**Technical Documentation of Credit-Card-Default Prediction**

**By- Rajvee Sharma**

**Data Science Trainee at Alma Better**

**Abstract-**

*Aiming at the problem that the credit card default data of a financial institution is unbalanced, this leads to unsatisfactory prediction results. 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.*

*The model effectively solves the problem of sample data imbalance. At the same time, this paper constructs some common machine learning models like logistics, SVM, random forest and Decision Tree and then compares the classification performance of these prediction models. The experimental results show that the proposed algorithm can greatly improve the prediction performance of the model.*

*Traditional 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 create the feature within one range. This model will help commercial banks, financial organization, loan institutes, and other decision-makers to predict the loan defaulter earlier.*

**Introduction-**

*The rapid growth in E-commerce industry has led to an exponential increase in 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. For predicting these transactions banks make use of various ML algorithms, past data has been collected and new features are been used for enhancing the predictive power. 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. *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 result with ML classifier.*

*Limited work was done on resampling technique because only few resampling techniques were employed and less efficient results were obtained.*

1. *Lastly, the interpretable model is also deployed on the web to ease the different stakeholders and many other people.*

**Description of data-**

*Data is 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 |  |
| SEX |  | Gender (1=Male, 2=Female) |  |
| LIMIT\_BAL |  | Amount of given credit in NT dollar |  |
| Education |  | 1=graduate school, 2=university,  3= school, 4=others,  5&6= unknown |  |
| Marriage |  | 1= Married, 2=Single, 3=Others |  |
| Age |  | Age in years |  |
| Pay\_0 |  | Repayment status in September |  |
| Pay\_2,  …  Pay\_6 |  | Repayment status in August  Repayment status in April |  |
| Bill\_Amt\_1 |  | Amount of bill statement in September |  |
| Bill\_Amt\_2  ….  Bill\_Amt\_6 |  | Amount of bill statement in August  Amount of bill statement in April |  |
| Pay\_Amt\_1 |  | Amount of previous payment in September |  |
| Pay\_Amt\_2,  …  Pay\_Amt\_6 |  | Amount of previous payment in August  Amount of previous payment in April |  |
| Default Payment Next Month |  | Default payment (1=yes, 2=no) |  |

**Machine Learning Models**

1. **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). 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 variables and also it predicts the probability of occurrence of an event by fitting data to a logistic regression function. Such as*

*0 = e^ (l0 +l1\*x) / (1+e^ (l0+l1\*x)) (3.1)*

*Where, 0 is predicted output, l0 is bias and l1 is coefficient terms for the single input (x).*

1. **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. Sampling over features has indeed the effect that all tress 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 concept of bagging and random feature subspace selection to create more robust models.*

**3)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.*

1. **SVM**

*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 forum. There are many ways of achieving this goal. Clustering may proceed according to some parametric model or by grouping points according to some distances or similarity measure as in hierarchical clustering. This is 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 data using Support Vector Domain.*

**Methodology**

1. *Exploratory Data Analysis*
2. *Baseline Model*
3. *Performance Metrics*
4. *Optimization*
5. *Feature Importance*
6. *Hyper parameter Tuning*
7. *Class Imbalance*
8. *Analyze Results*

**Result-**

*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.*

1. **Cleaning the data**

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

1. **Exploratory Data Analysis**

*In statistics, EDA 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.*

1. **Training Model**

*Even though there are many machine learning methods available for certain machine learning problem, 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**

*To identify the default payment of credit card client 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 technique 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 techniques like logistic regression, XGBoost , 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 proposed system.*

**Reference-**

*1. https://www.researchgate.net/publication/344914401\_An\_Investigation\_of\_Credit\_Card\_Default\_Prediction\_in\_the\_Imbalanced\_Datasets*

*2. https://www.kaggle.com/code/ainslie/credit-card-default-prediction-analysis/notebook*