Loan Approval Prediction Using Machine Learning Algorithm- Decision tree

N.Guna sree, M.Divya, N.Reshma, Mr.K.Rajendra Prasad, Department of Electronics and Communication Engineering, Vignan's Institute of Engineering for Women

ABSTRACT

The project is to predict whether assigning the loan to a particular person will be safe or not.With the enhancement in the banking sector lots of people apply for the bank loans, but banks have limited assets which it can grant to limited people only. So, finding out to whom the loan can be granted which will be a safer option to the bank is a difficult process. So, we try to reduce the risk factor while selecting the safe person to save their efforts and assets.

Machine learning focuses on applications that learn from experience and improve their decision-making or predictive accuracy over time. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.

In the lending industry,investors provide loans to borrowers in exchange for the promise of repayment with interest.If the borrower repays the loan,then the lender would make profit from the interest.However,if the borrower fails to repay the loan,then the lender losses money.Therefore,lenders face the problem of predicting the risk of a borrower being unable to repay a loan.In this study,the data from lending club is used to train several Machine Learning models to determine if the borrower has the ability to repay its loan.In addition,we would analyze the performance of the models.

Keywords: Machine learning, information Extraction, Decision tree, logistic Regression

Software used:jupyter notebook,python language

INTRODUCTION

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people. Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problemsolve.

Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs. Any technology user today has benefittedfrommachinelearning.Facialrecognitiontechnologyallowssocialmediaplatforms to help users tag and share photos of friends. Optical character recognition (OCR) technology converts images of text into movabletype.

Recommendation engines, powered by machine learning, suggest what movies or televisionshowstowatchnextbasedonuserpreferences.Self-drivingcarsthatrelyonmachine learning to navigate may soon be available to consumers. Machine learning is a continuously developing field. Because of this, there are some considerations to keep in mind as you work with machine learning methodologies, or analyze the impact of machine learning processes. In this tutorial, we'll look into the common machine learning methods of supervised and unsupervisedlearning,andcommonalgorithmicapproachesinmachinelearning,includingthe k-nearest neighbor algorithm, decision tree learning, and deeplearning.

2. LITERATURE REVIEW

[1] Sarwesh Site, Dr. Sadhna K. Mishra proposed a method in which two or more classifiers are combined together to produce a ensemble model for the better prediction. They used the bagging and boosting techniques and then used random forest technique. The process of classifiers is to improve the performance of the data and it gives better efficiency. In this work, the authors describe various ensemble techniques for binary classification and also for multi class classification. The newtechnique that is described by the authors for ensemble is COB which gives effective performance of classification but it also compromised with noise and outlier data of classification. Finally they concluded that the ensemble based algorithm improves the results for training data set.

[2]Amira Kamil Ibrahim Hassan, Ajith Abraham constructed a loan default predication model using three several neural network training algorithms. The aim is to test accuracy using attribute filter technique and develop a model called ensemble model by combining the results of those three algorithms. The experiment did on several parameters like training time, MSE, R, iteration for comparison. The best algorithm was Levenberg -Marquardt (LM) because it had largest R and the slowest algorithm is One Step Secant (OSS). For the accuracy purpose, the filtering function was applied on original dataset that produced two another dataset. Then for each data set different training algorithm of neural network is applied and the filtering function gave the better model among all the models.

[3] A.R. Ghatge, P.P. Halkarnikar develops the artificial neural network model for predict the credit risk of a bank. The Feed- forward back propagation neural network is used to forecast the creditdefault. They also compare the results with the manual calculations of the bank conducted in year 2004, 2005 and 2006. The results give the better and higher performance over manual calculations of bank.

[4] Suresh Ramakrishnan, Maryam Mirzaei and Mahmoud Bekri explores Adaboost ensemble method and makes an empirical comparison. The main goal is to compare ensemble classifiers. This study explores Ada Boost and bagging ensemble for default prediction to contrast with several classifiers including learning Logistic Regression (LR), Decision Tree (DT), artificial Neural Networks (NN) and support vector machine (SVM) as base learner.

[5] Dr. A. Chitra and S. Uma introduces a two-level ensemble model for prediction of timeseries based on racial bias function network (RBF), k nearest neighbor (KNN) and self-organizing map (SOP). The aim is to increasing the prediction accuracy. They construct a model named PAPEM i.e. Pattern prediction Ensemble Model that uses Mackey dataset, Sunspots dataset and Stock Price dataset as dataset and shows the proposed model performs better than the individuals. The Comparison of various classifiers done on root mean square, mean absolute percentage error and prediction accuracy. The results show that the PAPEM model is better than standalone classifier.

3.Methodology

3.1Machine Learning

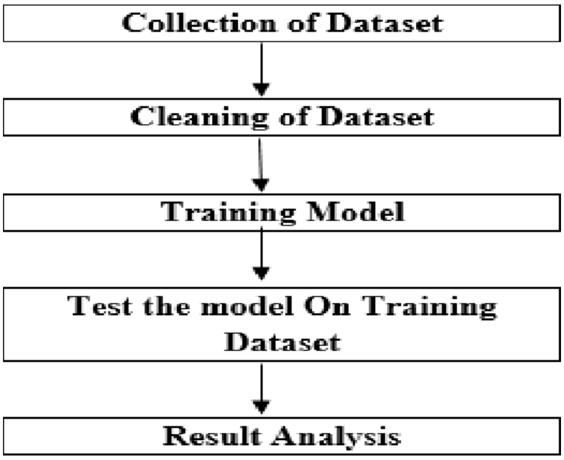
Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. Machine learning enables analysis of massive quantities of data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly. Combining machine learning with AI and cognitive technologies can make it even more effective in processing large volumes of information..

3.2Supervised learning

Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized

Supervised learning is the machine learning task of learning a function that maps an input toanoutputbasedonexampleinput-outputpairs.Itinfersafunctionfromlabeledtrainingdata consisting of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisorysignal). Asupervisedlearningalgorithmanalyzesthetrainingdataandproducesan inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseeninstance.

Loan prediction Methodology



Loan Prediction System allows jumping to specific application so that it can be check on priority basis. This Paper is exclusively for the managing authority of Bank/finance company, whole process of prediction is done privately no stakeholders would be able to alter the processing.

3.2Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

Logistic Function (Sigmoid Function):

The sigmoid function is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range of 0 and 1.

The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.

In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Assumptions for Logistic Regression:

The dependent variable must be categorical in nature.

The independent variable should not have multi-collinearity.

Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

We know the equation of the straight line can be written as:

In Logistic Regression y can be between 0 and 1 only, so for this divided the above equation by (1-y):

$\frac{y}{1-y}$; 0 for y=0, and infinity for y=1

But we need range between –[infinity] to +[infinity], then take logarithm of the equation it will be become:

$\text{Log}[\frac{y}{1-y}]=b_{0+}b_1x_1+b_2x_2+b_3x_3+\ldots\ldots\ldots\ldots..+b_nx_n$

The above equation is the final equation for logistic regression.

Type of Logistic Regression:

On the basis of the categories, Logistic Regression can be classified into three types:

Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.

Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"

Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

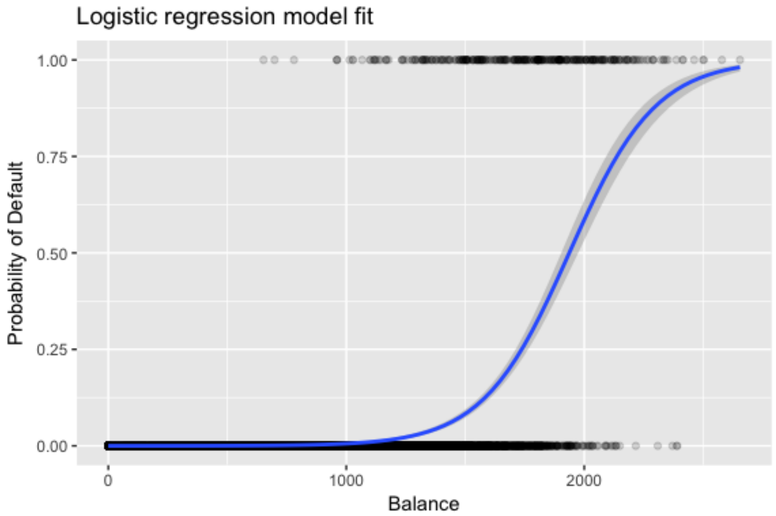
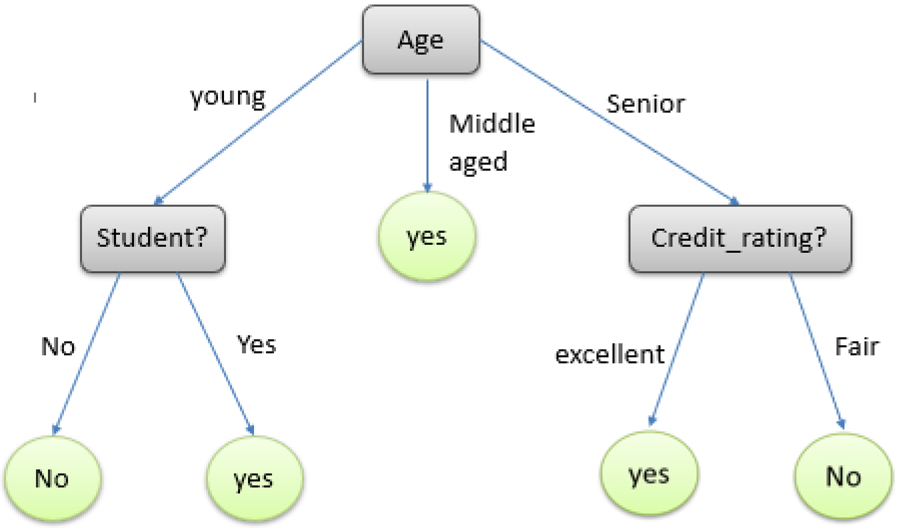


Fig: Logistic Regression graph

3.3 DECISION TREE

Decision tree learning or induction of decision trees is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Decision trees are among the most popular machine learning algorithms given their intelligibility and simplicity.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data (but the resulting classification tree can be an input for decision making). This page deals with decision trees in data mining



3.4How does the Algorithm Work

Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problems too.

The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data(training data).

The understanding level of Decision Trees algorithm is so easy compared with other classification algorithms. The decision tree algorithm tries to solve the problem, by using tree representation. Each internal node of the tree corresponds to an attribute, and each leaf node correspondsto a class label.

4.RESULTS

Algorithm	Accuracy
Logistic regression	77.77
Decision tree	79.16

CONCLUSION

The analytical process started from data cleaning and processing, Missing value imputation with micepackage, then exploratory analysis and finally model building and evaluation. This brings some of the following insights about approval. Applicants with Credit history not passing fails to get approved, Probably because that they have a probability of a not paying back. Most of the Time, Applicants with high income sanctioning low amount is to more likely get approved which make sense, more likely to pay back their loans. Some basic characteristic gender and marital status seems not to be taken into consideration by the company.

REFERENCES

[1]. Sharma, K.K. Sharma, “ Baseline Wander Removal of ECG Signals using Hilbert Vibration Decomposition”, IEEE Electronics Letters, vol. 51 no.6 pp. 447-449, March 2015.

[2] P. Kolios, C. Panayiotou, G. Ellinas, and M. Polycarpou, “Data-driven event triggering for IoT applications,” IEEE Internet Things J., vol. 3, no. 6, pp. 1146–1158, Dec. 2016.

[3] Kumar Arun, Garg Ishan, Kaur Sanmeet, MayJun. 2016. Loan Approval Prediction based on Machine Learning Approach, IOSR Journal of Computer Engineering (IOSR-JCE)

[4] Research on bank credit default prediction based on data mining algorithm, The International Journal of Social Sciences and Humanities Invention 5(06): 4820-4823, 2018.

[5]Research on bank credit default prediction based on data mining algorithm, The International Journalof Social Sciences and Humanities Invention 2018.