RESEARCH ARTICLE OPEN ACCESS

Accuracy Prediction for Loan Risk Using Machine Learning Models

Anchal Goyal [1], Ranpreet Kaur [2]
Research Scolar [1], Assistant Proffesor [2]
Department of Computer Science
RIMT –IET (PTU),Mandi Gobindgarh
Punjab - India

ABSTRACT

Extending credit to individuals is essential for markets and society to act efficiently. Estimating the probability that an individual would default on their loan, is useful for banks to make a decision whether to approve a loan to the individual or not. In this paper, we find the accuracy of several models in R language and evaluate it to establish the finest model to forecast the finance status for an organization. We did the experiment five times on the same data set and find the experimental results that show the Tree Model for Genetic Algorithm is the best model for forecasting the finance for costumers.

Keywords: - Accuracy, Prediction, Genetic algorithm, Finance.

I. INTRODUCTION

A. Introduction to Machine Learning

Machine learning is a arena of computer science that involves the learning of pattern identification and computational learning theory in AI. Machine learning generally refers to the changes in systems that carry out tasks linked with artificial intelligence (AI). Such tasks include recognition, analysis, planning, robot control, forecasting, etc. It explores the study and construction of algorithm that can make prediction on data. Machine Learning is used to build programs with its tuning parameters that are adapted consequentially so as to increase their functioning by adapting to earlier data.

Machine learning can be broken into two categories:

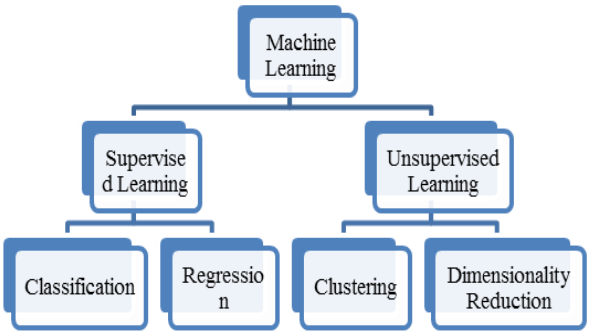

Fig.1 Categories of Machine Learning

1) **In Supervised Learning**, a data set includes both *features* and *labels*. The task is to build an estimator which is able to forecast the label of an object with the set of features. Supervised learning is further broken down into two parts: classification and regression.

Classification is the task of forecasting the value of a categorical variable given some input variables.

Regression is the task of forecasting the value of a continuously changeable variable (e.g. a price, a temperature) given some input variables.

2) **In Unsupervised Learning,** a data set has no label and we find similarities among the objects. We can use this technique to display the best arrangement of data. It includes tasks such as dimensionality reduction, clustering.

Dimensionality reduction is the task of derive a set of new features that is smaller than the original feature set while hold large of the changing of the original data.

Clustering is the method of gathering samples into groups of analogous samples according to some predefined similar or dissimilar measure.

B. Introduction to R Language

R is a programming language produced by Ross Ihaka and Robert Gentleman at the University of Auckland. R is a GNU project and the source code of R is written in

C, FORTRAN. R is basically a data analysis software environment for statistical computing i.e. interface between the statistics and computer science. R can be extended easily by packages that are available in CRAN package repository. R and its libraries implement a large variety of statistical and graphical techniques including time series analysis, clustering, classification, linear and non linear modeling. Basically R is freely available under GNU General Public License and precompiled binary versions are provided for various operating system. This language is mostly used by data miners and statisticians for developing the software.

1) The R environment

R is integrated software for manipulation of data, doing calculation and for graphical display. It includes

- an effective storage resource
- an effectual data handling resource
- a well developed programming language that include loops, functions and other input output facilities.
- Simple and effective language for users
- a huge and integrated collections of tools for analysis of data

C. Accuracy

Accuracy depends on how data is collected, and judged on basis of comparison of several parameters. True positive (TP) depicts amount of predictions which are positive, the actual value being positive. Similar in the case of true negative (TN). The accuracy is computed as:

$$Accuracy = \frac{TP + TN}{Toatal\ Data} * 100$$

When we build a predictive model, we need a way to evaluate the capability of the model on the data. This is done by estimating the accuracy of the model. The Caret Package in R provides a large no. of methods for estimating the accuracy of machine learning algorithms.

II. DATA SET AND FEATURES

The data set include 13 attributes such as Gender, Marital Status, Education, Income, Loan Amount, Credit History and others which are shown in table1 & the sample date set is shown in table2:

Table1: Feature Description

Feature ID	Features	Information
F1	Loan ID	Unique Loan ID
F2	Gender	Male/ Female
F3	Married	Applicant married (Y/N)
F4	Dependents	Number of dependents
F5	Education	Applicant Education (Graduate/ Under Graduate)
F6	Self-employed	Self employed (Y/N)
F7	Applicant Income	Applicant income
F8	Co applicant Income	Co applicant income
F9	Loan Amount	Loan amount in thousands
F10	Loan Amount Term	Term of loan in months
F11	Credit History	credit history meets guidelines
F12	Property Area	Urban/ Semi Urban/ Rural
F13	Loan Status	Loan approved (Y/N)

III. MACHINE LEARNING MODELS

Various machine learning models that have been applied for the prediction of accuracy as explained below:

1. Decision Tree Model

A decision tree model is one of the most frequent data mining models. It is popular because it is easy to understand. Decision trees are one of the useful algorithms that are used for regression and classification. They are also known as glass-box model. When the model once found the template in the data then we can see what the decision will be made for that data which we want to predict.

2. Linear model

A linear model is the one of the method for fitting a statistical model to data. It is appropriate when the target variable is numeric and persistant. This model helps to analyze the data and also helps to recognize and predict the performance of the complicated system.

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	0
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	1
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	1
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	1
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	1

Table 2: Sample dataset

Table 3: Machine learning models used.

Models	Packages	Tuning Parameter(s)
Bagged CART	Iperd	None
Random Forest	Randomforest	Mtry
Tree Model For Genetic Algorithm	evtree	Alpha
Decision Trees	rpart	Min split, max depth, min bucket
Linear Model	Car, nnet	Size
Neural Network	nnet	Size, decay
SVM	Kernlab	C
Extreme Learning Machine	elmNN	Nhid,actfun
Multivariate Adaptive Regression spline	earth	Degree
Bayesian Generalized Linear Model	arm	None
Model Tree	tree	None

3. Random Forest Model

A random forest model is basically a collection(i.e. ensemble) of tens or hundrerds of decision trees. These models are mainly used if we have large no. of input variables i.e. in hundreds and thousands and if we have very vast dataset. This model is very efficient if we have large no. of variables and it distributes the variable into different subsets. Ensemble models are robust to variance and bias.

4. Neural Network Model

This model is basically based on various layers that are connected to each other like neurons. This model combines the numbers and provides the numeric data to produce the final results throughout the network. These models are identical to biological neural network in order to perform functions parallel and collectively rather than individually.

5. Support vector machine

SVM is supervised machine learning model with learning algorithms which examine the data and uses that data for regression and classification. This model uses a technique namely a kernel trick to transform the data and based on these transforms of data, it finds the best optimum results. It is not considered as better as than the other machine learning models because it works on less data set.

6. Extreme learning machines

ELM is a modification is a feed forward network with single layer which have a hidden nodes for single layer. The Weights are randomly given to hidden nodes and it never be updated. The name to this model was given by Guang-Bin Huang. Different from other traditional models, the extreme learning model not only provide the smaller training error but also better performance.

7. Multivariate Adaptive Regression Splines

This model is established by Jerome H. Friedman in 1991. This model is used for both regression and classification type problem with the purpose to predict the values. The 'earth' package is used in implementation of this model. The earth source code is licensed under the GPL. This technique has popular in data mining because it is used to find the difficult data mining problems.

8. Model tree

Model tree is a classification model that is combination of decision tree learning and logistic regression model. The package named 'tree' is used in implementation of this model. This model tree works on when have to predict the numeric quantities. It is a tree that include linear regression function at their leaves.

9. Bayesian Generalized Linear Model

BGLM is most generally used technique for creating the relationship. This model is used when have huge dataset and BGLM is used to fit the dataset into

pragmatic size and remove the problem of over fitting. This model is included in package "arm" in r language.

10. Bagged Cart Model

This model is used for classification and regreession problems. This model build under the package 'iperd' and 'plyr'. Bagging for classification and regression trees were suggested by Breiman in 1996.

11. Tree model form Genetic Algorithm

Genetic algorithm is a search heuristic i.e. it is an algorithm for finding and solving a problem more quickly and produces the result in reasonable time. This model is very efficient, flexible and finds optimal solutions for given problem. This model builds under the package 'evtree'. This algorithm is usually based on theory of natural selection and survival of fittest. The larger the value of fitness is the most the optimal result will be.

IV. RESULTS

In it, we analyze the results of various different machine learning models which are implemented in R to find the accuracy of each model and find the best model for the bank that provides loan to the costumers. Accuracy is the most important aspect for any organization. The experimental results show that the Tree Model from Genetic Algorithm is the best model among the entire model for the given dataset for predicts the loan .Table 4 shows the accuracy of all the models.

Table 4: Accuracy Results

Models / No. of runs	**RUN1**	**RUN2**	**RUN3**	**RUN4**	**RUN5**
Decision Tree	75	80.56	76.39	81	79.86
Linear Model	73.61	75	70.14	79.17	78.47
Neural Network	75.69	81.25	77.02	83	79.86
Random Forest	76.39	79.86	77.08	82.64	80.56
SVM	77	79.56	76.39	81.94	80.56
Bagged Cart	77.08	74.47	76.39	79.17	78.47
Tree model for genetic algo	79.5	81.75	78.5	83.33	81.25
model tree	73.61	74.31	69.44	70.83	79.86
Extreme learning machine	57.64	69.44	65.97	73.61	68.75
Multivariate Adaptive Regression Spline	75.69	79.17	77.5	81.23	79.86
BGLM	76.39	81.25	76.39	83.33	79.86

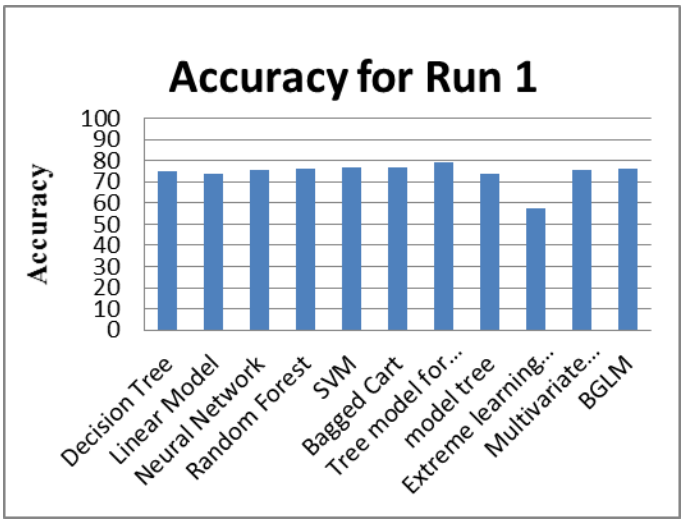

Fig. 2 Accuracy for Run 1

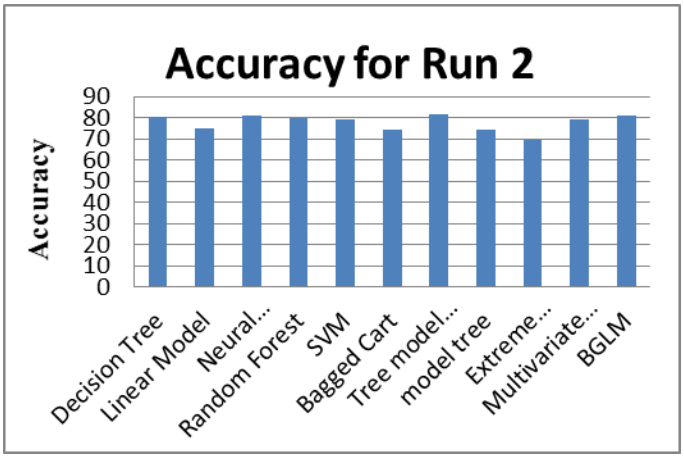

Fig. 3 Accuracy for Run 2

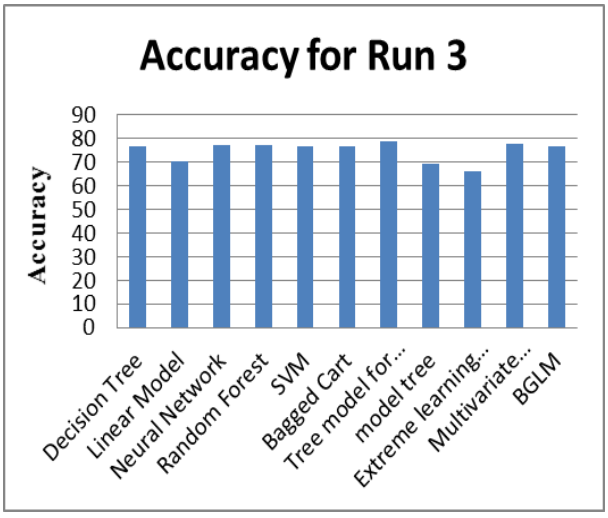

Fig. 4 Accuracy for Run 3

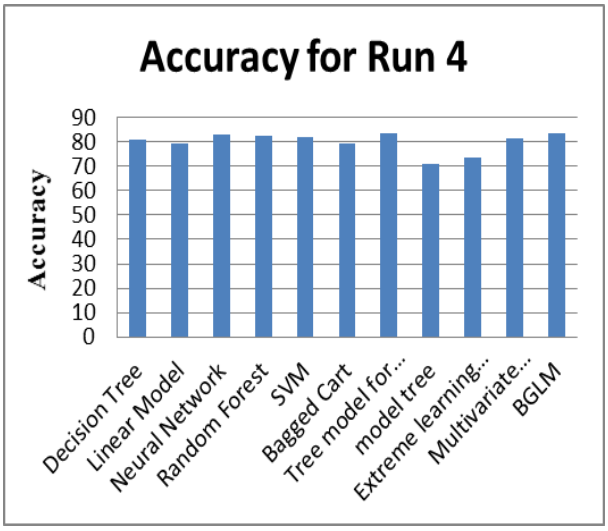

Fig. 5 Accuracy for Run 4

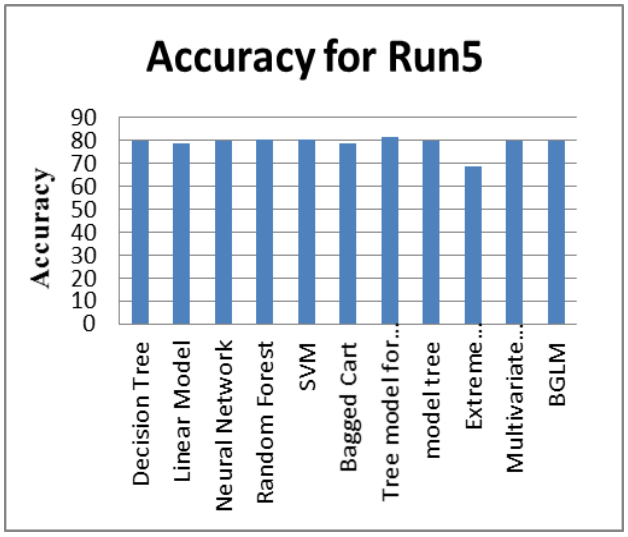

Fig. 6 Accuracy for Run 5

V. CONCLUSION

In this paper, we find the accuracy of several models in R language and evaluate it to establish the best model to predict the finance status for an organization. We did this experiment five times on the same data set having different seed values and the accuracy varies according to its seed value that is shown in figures 2 to 6. The experimental results that show the Tree Model for Genetic Algorithm is the best model for forecasting the loan for costumer.

REFRENCES

[1] Wo- Chiang Lee, " Genetic Programming Decision Tree for Bankruptcy Prediction", JCIS-Oct 2006, 1951-6851.

[2] Breiman, L, "Bagging predictors". Machine Learning. .

[3] Breiman, L, "Random forests". Machine Learning, 2001, 5–32.

[4] Lim, T.-S., Loh, W.-Y., & Shih, Y.-S. ." A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms,0ct 2000, 203–228.

[5] Niculescu-Mizil, A., & Caruana, R." Predicting good probabilities with supervised learning". Aug2005.

[6] Jason Brownleek. "Compare Machine Learning Models", Sep24, 2014.

[7] K. Bache and M. Lichman, UCI machine learning repository.

[8] H. Faris, B. Al-Shboul and N. Ghatasheh, "A Genetic Programming Based Framework for Churn Prediction in Telecommunication Industry",Sep 2014, 253-362.

[9] D. Fantazzini and S. Figini, "Random Survival Forests Models For SME Credit Risk Measurement," Methodology and Computing in Applied Probability, Jan 2009, 29-45.