variance = standard deviation squared. And 1 squared = 1. StandardScaler makes the mean of the distribution approximately 0.

Train Test Split

The train_test_split() method is used to split our data into train and test sets. First, we need to divide our data into features (X) and labels (y). The dataframe gets divided into X_train, X_test , y_train and y_test. X_train and y_train sets are used for training and fitting the model. The X_test and y_test sets are used for testing the model if it's predicting the right outputs/labels. Divided the data into 80% and 20% for training and testing purposes respectively.

Models used for training

Logistic Regression: It is a classification technique used in machine learning. It uses a logistic function to model the dependent variable. The dependent variable is dichotomous in nature, i.e. there could only be two possible classes (eg.: either the user will default or not).

k-nearest neighbors: It is also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications

or predictions about the grouping of an individual data point.

Gaussian Naive Bayes: Naive Bayes is a probabilistic machine learning algorithm used for many classification functions and is based on the Bayes theorem. Gaussian Naïve Bayes is the extension of naïve Bayes. While other functions are used to estimate data distribution, Gaussian or normal distribution is the simplest to implement as you will need to calculate the mean and standard deviation for the training data.

Support Vectors Classifier: It tries to find the best hyperplane to separate the different classes by maximizing the distance between sample points and the hyperplane.

Random Forest Classifier:
Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

Classification Evaluation Metrics

confusion matrix: It is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. Some terms used are explained below

- **true positives (TP):** These are cases in which we predicted yes (they will default), and they do have default.
- **true negatives (TN):** We predicted no, and they are non defaulters..
- **false positives (FP):** We predicted yes, but they don't actually default. (Also known as a "Type I error.")
- **false negatives (FN):** We predicted no, but they actually do default. (Also known as a "Type II error.")

classification report: It is the report which explains everything about the classification. This is the summary of the quality of classification made by the constructed ML model. It

comprises mainly 5 columns and (N+3) rows. The first column is the class label's name and followed by Precision, Recall, F1-score, and Support. N rows are for N class labels and the other three rows are for accuracy, macro average, and weighted average.

- **Precision:** It is calculated with respect to the predicted values. For class-A, out of total predictions how many were really belong to class-A in actual dataset, is defined as the precision. It is the ratio of the [i][i] cell of confusion matrix and sum of the [i] column.
- **Recall:** It is calculated with respect to the actual values in the dataset. For class-A, out of total entries in the dataset, how many were actually classified in class-A by the ML model, is defined as the recall. It is the ratio of the [i][i] cell of confusion matrix and the sum of the [i] row.
- **F1-score:** It is the harmonic mean of precision and recall.
- **Support:** It is the total entries of each class in the actual dataset. It is simply the sum of rows for every class-i.

Conclusion

The accuracy score of all the models tested on the dataset are given below

	Model Name	accuracy
0	logistic regression	0.723304
1	k nearest neighbor	0.754761
2	Naive Bayes	0.543762
3	Support Vector Machine	0.754868
4	Random Forest	0.841108

Out of all the models Random forest gives the highest accuracy.

References

1.Rpubs
2.Kaggle
3.Towards data science
4.Medium
5.Analytics Vidhya
6.Geeks for geeks