**Credit Card Default Prediction**

**Kartike, Animesh Chakraborty, Anupam Mishra**

**Data science trainees,**

**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 faced 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.

With the vigorous development of the financial sector, financial risks are showing a tendency of diversification, which the customer credit risk of commercial Banks in particular. As a result, the customer credit risk is generally taken into account by financial institutions, and the credit evaluating model emerges as the times require. At present, many researches have concentrated on enhancing the precision of the model, ignoring the interpretability, which makes it difficult to apply in industry; Compared to Precision, the study of interpretable model is relatively small, moreover, due to imbalanced data sets, the accuracy of existing models is not high .

To test performances of the model, we use the real Taiwan credit card customer datasets for the empirical research .Based on the open data set of credit card in Taiwan, Six data mining methods, Logistic regression, SVC, KNN, Decision Tree, XGboost and Random Forest, are compared in this paper.

**Keywords:**-Machine learning , Logistic Regression, gradient boosted decision tree, Random forest Correlation Analysis, customer credit risk, K-Nearest Neighbor, Support Vector Classifier, credit card default model, Taiwan.

**Problem Statement:**

In general, we can refer to a customer’s inability to pay, or their default on a payment, or personal bankruptcy, all as potential issues of non-payment. However, each of these scenarios is a result of different circumstances. Sometimes it is due to a sudden change in a person’s income source due to job loss, health issues, or an inability to work. Sometimes it is deliberate, for instance, when the customer knows that he/she is not solvent enough to use a credit card anymore, but still uses it until the card is stopped by the bank. In the latter case, it is a type of fraud, which is very difficult to predict, and a big issue to creditors.

To address this issue, credit card companies try to predict potential default, or assess the risk probability, on a payment in advance. From the creditor's side, the earlier the potential default accounts are detected the lower the losses. For this reason, an effective approach for predicting a potential default account in advance is crucial for the creditors if they want to take preventive actions. In addition, they could also investigate and help the customer by providing necessary suggestions to avoid bankruptcy and minimize the loss.

**Introduction:**

Credit card default prediction is based on the historical data of credit card customers. The use of corresponding methods to predict and analyze credit card customer default behavior is a typical classification problem. Data mining algorithms have long been applied to the study of credit card default prediction problems. We have used discriminate analysis to score the credits and behaviors of borrowers and used Logistic regression, decision trees, Random Forest and other algorithms to predict customer default payments in Taiwan, and compared the predictions of these algorithms. We have explored the key factors affecting customer credit by using Logistic Regression models, random forest models and other models. The results show that the accuracy of random forest prediction is higher than that of other models.

To test performances of the model, we use the real Taiwan credit card customer datasets for the empirical research. This dataset contains features which are in the form of table:

| Variable | Feature description |
| --- | --- |
| X1 | Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit. |
| X2 | Gender (1 = male; 2 = female). |
| X3 | Education (1 = graduate school; 2 = university; 3 = high school; 4 = others). |
| X4 | Marital status (1 = married; 2 = single; 3 = others). |
| X5 | Age (year). |
| 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. |
| 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. |
| 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. |
| Default | 1 for fraudulent transactions, 0 otherwise. |

**Steps involved:**

* **Exploratory Data Analysis**

Exploratory data analysis is a statistical way of understanding the data which is usually done in a visual way. The graphs plotted in exploratory data analysis are for better understanding of data to the analyst.

After loading the dataset we performed this method by comparing our target variable that is Default Payment with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

* **Null values Treatment**

After the data is loaded, the missing data is checked using is.na() or isnull() function .The output depicted that there were no missing values in our dataset.

So our dataset does not contain any missing values.

We have also checked there is no duplicate value in our dataset.

* **Encoding of Categorical features**

We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

Categorical variables- Sex, Education and Marriage- were converted coded into numerical depictions to fit our Model to predict Credit card defaulter.

* **Feature Engineering**

To make the data tenable for understanding and further analysis , the data set was analyzed for identifiable statistical trends and patterns. After preliminary analysis, the following steps were undertaken to transform the data into a systematically workable dataset:

In the data preprocessing the first step to be done is cleaning of the data. First we check for null values and then we check if some categories of variables are mislabeled or undocumented. Then the categories are labeled correctly so that it does not lead to overfitting of models in future. Renaming of certain variables also takes place so that it is simpler to understand for future processing.

**Oversampling Methods:**

SMOTE (Synthetic Minority oversampling approach) is an oversampling approach in which rather than replacing the existing samples here new samples are synthetically generated. The algorithm for SMOTE is:

1. We find the k-nearest neighbors for our minority class instance Y by applying Euclidean distance to find the distance between instance Y and others in the same class.

2. Based on the sampling ratio, an instance among k - nearest neighbors is randomly selected. Here Yn, represents the selected nearest samples which vary from 1 to sampling ratio(N).

3. Here the new synthetic data is created by taking the difference between chosen k-nearest neighbor and current data and then multiplying it with the random numbers which vary from 0 to 1.

Ysyn = Y+ (Yn – Y ) \* rand

Here rand - stands for random numbers

* **Standardization of features**

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

* **Fitting different models**

For modeling we tried various algorithms like:

1. **Logistic Regression**
2. **Decision Tree Classifier**
3. **Random Forest Classifier**
4. **eXtreme Gradient Boosting Classifier**
5. **Support Vector Classifier**
6. **K-Nearest Neighbor**

After performing the various models we get the best recall from the KNN and best F1 score from the SVC classifier.

* **Algorithms:**

**1.Logistic Regression**

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

It is used in statistical software 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.

**2.Decision Tree**

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

**3.Random Forest**

Random forest is an almighty tool which ensembles decision trees and bagging .The base learner of random forests is a binary tree constructed by recursive partitioning (RPART) and then developed using classification and regression trees. Binary splits of the parent node of a random forest splits data into two children’s nodes and increases homogeneity in children nodes compared to parent nodes. Note that a random forest does not split tree nodes based on all variables; instead, it chooses random variable subsets as candidates to find the optimal split at every node of every tree. Then the information from the n trees is aggregated for classification and prediction. Random forests also provide the importance of each feature by accumulated Gini gains of all splits in all trees representing the variable discrimination ability:

Import =1#trees∑v∈xjGain(xj,v)

Where Gain(xj,v) is the gain of the Gini index of feature xj combined with node v.

**4. eXtreme Gradient Boosting Classifier**

XGBoost is a machine-learning system centered on the “lifting tree” proposed by Chen, on the basis of a great deal of previous research work on the gradient lifting algorithm. It contains a set of iterated residual trees. Each tree is the residual of the N-1 tree before learning. Adding the output values of new samples predicted by each tree produces the predicted value of the final sample (2016). Nonetheless, unlike the commonly used gradient-boosting decision tree (GBDT), which only uses the first derivative information in optimization, XGBoost expands the cost function in the second-order Taylor expansion and uses the first and second derivatives at the same time, which allows XGBoost to obtain credible results. Its main characteristics are as follows.

1) By supplementing the loss function with a regularization term, XGBoost acquires a formidable anti-overfitting characteristic.

2) The second-order Taylor expansion is used to make the loss function more accurate.

3) The efficacy of the model iteration is greatly enhanced by the qualities of other parallel operations.

4) XGBoost supports column sampling, which reduces overfitting, reduces calculation and improves iteration efficiency.

Because of its advantageous algorithm, XGBoost has been deployed in the financial sector and elsewhere with increasing frequency in recent years. The present context allows its popularity and practicality to be more objectively assessed. Therefore, this paper takes XGBoost as its basic tool to complete the users’ credit-card characteristic analysis.

**5. Support Vector machines**

A Support Vector Machine (SVM) is a supervised model which can be used for regression as well as classification tasks. The algorithms associated with SVM are used to recognize and analyze different patterns for the purpose of classification. The fundamental idea behind SVM is finding an optimal hyperplane that could separate two classes instances which are support vectors. But as some data are linearly inseparable, a new concept of kernel functions was introduced. These functions are used to transform a low dimensional input space into high dimension input space which means that it is used to solve the problem of linearly inseparable data. In this paper we are making use of a kernel function known as radial basis function (RBF) to solve complex problems.

The RBF kernel makes use of two important parameters such as C and gamma where C is the penalty. Parameter which by default is kept as 1.0 and gamma is a kernel coefficient.

1. **K-Nearest Neighbor**

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. The K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using K- NN algorithm. The K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

The KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

**Model Performance:**

Model evaluation is of paramount importance in any predictive modeling task. It becomes even more critical in ensemble predictive modeling, where the relative performance and diversity of models must be thoroughly evaluated. All the evaluation metrics are built on four types of classiﬁcations:

**1.Accuracy**

Typically, accuracy is used to assess the effectiveness of a model with the help of the confusion matrix. The accuracy of the model has been computed through.

Accuracy = (TP + TN)/(TP + TN + FP + FN)

**2.Precision**

Precision compares the number of true positives to the number of true positives and the number of false positives. That is,of all the instances the classiﬁer said were positive, precision measures how many of them were positive. The Precision of the model has been computed through

Precision = TP/(TP + FP)

**3.Recall**

Recall compares the number of true positives to the number of true positives and false negatives. The Recall of the model has been computed through

Recall = TP/(TP + FN)

**4.F-Measure**

F-Measure combines precision and recall as the harmonic mean. The precision and recall trade-off with each other:higher precision generally associated with low recall. The Value of F-Measure has been computed through

F-Measure = (2\*Precision\*Recall)/(Precision + Recall)

**5.Receiver Operating Characteristic (ROC) CURVE**

A receiver operating characteristic (ROC) curve plot is also a widely used measure to evaluate the performance of classiﬁers. Speciﬁcally, the plot is created by plotting the true positive rate (recall) against the false positive rate at various threshold levels.

**5.Conclusion**

We estimated six models, Logistic Regression, Decision Tree, Random Forest, SVC, KNN and XGB, and reviewed their results using performance metrics such as Accuracy, Recall, Confusion Matrix, ROC curve/ AUC-Score, and F1-Score. We found that after performing the various models the best precision-recall balance from the SVC and best F1 score from KNN classifier.

▶ KNN has the highest F1 score as compared to other models performed.

▶ SVC has the best precision-recall balance.

▶ Higher recall can be achieved if low precision is acceptable.

▶ We can deploy the model and can serve as an aid to human decision.

▶ Model can be improved with more data and computational resources