Credit Card Default Prediction using SMOTE and Machine Learning Algorithms

Anvesha Dasgupta
Department of Electronics
R V College of Engineering
Banglore, India
anneshadasgupta.ei17@rvce.edu.in

Dona Ghosh
Department of Mathematics
Jadavpur University
Kolkata, India
goshdona51@gmail.com

Jatin Vyäs
Department of Data Science
S P Jain Global School of Management
Mumbai, India
jatin.bj19bdv009@spjian.org

Abstract— A matchless way to make payments broadly used across the globe is via a credit card. All the credit cards applications don’t get approved by the banks, multiple factors of the customer such as credit bureau, income, demographics etc. are taken into consideration for approval of a credit card. Over few years the credit card companies are encountered with a credit loss. So as to avoid possibility of such risk, this analysis proposes a machine learning model for prediction of default payments considering real world user credit card data. After analyzing crucial factors for credit risk from the credit card provider’s historic data, policies can be modeled to reduce the risk and assess the financial gain which the business may procure from the model. The problem of imbalance class in the dataset is abolished with help of balancing algorithms namely SMOTE and ADASYN. The effectiveness of the model is increased with a balanced dataset. Later this balanced data set is given as an input to machine learning algorithms like random forest, SVM to predict default credit cards. Accuracy, Confusion Matrix, F-Score and Precision are been calculated for performance comparison of the machine learning models. The model with best accuracy is implemented to predict credit card default.

Keywords—Machine Learning Classifiers, Credit Card Default, SVM, Random Forest, SMOTE.

I. INTRODUCTION

The credit card emerged into the standard payment as use of credit card increases the purchasing power of the customer. With consumption of credit cards is expanding day by day the number of defaulters also increasing. The customer who miss out making the minimal payment for months is termed as default customer. In comparison with traditional loans the dues reimbursement in credit card loan is minimal in comparison with the credit balance which allocates a greater risk on the lenders. The state economy is extremely impacted with default cards which evoked many researchers to investigate this issue as even a minute increase in the accuracy will lead to massive economic gains to the credit card providing companies.

Data mining includes various techniques which can be used for inspecting available data and modify them into required format. It is also used for recognizing patterns and relationship among dissimilar data on large datasets. In recent time, banking sectors have discovered its relevance with data mining. The foremost objective of applying data mining techniques is to build predictive models. For controlling risk, it is important to have default prediction of credit cards. The aim of this research is to increase the accuracy and the precision of the prediction model by balancing the dataset during data pre-processing stage.

II. LITERATURE SURVEY

The banking sector have incurred a massive loss a short time ago on account of borrowers disinclined to pay off their dues leading to bad debts. The primary reason for this is unsuitable decisions and acknowledging credit card applications to any un qualified credit card applicant. Risk control and recognition of credit risk with customer requires handling of significant quantity of data with large variety of features. Now a days a consequential ultimatum bank manages face in distinguishing of potential defaulters as providing of credit cards to unsuitable applicants has led to huge losses for banks. Furthermore, customers with reimbursing capacity but stockpile heavy credit and over exhaustion of the credit limit also can lead to dreadful losses for the banks. Credit cards rotates around credit, the amount to be paid fluctuates every month. Therefore, reasonable monitoring of these features can assist the banks to keep an eye on the accounts that have the probability of default and accordingly take required measures. Integrating machine learning techniques for prediction of defaulters can be beneficial for the banks in scheming preventive measures and thus avoid losses.

Machine learning is programming of computers to optimize a performance measures using past historic data or experience. Machine Learning is used to make precise recommendations based on observations and predictions. Machine Learning examines the areas of algorithms that can make high-end predictions on data [3]. The learning process in machine learning is classified into Training and Testing. The model is built on the training data and the model is validated with the testing data.

A credit card amount that is not retrieved for a spell of more than 90 days is observed as non-recoverable or a default. The forecast is therefore executed for this type of dependent variable. These credit cards accounts are then classified as bad accounts and good accounts. Use of various machine learning techniques are required for the prediction of credit risk. Many researchers work is summarized below:

Dananis et al [1] has described a technique for credit risk estimation which uses linear support vector machines classifiers. It uses particle swarm optimization technique so as to concentrate on imbalanced data. This model has inflated accuracy but also it is less stable in contrast to other techniques.

Ling-ying et al [2] considered several techniques to abstract features from input variables to assess good from bad accounts. Utilizing these features, a new and more complex impression of credit cardholder’s behavior can be predicted.
Singh et al [10] in this research feature selection and feature reduction techniques were implemented to analyze the capabilities of base classifiers. Feature selection techniques such as Feature filter and Wrapper based techniques were used to get least number of features by eliminating unused features. These features are then applied to various machine learning algorithms out of which SVM had highest accuracy.

Koutanæi et al [11] has suggested a hybrid data model were feature selection in addition to algorithm classification is executed in three separate phases. In the first phase the typical data preprocessing is completed. In the second and third phase four feature selection algorithms namely genetic algorithm, information gain ratio and relief attribute evaluation have been implemented. At this moment feature selection methods criteria are set based on the accuracy of various classification algorithms and then the feature selection method which is the best for a specific classification is used. This hybrid model gives results with good accuracy.

In Synthetic minority oversampling technique (SMOTE) the evaluation is done by increasing the subordinate sample class although in Tomek link technique the number of dominant class is diminish so that it becomes equal to the minority class. When both SMOTE and Tomek Link methods are combined together along with svm classification method. It was found at that the accuracy when both the sampling techniques were combined were better as compared to when they were individually tested [7].

III. PROPOSED METHODOLOGY

The proposed model is about predicting whether a credit cardholder will default using SMOTE and various machine learning classifiers. The first step in building our prediction model is to do data cleaning. After cleaning of the dataset, now the dataset is divided into two individual set that is the training set and the test set. The next step is to apply multiple data mining algorithms. Later on, data balancing techniques such as SMOTE and ADASYN are applied on the training data and then the training dataset is classified using machine learning algorithms in order to predict whether the credit card is defaulted or not. Finally, the accuracy of this model is calculated.

1. Data Collection

The dataset which is used in this analysis is downloaded from UCI repository website. Taiwan’s credit card user data has been used for in this paper. This dataset consists a total of 24 features and 24000 rows. For predicting if a credit card is defaulted or not, these datasets contains a binary column which is labeled as 1 if the card is defaulted otherwise it is labeled 0. Rest of the features are explanatory variables which helps in prediction purpose.

2. Data Pre-Processing

Data pre-processing operations are performed to enhance the quality of default input datasets and to assess mis representation of data in the dataset. The first step in data preprocessing is to clean the data, for this step null values are to be checked for continuous variables and then categorical variables are checked for mislabeled values in the dataset. Adding correct labels to the categories is important as it helps the model from getting overfit. Renaming of some variables also takes place because it is simpler to understand during processing.

3. Splitting Dataset

Dataset Splitting appears essential step to eradicate bias to training data for machine learning algorithms. Transforming variables of a machine learning algorithms to best fit the training data usually results in an overfitting of algorithm that performs badly on actual test data. For this purpose, the input dataset is split into multiple distinct subsets on which the machine learning algorithms are trained with different variables.

3.1. Training Dataset

Machine learning algorithms uses the observations from the training dataset to learn. In supervised learning algorithms, each observation also includes one observed output variable and one or more observed input variables. In unsupervised machine learning algorithms, there is no observed output variable. The algorithm learns the training dataset and identifies patterns and clusters.
3.2. Test Dataset

A set of observations that are used to assess performance of the model by applying some performance measures. In the test dataset, its essential that no observations should be present from the training set. If this situation occurs it will be unfavorable to evaluate whether the model is capable to make predictions based on its learnings or not.

4. Data Balancing Methods

The most important element of any machine learning model is features and observations. The machine learning algorithms are essential but the data is believed to be most essential. Hence, the data preprocessing is very crucial in machine learning. The dataset used in the analysis contains very less observations where the credit card is defaulted as compared to non-defaulted observations. The class which has a less number of observations is known as minority class and the class which has a more number of observations are known as majority class. Because of this imbalance data, many machine learning algorithms are inclined towards the majority class due to which the observations under minority class aren’t classified accurately which affects the overall models. In this research, to overcome the problem of imbalance class we make use of oversampling methods namely SMOTE and ADASYN.

4.1. SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a sampling algorithm that generates artificial data points of the minority class in place of sampling existing observations with substitution [8]. This method is tried on the continuous variables for each selected in the minority class. There are some instances that should be observed when implementing SMOTE method with cross-validation.

The algorithm for SMOTE is:

a) Take the k nearest neighbors which belongs to the identical class of the observed sample.

b) Randomly select n samples of these k neighbors. The number n is based on the prerequisite of the oversampling, if 100% over-sampling is needed, then n = 1.

c) Evaluate variation among the feature vector of the observed sample and the individual n neighbor feature vectors.

d) Each of the variation is multiplied with a random number between 0 and 1.

e) Add these numbers separately to the feature vector of the observed sample so as to achieve n new synthetic observations.

f) Return the new synthetic samples

4.2. ADASYN

Adaptive synthetic minority oversampling technique is reviewed as reduction to the SMOTE algorithm. In this method the minority samples are generated on the basis of density distribution. The contrast between SMOTE and ADASYN is that in SMOTE the total number of mocked data created will always be equivalent for every minority sample. Although, the ADASYN algorithm utilizes density function which determines how many numbers of mocked data has to be produced for every minority instance. This algorithm issues distinct weights to individual minority samples to create mocked data. ADASYN is fundamentally used for creating mocked data observations for the minority class observations which were difficult to learn.

5. Classification Algorithms

5.1. Random Forest

Random Forest is a supervised learning algorithm, which makes use of labelled data and learns to analyze any input unlabeled data. Random Forest is used to answer classification along with regression problems, by making it a diversified model. Random Forest is a collection of huge number of decision trees that can be used as ensemble methods. Random Forest also assists bagging where the building of each tree is constructed as uncorrelated forests with feature randomness. Random Forest observes difference from decision trees, DT is built with all features on the complete dataset while random forest selects the features randomly to build the tree and to result the outcome.

5.2. K Nearest Neighbor

The K-nearest neighbors (KNN) algorithm is an uncomplicated and easy-to-implement supervised machine learning algorithm that can be used to solve both regression and classification problems. KNN algorithm makes use of similarity among the features to predict the values of new datapoints which further means that the new data point will be allocated a value based on how strictly it matches the
5.3. Ada Boost

AdaBoost algorithm is an ensemble learning with contemporary technique to solve complex classifications in combining simple weak classifiers to strong classifiers. Ensemble learning models are outlined on two techniques that is Boosting and Bagging. AdaBoost is one of the boosting algorithms where the weights of the individual observation are iteratively discovered on depending the accuracy of the last classification result. AdaBoost is highly accurate machine learning classifier which has error rate close to zero. AdaBoost appears to be sensitive with corrupt data and outliers. AdaBoost performs two steps, the classifier is trained respectively on various weighed training data and then in each iteration it yields good fit for the observations by reducing the training error. AdaBoost can be used with any base classifiers as its not vulnerable to overfitting.

IV. RESULT AND DISCUSSION

The actual dataset is split into training and testing datasets which comprises of 80% of the actual data as training and the remaining 20% data as test dataset. SMOTE and ADASYN balancing techniques are applied on the training dataset. The training dataset is required to balance as the frequency of observations which are defaulted are less as compared to the frequency of observations that are non-defaulted credit cards.

<table>
<thead>
<tr>
<th></th>
<th>Number of Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default Credit Card</td>
<td>5328</td>
</tr>
<tr>
<td>Non-Default Credit Card</td>
<td>18672</td>
</tr>
</tbody>
</table>

Table. 1. Training Observations

The SMOTE algorithm is implemented on the training dataset in such a way that the minority sample which is the defaulted credit cards of the dataset becomes equivalent to the majority class. This algorithm creates new mocked data by drawing both default and non-default samples.

<table>
<thead>
<tr>
<th></th>
<th>Number of Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default Credit Card</td>
<td>18672</td>
</tr>
<tr>
<td>Non-Default Credit Card</td>
<td>18672</td>
</tr>
</tbody>
</table>

Table. 2. Training Observations using SMOTE

SMOTE and ADASYN balancing techniques makes use of the same formula for generating mocked observations. The only difference is that ADASYN method creates more mocked observations for the instances that were complex to learn depending upon the weights which are assigned to minority instances. After applying this technique, we get the following results.

This research makes credit card default models using three machine learning algorithms Random Forest, K-Nearest Neighbor and AdaBoost. The main objective of the research is to compare and predict the best accuracy for any given input dataset.

Table. 3. Training Observations using ADASYN

![Proposed Model Credit Card Default Prediction](image)

Fig. 3. Proposed Model Credit Card Default Prediction

The primary benchmarks in comparing the execution of the classifiers is to calculate the effectiveness of the algorithms.

- Confusion Matrix

Confusion Matrix is the most common way to assess the performance of the model with binary responses. The default observations are defined as positives and non-default observations as negatives [10]. When the defaulted credit cards are been predicted to be default by the model then this type of possible outcomes are called as True Positives (TP). True Negatives (TN) outcomes are when non-default credit cards are predicted to be non-default. False Positives (FP) outcomes are when the non-default credit cards are predicted to be defaulted. False Negatives (FN) outcomes are when the default credit cards are predicted to be non-default.

![Confusion Matrix](image)

Fig. 4. Confusion Matrix
Precision estimates the output quality of the model by evaluating the below mentioned formula. Precision is calculated by dividing the true positive outcomes to the sum total of true positive and false positive values. It is a measure of result relevancy [9].

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

- **Recall**

Recall is measure to find the output quality to find how many true relevant results are obtained. Recall is calculated by dividing the true positive values by the sum total of true positive and false negative values of the observations. It is a measure of sensitivity.

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

- **F1 Score**

F1 score is calculated as the weighted average of recall and precision. It provides with the single value that balances precision and recall.

\[
\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

- **Accuracy**

Accuracy is described as the ratio of the number of samples correctly classified by the classifier to the total number of samples for a given test data set.

\[
\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN}
\]

The performance measures precision, recall, f1 score and accuracy are calculated for Random Forest, KNN and AdaBoost algorithms, respectively.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.608</td>
<td>0.404</td>
<td>0.486</td>
<td>0.813</td>
</tr>
<tr>
<td>KNN</td>
<td>0.272</td>
<td>0.550</td>
<td>0.364</td>
<td>0.582</td>
</tr>
<tr>
<td>Ada-Boost</td>
<td>0.629</td>
<td>0.40</td>
<td>0.489</td>
<td>0.818</td>
</tr>
</tbody>
</table>

Table 5. Performance of Algorithms using SMOTE

Table 5 and Table 6 shows the performance measures of the algorithms after the sampling methods namely SMOTE and ADASYN have been applied on the training datasets.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.610</td>
<td>0.402</td>
<td>0.485</td>
<td>0.814</td>
</tr>
<tr>
<td>KNN</td>
<td>0.265</td>
<td>0.590</td>
<td>0.366</td>
<td>0.554</td>
</tr>
<tr>
<td>Ada-Boost</td>
<td>0.60</td>
<td>0.411</td>
<td>0.488</td>
<td>0.812</td>
</tr>
</tbody>
</table>

Table 6. Performance of Algorithms using ADASYN

From the above tables we can analyze that the accuracy of the algorithms when applied on the unbalanced dataset is impartially higher than the accuracy on the balanced dataset. Out of the three machine learning models AdaBoost had highest accuracy for unbalanced data. After balancing the dataset using SMOTE and ADASYN it can be seen that the Recall and F1 Score values have both increased.

V. CONCLUSION

The use of machine learning techniques for the prediction of credit card defaulters is crucial for the recognition of credit risk. This will assist the financial institutions in scheming their future policy. In this research, data is cleaned first to remove uncollaborated and falsely labeled categories of the variables to avoid overfitting of the model. Oversampling techniques like SMOTE and ADASYN are implemented to balance the data. Machine learning methods such as Random Forest, K-nearest neighbors and AdaBoost have been used for classification purpose. Here we have analyzed the performance of three classification methods by applying it on both balanced and unbalanced dataset. After, comparison is also done between the two sampling methods.
methods. In further research, we will try to conduct experiments on larger data sets and try to tune the model with other classification algorithms so as to achieve exceptionally high performance of the model.

VI. PAPER REFERENCES


