ISSN: 2664-9640

Minimization of risk and maximization of profit on behalf of bank by using machine learning algorithm

Tande Tadiwanashe Wendy, Hu Yue *

Zhejiang University of Science and Technology, China;

* Corresponding Author

Email addresses: tadiwanashetande@gmail.com (Tande Tadiwanashe Wendy),

huyue@zust.edu.cn (Hu Yue)

Abstract

The main aim to develop a loan approval prediction model for a bank by utilizing machine learning algorithms. [Objective] The primary objective is to minimize the bank's loss by developing a decision rule for approving or rejecting a loan application based on an applicant's demographic and socio-economic profile. [Method] Three machine learning algorithms, namely Logistic Regression, Random Forest, and Decision Tree, were applied to a German credit dataset to achieve this objective. The data were preprocessed, including converting the target variable to binary, converting character variables to factors, and selecting only numeric variables. The accuracy, sensitivity, and precision of the models were evaluated using cross-validation, and the results were compared. [Result] The Random Forest model showed the highest accuracy, with an average accuracy of 76.07%, followed by the Decision Tree model with an average accuracy of 71.99%, and Logistic Regression with an average accuracy of 71.62%. The sensitivity evaluation showed that the Decision Tree model had the highest sensitivity, with an average sensitivity of 70.88%, followed by the Logistic Regression model with an average sensitivity of 68.35%, and the Random Forest model with an average sensitivity of 65.86%. [Conclusion] In conclusion, the Random Forest model showed the highest accuracy, while the Decision Tree model had the highest sensitivity. However, all three models showed similar precision scores. Therefore, based on the objective of minimizing the bank's loss, the Decision Tree model may be more suitable as it has higher sensitivity, which means it has a lower chance of approving a loan for a potentially risky applicant. However, further research and testing are necessary before implementing these models in real-world applications.

Keywords

Machine learning, logistic Regression, Decision Tree, Credit Card Fraud.

1. Introduction

Most of the time, a credit card is a card that is given to a customer (the cardholder) and lets them buy goods and services up to their credit limit or get cash in advance. Credit cards give the cardholder the benefit of time, in that customers can pay later in a set amount of time by carrying it over to the next billing cycle. Credit cards are easy to steal from. In a short time and without any risk, a lot of money can be taken out without the owner's knowledge. Fraudsters always try to make every fraudulent transaction look like it was done legally, which makes it hard to spot fraud. In 2017, there were 1,579 data breaches that exposed nearly 179 million records. Credit card fraud was the most common type, with 133,015 reports, followed by fraud related to employment or taxes with 82,051 reports, phone fraud with 55,045 reports, and bank fraud with 50,517 reports, according to statistics from the FTC [1]. Since frauds, especially credit card frauds, have been in the news a lot in the past few years, most people around the world are very aware of them. The data set for credit cards is very unbalanced because there will be many more legitimate transactions than fraudulent ones. As technology improves, banks are switching to EMV cards, which are smart cards that store their information on integrated circuits rather than on magnetic stripes. This has made some payments made with cards safer, but frauds that happen when the card is not present are still happening at a higher rate. As chip cards have made transactions more secure, criminals have turned their attention to CNP transactions, according to a 2017 report from the US Payments Forum.

2. Related Work

In paper [2], fraud with credit cards or any other source of money used in a transaction is talked about. In this paper, we talk about different ways to commit fraud. To figure out how they work, we use algorithms like machine learning, genetic algorithms, and neural networks. [3] suggests a different way to do things by using the Hidden Markov Model (HMM). [4] looks at Bayesian decision theory and uses cost-sensitive classification to find credit card fraud. Using a mix of Hidden Markov Model, Behavior Based Technique, and Genetic Algorithm, [5] shows a new way to stop fraudulent transactions. The problem of finding credit card fraud seems hard to solve from a learning point of view because there is a big difference between the number of real transactions and the number of fake transactions [6]. In this field, KNN has been used a lot. The training data are used by supervised learning to classify the dataset, while the clustering technique is used by unsupervised learning. Recently, Auto Encoder and Restricted Boltzmann Machine were used with deep learning to try to solve this problem. The authors concluded that supervised learning works better for finding credit card fraud in a historical database [7]. Bayesian minimum risk theory is used to figure out how Under sampling affects the posterior probability of a machine learning model. Random Forests and Support Vector Machines are both talked about in [8]. The system used in paper [9] is based on a cost-sensitive decision tree. In this paper the main aim of author to minimize loss from the bank's perspective, the bank needs a decision rule regarding who to give approval of the loan and who not to. An applicant's demographic and socio-economic profiles are considered by loan managers before a decision is taken regarding his/her loan application.

3. Machine Learning Algorithm

Learning Machines In general, machine learning is a branch of artificial intelligence that enables a system to learn from experience automatically and without human intervention, with the goal of predicting future outcomes as accurately as possible using various algorithmic models. [10] Machine Learning differs significantly from traditional computation approaches in which systems are explicitly programmed to calculate or solve a problem. Machine learning is concerned with the input data used to train a model, in which the model learns various patterns in the input data and applies that knowledge to predict unknown results. Machine learning has a plethora of applications. [11] It is used in a variety of applications such as spam filtering, weather forecasting, stock market forecasting, medical diagnosis, fraud detection, autopilot, house price prediction, face detection, and many others. Machine learning is typically classified into three types: supervised, unsupervised, and reinforcement learning. This thesis is about supervised learning, which we will go over in the following section. For the time being, we can define supervised learning as a method in which the model is trained with both input and output labels. Unsupervised learning, on the other hand, is when the dataset has input labels (i.e., a model is trained with unlabeled data) from which it learns different patterns and structures. It is typically used in applications such as visual recognition, robotics, and speech recognition.

3.1. Supervised Learning

Supervised learning is a machine learning approach in which the model is given both input and output labels to train. The supervised model trains labelled input and output data and extracts patterns from the input data. These extracted patterns will be used to support future decisions. Supervised learning can be formalized as follows:

$$Y = f(x) \tag{1}$$

where x is an input variable, Y is an output variable, and $f(X)$ is a mapping function.

The goal is to approximate the mapping function so that it can correctly predict the output variable (Y) when given an unknown input. Furthermore, there are two types of supervised learning: classification and regression. The output variable in a classification problem is a category (e.g., fraud or genuine, rainy or sunny, etc.). The output variable in a regression problem is a real value (e.g., the price of a house, temperature, etc.). This thesis only addresses the classification issue.

3.2. Classification

In machine learning, [12] a classification problem is defined as the task of predicting the class label of a given data point. Fraud detection, for example, can be identified as a classification problem. The goal in this case is to predict whether a given transaction is fraudulent or genuine. In general, there are three types of classification: binary classification (e.g., classifying a transaction as either fraudulent or genuine), multi-class classification (e.g., classifying a set of images of flowers as Rose, Lilly, or Sunflower), and multi-label classification (where the data samples are not mutually exclusive, and each data sample is assigned a set of target labels). This thesis is concerned with the binary classification problem, in which the output label is either normal or fraudulent.

3.2.1. Resampling approach

Most predictive models perform poorly in the presence of an unbalanced class distribution. As a result, some data preprocessing must be performed prior to providing data as an input to the model. In the case of a class imbalance problem, such data preprocessing is carried out using a data level approach known as resampling. There are three types of resampling methods: under sampling, oversampling, and hybrid.

3.2.2. Bagging

Bagging, an abbreviation for Bootstrap Aggregation, is a straightforward ensemble method with considerable power. Bootstrapping is used to produce additional training samples from the original training set by randomly replacing some of the data. Bootstrap training samples are the new sets of data used for learning. The prediction models are trained using data from each bootstrap sample independently. Finally, we take an aggregate view of the predictions by either averaging (in the case of regression) or voting on the results from each of the bootstrapped models (for classification).

3.2.3. Boosting

The process of "boosting" is yet another potent ensemble method. A weak learner, also known as a basic learner, is combined with other weak learners to form a strong learner that can produce greater outcomes than those produced by a single learner. Boosting trains, the weak learners sequentially, with each learner trying to correct its predecessor by adding extra weights to the samples that were previously misclassified, in contrast to bagging, where each model runs in parallel and then the outputs are merged at the end. Consequently, the misclassified situations will become increasingly important to the prospective poor learner.

3.3. Selected Models

In this section, we will discuss the different models selected for the predictive analysis. Depending on the nature of the classification problem, we chose three very popular predictive model. They were logistic regression and decision tree.

3.3.1. Logistic regression

One of the most widely used classification methods in machine learning is logistic regression. Despite having "regression" in the title, this is not a regression method. In order to solve the regression problem, logistic regression was developed from another widely used machine learning algorithm, linear regression. The prediction in logistic regression is the percentage of outcomes that will fall into each class. A linear regression model makes predictions for real-valued outcomes by integrating input variables (x) with weights. If it helps, imagine there is only one input (or independent) variable (x) and one output (or dependent) variable (y). When x is the only input variable, the hypothesis of linear regression can be written as

$$y = a_o + a_{1x} \tag{2}$$

where a0 is a bias term and a1 is the weight for x. During training, you'll pick up these kilos. It's possible here that the hypothesis' value be negative or positive. Similarly, a linear equation is used in logistic regression. It employs a sigmoid function or a logistic function, as indicated in equation 2, to compress the predicted real values between 0 and 1, as is appropriate for predicting the chance of belonging to each class. The sigmoid function is depicted in Figure, and its formula is:

$$sigm(z) = 1\,1 + e_z \tag{3}$$

Equation 3 represents logistic regression for a classification problem with a single independent variable (x) and a single dependent (y) variable. Logistic regression's default threshold is 0.5, therefore results with probabilities below 0.5 are assigned to class 0 and those above 0.5 are assigned to class 1. This cutoff level is flexible and can be modified as needed. Where a0 and a1 are the parameters of the logistic regression model that are acquired during training,

$$P\,(y = 1) = sigm\,(a_o + a_{1x}) \tag{4}$$

Consequently, the following represents the expected result when the threshold is set to 0.5.

$$y = 1\,if\;P(y = 1) \geq 0.5$$
$$y = 0\,if\;P(y = 1) < 0.5$$

3.3.2. Decision tree

Because of its clear presentation and interpretability, the decision tree model is widely employed in data mining and machine learning. To learn how to accurately predict a dependent variable's value using a set of independent variables (called features), a tree is constructed or trained. Each leaf node in the tree represents a different independent variable, and the edges leading to their parent nodes indicate the range of values for that variable. [13] The dependent variables are represented by leaf nodes, and the progression from root to leaf represents the accuracy of predictions. To learn a tree, a dataset must be segmented into smaller datasets according to predetermined criteria. Recursive partitioning involves doing this operation recursively on each subset. Various decision-tree algorithms, such as ID3, C4.5, CART, MARS, and so on, have been proposed for tree learning. [14] Decision trees are easily interpretable since they are a white box model. It works effectively with huge datasets despite requiring minimal preprocessing. [15] However, compared to other machine learning methods, trees are often inaccurate, and a single tree is not extremely resilient or stable. To begin examining our data, we construct a decision tree. Based on the values of many parameters including duration, employment, purpose, credit, etc., our tree model classified the samples as good or poor.

Overfitting, however, is a problem that inevitably arises when a tree has too many branches and nodes. Overfitting can be avoided by limiting the total number of branches to a manageable quantity (called pruning). To strike a balance between predictive power and overfitting during the pruning process, the complexity parameter was implemented. Decide on the best possible tree size by adjusting the complexity parameter (cp). The cp value that corresponds to the moment where the error is minimal is the one, we should use to ground our model.

Data

The German Credit Data is available to the public at from the UCI Repository of Machine Learning Databases. This dataset sorts of people into good or bad credit risks based on a set of characteristics about them. There are 700 samples that have a good credit risk, and 300 that have a bad credit risk. As we said in the beginning, there are more people with bad credit than with good credit. This is an example of how credit data is not balanced. The German Credit Data measures 20 different things, 7 of which are numbers and 13 of which are groups of things. These are the risk factors that are used to tell the difference between groups with high risk and groups with low risk. Author split the data into two groups: one for training and one for testing. To make sure the model could be used in the real world, we randomly chose 700 out of 1000 samples and put them into the training set. The 300 samples that were still there were added to the set to be tested.

4. Data Analysis

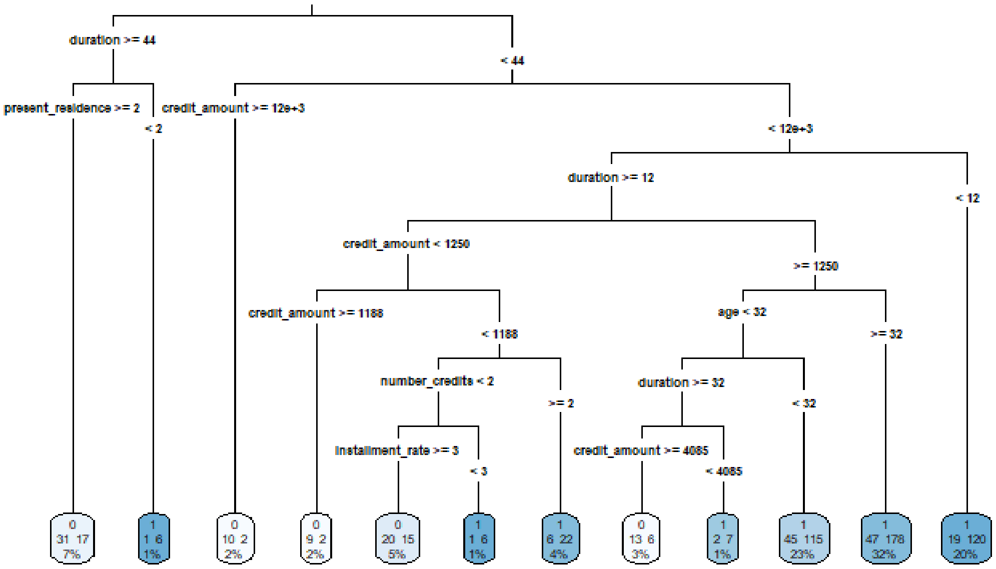

Figure 1: Decision tree for German Data set

Figure 1 shows the decision tree for German Credit data set is created by decision tree algorithm, the present residence ≥ 2 and ≤ 2 shows classification as 0 and 1respectively. if first leaf node is considered than 31 values are classified to correct, and 17 values are considered as incorrect with about 7% of data belong to the present residence classification nodes. Other nodes are showing in the decision tree is calculated in a same manner.

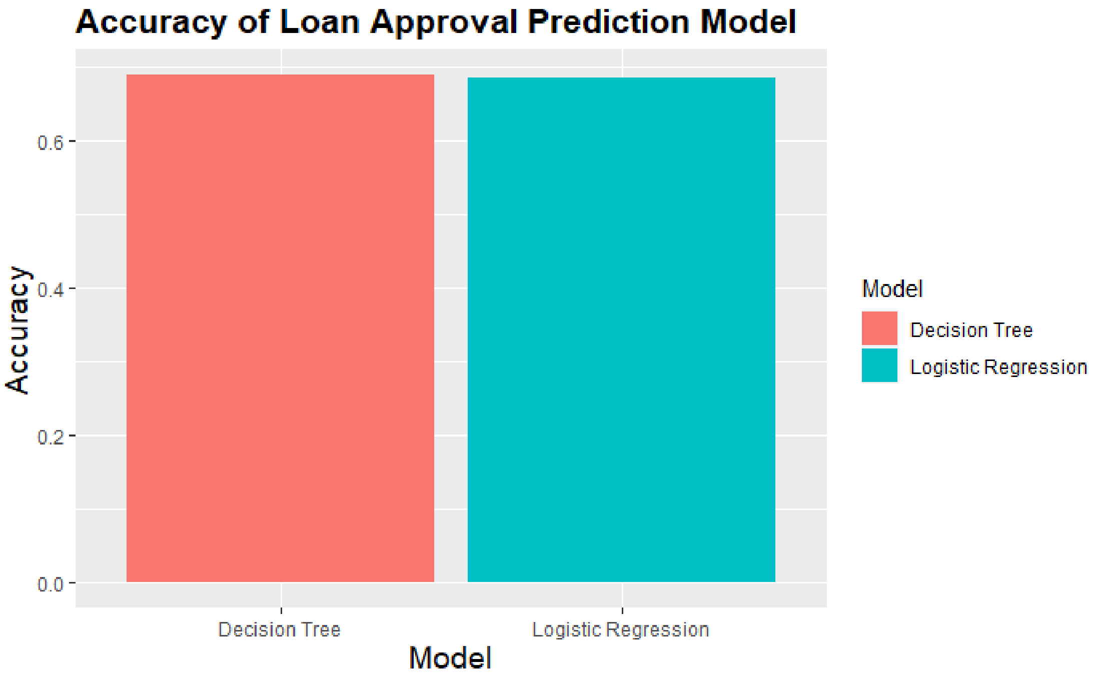

Figure 1 : Accuracy of Loan Approval Prediction Model

Figure 2 shows the Accuracy of loan approval of prediction model, the Decision tree shows the higher accuracy of loan approval prediction model as compared to the logistic regression model.

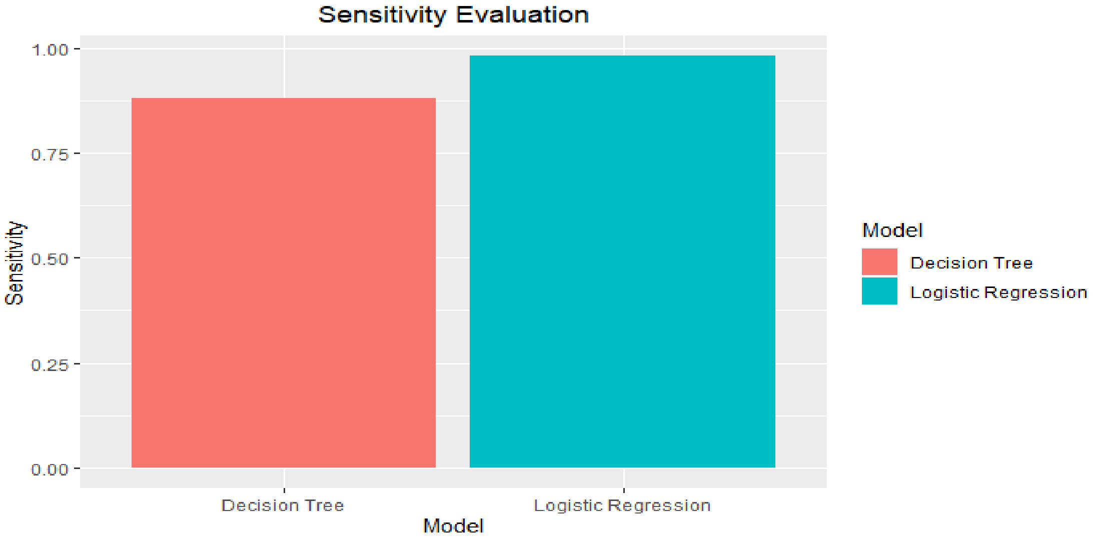

Figure 2 : Sensitivity Evaluation of loan for Prediction Model

Figure 3 shows the sensitivity evaluation of loan prediction model, logistic regression shows the sensitivity evaluation of loan for prediction model shows higher sensitivity evaluation as compared to the decision tree.

Table: Accuracy for Machine Learning Algorithms

Model	Accuracy	Sensitivity	Precision
Logigistic Regression	0.686667	0.980392	0.689655
Decision Tree	0.69	0.882353	0.722892

The above table shows the accuracy of machine learning algorithms by logistic Regression and Decision Tree.

5. Results

This research develops two different classification algorithms: the Logistic Regression algorithm, , and the Decision Tree algorithm. To conduct an analysis of the models, The dataset is divided in a 70:30 ratio, with 70% of the transactions being considered for training the models and 30% of the remaining transactions being used for evaluating the models' accuracy. The random sampling strategy is utilised in order to transform an unbalanced dataset into a balanced dataset for the purpose of achieving more precise outcomes. After that, the classifier algorithms are run on the dataset after it has been adjusted. Metrics like precision, sensitivity, and specificity are used to examine the efficacy of these machine learning techniques.

$$\text{Accuracy} =(TP + TN)/((TP + FP + TN + FN))$$

$$\text{Sensitivity}=TP/((TP + FN))$$

$$\text{Specificity}=TN/((FP + TN))$$

From the table 1 accuracy of the loan prediction model by decision tree is higher than logistic regression, sensitivity analysis for loan prediction model is higher by logistic regression, and the precision analysis for loan prediction is higher by decision tree as compared to logistic regression.

6. Conclusion

Several popular machine learning algorithms logistic regression and decision tree are compared and contrasted in this article. After training two classifiers using distinct machine learning approaches, the resulting outputs are compared using the aforementioned accuracy measures. In terms of Accuracy, the decision tree is considered as the best model, for sensitivity analysis of loan prediction logistic regression is considered to be used. Credit card fraud monitoring has risen to prominence in this age of online purchasing. You can't run a business and deny the reality that we're moving away from cash as a payment method. This makes it impossible to continue using the same means of payment as before. If you don't put it to use, your business won't grow. Not everyone who walks through your door intends to make a purchase will have currency on hand. In recent times, they have prioritized the use of plastic over cash. Therefore, your business will need to modify its payment systems to support a wide variety of currency types. Considering the current situation, it is obvious that businesses use a broad range of payment terminals, including those that accept debit cards, credit cards, and a plethora of other payment methods. Despite this, credit card theft has become a serious problem in recent years. Therefore, in this day and age of digital payment, it is crucial to understand the significance of credit card fraud detection if you want to successfully run your company.

References

[1] Alzaidi, A. a. (2008). Artificial Intelligence for Islamic Banking. The Journal of Muamalat and Islamic Finance Research.

[2] Bahnsen, A. C. (2013). Cost sensitive credit card fraud detection using Bayes minimum risk. In 2013 12th international conference on machine learning and applications (Vol. 1, pp. 333--338). IEEE.

[3] Bhattacharyya, S. a. (2014). Data mining for credit card fraud: A comparative study. Decision support systems, 4(9).

[4] Chaudhary, K. a. (2012). A review of fraud detection techniques: Credit card}. International Journal of Computer Applications, 45(1), 39--44.

[5] Dal Pozzolo, A. a. (2017). Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE transactions on neural networks and learning systems, 29(8), 3784--3797.

[6] de Souza, M. J. (2019). Can artificial intelligence enhance the Bitcoin bonanza. The Journal of Finance and Data Science, 5(2), 83--98.

[7] Dighe, D. a. (2018). Detection of credit card fraud transactions using machine learning algorithms and neural networks: A comparative study. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 1--6.

[8] Gupta, M. a. (2013). Outlier detection for temporal data: A survey. IEEE Transactions on Knowledge and data Engineering, 26(9), 2250--2267.

[9] Haider, A. a. (2022). Predictive Market Making via Machine Learning. Operations Research Forum, 3(1), 5.

[10] Hoda, M. (2015). Computing for Sustainable Global Development (INDIACom).

[11] Jansen, S. (2018). Hands-On Machine Learning for Algorithmic Trading: Design and implement investment strategies based on smart algorithms that learn from data using Python.

[12] Motwani, M. a.-M. (2017). Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. European heart journal, 38(7), 500--507.

[13] Pumsirirat, A. a. (2018). Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. International Journal of advanced computer science and applications, 9(1).

[14] Sircar, A. a. (2021). Application of machine learning and artificial intelligence in oil and gas industry. Petroleum Research, 6(4), 379--391.

[15] Talekar, D. L. (2014). Credit Card Fraud Detection System: A Survey. Intern.