Credit Card Default Prediction

Shivraj Saude,
Prerit Tyagi.
Data Science Trainees,
AlmaBetter, Bangalore.

Abstract:

Financial threats are displaying a trend in 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. In classification problems, an imbalanced dataset is also crucial to improve the performance of the model because most of the cases lied in one class, and only a few examples are in other categories. Traditional statistical approaches are not suitable to deal with imbalanced data. There is often as significant difference between the minimum and maximum values in different features, so Min-Max normalization is used to scale the features within one range. Data level resampling techniques are employed to overcome the problem of the data imbalance. Various under sampling and oversampling methods are used to resolve the issue of class imbalance. Different machine learning models are also employed to obtain efficient results. This model will help commercial banks, financial organizations, loan institutes, and other decision-makers to predict the loan defaulter earlier.

Keywords: EDA, Data Analysis, NumPy, pandas, Data visualization, Different types of Machine learning Models.

1. Problem Statement:

Fraud Detection Prediction. 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.

2. Introduction:

The rapid growth in the E-Commerce industry has led to an exponential increase in the use of credit cards for online purchases and consequently they has been surge in the fraud related to it .In recent years, For banks has become very difficult for detecting the fraud in credit card system. Machine learning plays a vital role for detecting the credit card fraud in the transactions. For predicting these transactions banks make use of various machine learning methodologies, past data has been collected and new features are been used for enhancing the predictive power. The performance of fraud detecting in credit card transactions is greatly affected by the sampling approach on data-set, selection of variables and detection techniques used. The performance of the techniques is evaluated for different variables based on sensitivity, specificity, accuracy and error rate. The main idea is by analyzing the customer data and by combining machine-learning algorithm to identify the default credit card user. Default is a keyword, used for predicting the customer who can't repay the amount on time. Predicting future credit default accounts in advance is highly tedious task. Modern statistical techniques are usually unable to manage huge data.

This project possesses various contributions in the domain of credit risk prediction.

1) First, latest dataset has been used to build a machine learning model for credit risk prediction.
2) Second, the data imbalance problem has been explored by comparing the different resampling techniques and evaluate the performance that which the resampling technique has given effective results with a machine learning classifier.
3) Limited work was done on resampling techniques for data balancing in this domain because only a few resampling techniques were employed and also obtained less efficient results.
4) Lastly, the interpretable model is also deployed on the web to ease the different stakeholders. This model will help commercial banks, financial organizations, loan institutes, and other decision-makers to predict the credit defaulter earlier.

3. Dataset Description:

Data is the very prerequisite for any successful machine learning model. No matter how great your machine learning models are, you cannot get a reliable high-performance model from the prediction model without a sufficient amount of rich data.

We have records of 30000 customers. Below are the description of all features we have.

- 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, 0,4,5,6 = others)
- MARRIAGE: Marital status (0 = others, 1 = married, 2 = single, 3 = others)
- AGE: Age in years

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)

- PAY_0: Repayment status in September, 2005 (scale same as 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)

What do 0 and -2 mean in PAY_X columns?

From dataset documentation:

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). Notice code 0 and 2 are in the PAY_X columns but are not included in the data description. Also by doing the value_counts, it shows 0 is the most frequent observation.

-2: No consumption; -1: Paid in full; -0: The use of revolving credit, meaning the payment wasn't due, which makes sense that most customers were using the revolving credit.

A. CLEANING OF DATA :

Cleaning the data involves eliminating the outliers and taking attributes required for feature extraction post Exploratory Data Analysis (EDA).

B. EXPLORATORY DATA ANALYSIS:

In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Data visualization is the graphic representation of data. It involves producing images that communicate relationships among the represented data to viewers of the images. This communication is achieved through the use of a systematic mapping between graphic marks and data values in the creation of the visualization. This mapping establishes how data values will be represented visually, determining how and to what extent the property of a graphic mark, such as size or color will change to reflect changes in value of datum. Distribution of target classes is highly imbalanced, non-defaults far outnumber defaults. This is common in these datasets since most people pay credit cards on time (assuming there isn't an economic crisis). Credit Limit by Sex. The data is evenly distributed amongst males and females.

C. TRAINING MODELS:

Even though there are many machine learning methods available for certain machine learning problems, such as binary classification, for example, each method has its own strengths and weaknesses. Based on our demands and requirements, we may need to choose different methods. The models which we have used are Logistic Regression, SVC, Random Forest and XG Boost.

4. Machine Learning Models :

I. LOGISTIC REGRESSION

Logistic Regression is one of the classification algorithm, used to predict a binary values in a given set of independent variables (1 / 0, Yes / No, True / False). To represent binary / categorical values, dummy variables are used. For the purpose of special case in the logistic regression is a linear regression, when the resulting variable is categorical then the log of odds are used for dependent variable and also it predicts the probability of occurrence of an event by fitting data to a logistic function. Such as

O = e^(I0 + I1*x) / (1 + e^(I0 + I1*x))

(3.1) Where,

O is the predicted output

I0 is the bias or intercept term

I1 is the coefficient for the single input value (x).

II. RANDOM FOREST

The random forest approach is a bagging method where deep trees, fitted on bootstrap samples, are combined to produce an output with lower variance. However, random forests also use another trick to make the multiple fitted trees a bit less correlated with each other: when growing each tree, instead of only sampling over the observations in the dataset to generate a bootstrap sample, we also sample over features and keep only a random subset of them to build the tree. Sampling over features has indeed the effect that all trees do not look at the exact same information to make their decisions and, so, it reduces the correlation between the different returned outputs. Thus, Random forest algorithm combines the concepts of bagging and random feature subspace selection to create more robust models.

III. XG Boost

XG Boost is otherwise as extreme Gradient Boosting which is one of the machine learning boosting classifier models. The XG boost use plot importance() function which is a build in function to generate feature importance, which improves the performance and efficiency by algorithmic optimization and system optimization.

IV. SVC

The objective of clustering is to partition a data set into groups according to some criterion in an attempt to organize data into a more meaningful form. There are many ways of achieving this goal. Clustering may proceed according to some parametric model or by grouping points according to some distance or similarity measure as in hierarchical clustering. A natural way to put cluster boundaries is in regions in data space where there is little data, i.e. in "valleys" in the probability distribution of the data. This is the path taken in support vector clustering (SVC), which is based on the support vector approach. In SVC data points are mapped from data space to a high dimensional feature space using a kernel function. In the kernel's feature space the algorithm searches for the smallest sphere that encloses the image of the data using the Support Vector Domain Description Algorithm. This sphere, when mapped back to data space, forms a set of contours which enclose the data points. Those contours are then interpreted as cluster boundaries, and points enclosed by each contour are associated by SVC to the same cluster.

METHODOLOGY:

1. Exploratory Data Analysis
2. Baseline Model
3. Performance Metrics
4. Optimization
5. Feature Importance
6. Hyperparameter Tuning
7. Class Imbalance
8. Analyse Results

RESULTS:

The proposed hierarchy of the workflow model was loading the data, Cleaning the data, Training the model, Making Predictions, Tuning the hyper Parameters to increase Confidence.

CONCLUSION:

To identify the default payment of credit card clients of huge data set data analysis should be involved. Data analysis allows cultivation and learning based on model build, feature extraction, and various conditions that can improve the trait of customer acquirement. The four machine learning techniques mentioned can analysis the huge data set and to provide the accurate result. The boosting techniques which are included here can perform analysis for imbalanced dataset. By using Predictive analysis model for estimating the default payment and loss of extend and for predicting losses. In this paper, Machine learning technique like Logistic regression, XG Boost, SVC and Random Forest were used to detect the fraud in credit card system. Sensitivity, Specificity, accuracy and error rate are used to evaluate the performance for the proposed system. The accuracy for logistic regression, XG Boost, SVC and random forest classifier are 79%,85%,81% and 85% respectively. By comparing all the three methods, found that random forest classifier is better than the logistic regression and decision tree. XG Boost provided us the best results giving us a recall of 85% percent (meaning out of 100 defaulters 85 will be correctly caught by XG Boost).

FUTURE WORK:

In future work, the model will be tested on other datasets that are transactional and non-transactional in nature to prove its efficiency.

Moreover, the proposed model will be extended to customer credit scoring for consumer loans.

Given more time and resources, we would like to do work on the following:
1. Feature Engineering; coming up with new features given the current dataset that could potentially be better predictors of credit card defaulters

2. Hyper-parameter tunes other models

References:

1. https://www.record-evolution.de/en/blog/credit-default-risk-prediction/

2. https://www.ijrte.org/wp-content/uploads/papers/v8i6/F8897038620.pdf

3. https://www.researchgate.net/publication/344914401_An_Investigation_of_Credit_Card_Default_Prediction_in_the_Imbalanced_Datasets

4. https://medium.com/swlh/predicting-credit-card-defaults-with-machine-learning-fcc8da2fdafb