CREDIT CARD DEFAULT PREDICTION

KISHOR SHIVAJI PATIL
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. 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 a 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 undersampling 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.

INTRODUCTION

The rapid growth in E-Commerce industry has lead 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.

DATA 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.

Table Header	**Second Header**
ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and supplementary credit
SEX	Gender (1=male, 2=female)
EDUCATION	(1=graduateschool,2=university , 3=highschool,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 **…;** **PAY_6**	Repayment status in August, 2005 (scale same as above) Repayment status in April, 2005 (scale same as above)
BILL_AMT1	Amount of bill statement in September, 2005 (NT dollar)
BILL_AMT2 **…;** **BILL_AMT6**	Amount of bill statement in August, 2005 (NT dollar) Amount of bill statement in April, 2005 (NT dollar)
PAY_AMT1	Amount of previous payment in September, 2005 (NT dollar)
PAY_AMT2 **…;** **PAY_AMT6**	Amount of previous payment in August, 2005 (NT dollar) Amount of previous payment in April, 2005 (NT dollar)

Default payment next month	Default payment (1=yes, 0=no)

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

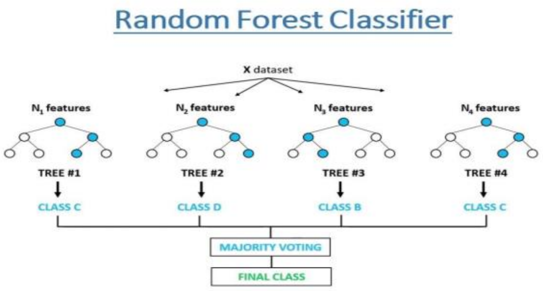
O = e^(I0 + I1*x) / (1 + e^(I0 + I1*x)) (3.1) Where,

O is the predicted output

I0 is the bias or intercept term

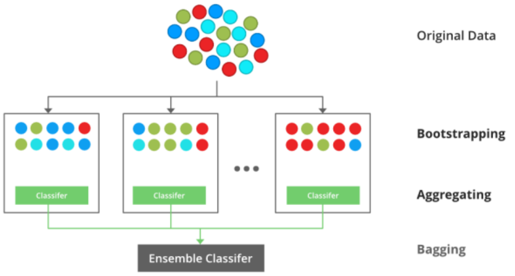
I1 is the coefficient for the single input value (x).

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. XGBoost



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

a. Exploratory Data Analysis
b. Baseline Model
c. Performance Metrics
d. Optimization
e. Feature Importance
f. Hyperparameter Tuning
g. Class Imbalance
h. Analyze 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.

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 XGBoost.

CONCLUSION

- ❖ XGBoost model has the highest recall, if the business cares recall the most, then this model is the best candidate.
- ❖ If the precision Score is the most important metric, then Random Forest is the ideal model.
- ❖ Since Random Forest has slightly lower recall but much higher precision than Logistic Regression,I would recommend Random Forest.
- ❖ There were not huge gap but female clients tended to default the most.
- ❖ Labels of the data were imbalanced and had a significant difference.
- ❖ The best accuracy is obtained for the Random forest and XGBoost classifier.
- ❖ XGBoost Provided Us the Best Results by Giving us Recall of 86% (Means Out of 100 Defaulters 86 Will be Correctly Caught by XGBoost.)