Prediction of Loan Approval using Machine Learning Algorithm: A Review Paper

Ms. Kathe Rutika Pramod
Information Technology Engineering, SVIT, Nashik Maharashtra, India

Ms. Panhale Sakshi Dattatray
Information Technology Engineering, SVIT, Nashik Maharashtra, India

Ms. Avhad Pooja Prakash
Information Technology Engineering, SVIT, Nashik Maharashtra, India

Ms. Dapse Punam Laxman
Information Technology Engineering, SVIT, Nashik Maharashtra, India

Mr. Ghorpade Dinesh B.
Information Technology (Assistant Proffessor) SVIT, Nashik Maharashtra, India

---

***Abstract*—** In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So they can earn from interest of those loans which they credits. A bank's profit or a loss depends to a large extent on loans i.e. whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non-Performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison. A very important approach in predictive analytics is used to study the problem of predicting loan defaulters: The Logistic regression model. The data is collected from the Kaggle for studying and prediction. Logistic Regression models have been performed and the different measures of performances are computed. The models are compared on the basis of the performance measures such as sensitivity and specificity. The final results have shown that the model produce different results.

I. INTRODUCTION

Small loan is an important aspect of our everyday life: it allows aspiring entrepreneurs to get started on ideas that could be grown into business; it allows curious students to afford higher education that is otherwise unavailable without a stable income; more importantly, it allows ordinary people who have no friends or relatives for support to obtaining short-term financial assistance and get back on their feet to fight for the American Dream. Nevertheless, with loan it comes with the possibility of default as well. Default is a financial term describing the failure of meeting the legal obligation of a loan - paying back the principal and interest. It's a common problem in the financial industries and one of the major risks of offering loans. Of course, default does not happen the majority of the time and the lending banks usually able to make up the loss from a defaulting loan from other fully paid loans and their accompanied interests. Furthermore, banks issuing loans with higher interest rate to individuals with high probability of default - the financial institutions are trading off an increased chance of default with an increased profit from the high interest. All things considered, default is a fact of life and most financial institutions have a well established practice to minimize its impact and absorbing the loss. But what about a situation where instead of a single bank is issuing the loan, the loan is comprised of funds from several investors? Lending Club is one of the many peer-to-peer lending company that gives rise to this peculiar situation. In plain words, peer-to-peer lending company acts as a broker between borrowers and investors. The company creates a platform where borrowers can create small unsecured personal loans, and investors can seek out these loans and decide which loans to invest from. Borrowers obtain the loan they want, investors get to profit from the loan interest, and the company gets a cut from both parties (origination fee from borrowers and service fee from investors). This also means that when a loan goes default, it's no longer a single bank that is absorbing the loss - single or multiple individual investors will be absorbing it instead. The overall profit might be positive if all the loans were originated from a single lender as other fully paid loans could cover the loss, but this is no longer the case as there will be winners and losers among this new form of lending practices if the investors did not diversify. An obvious solution to this problem is to predict whether a particular loan will go default based on initial information provided by the borrowers and their credit report. There's no doubt Lending Club already has an existing model in place to approve loans posted on their website. This paper will explore the process and result on formulating a new machine learning model that could predict a loan default; but more importantly, the model will focus on minimizing the overall loss in investment of bad loans in order to lessen the burden passed onto individual investors. As a side note, the paper will also explore privacy-preserving mechanism on sensitive information provided from the borrow's credit report. The end goal is to evaluate a simplified version of RAPPOR (Randomized Aggregately Privacy Preserving Ordinal Response) and determine whether data that have

been hashed by this algorithm could still be use to predict loan default as stated previously.

II. LITERATURE SURVEY

Amira Kamil Ibrahim Hassan, Ajith Abraham (2008) uses a prediction model which is constructed using three different training algorithms to train a supervised two layer feed-forward network. The results show that the training algorithm improves the design of loan default prediction model.

Angelini (2008) used a neural network with standard topology and a feed-forward neural network with ad hoc connections. Neural network can be used for prediction model. This paper shows that the above two models give optimum results with less error.

Ngai (2009) uses the classification model for predicting the future behaviour of costumers in CRM. In CRM domain, the mostly used model is neural network. He recognized eighty seven articles associated to data mining applications and techniques between 2000 and 2006.

Dr. A. Chitra and S. Uma (2010) introduced a ensemble learning method for prediction of time series based on Radial Basis Function networks (RBF), K - Nearest Neighbor (KNN) and Self Organizing Map (SOM). They proposed a model namely PAPEM which perform better than individual model.

Akkoç (2012) used a model namely hybrid Adaptive Neuro-Fuzzy Inference model, grouping of statistics and Neuro-Fuzzy network. A 10-fold cross validation is used for better results and a comparison with other models.

Sarwesh Site, Dr. Sadhna K. Mishra (2013) proposed a method in which two or more classifiers are combined together to produce an ensemble model for the better prediction. They used the bagging and boosting techniques and then used random forest technique.

Maher Alaraj, Maysam Abbod, and Ziad Hunaiti (2014) proposed a new ensemble method for classification of costumer loan. This ensemble method is based on neural network. They state that the proposed method give better results and accuracy as compared to single classifier and any other model.

AlarajM , AbbodM (2015) introduced a model that are based on homogenous and heterogeneous classifiers. Ensemble model based on three classifiers that are logistic artificial neural network, logistic regression and support vector machine.

III. MACHINE LEARNING

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. It's a science that's not new – but one that has gained fresh momentum.

While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data – over and over, faster and faster – is a recent development. Resurging interest in machine learning is due to the same factors that have made data mining and Bayesian analysis more popular than ever. Things like growing volumes and varieties of available data, computational processing that is cheaper and more powerful, and affordable data storage.

All of these things mean it's possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results – even on a very large scale. And by building precise models, an organization has a better chance of identifying profitable opportunities – or avoiding unknown risks. It is no doubt that the sub-field of machine learning / artificial intelligence has increasingly gained more popularity in the past couple of years. As Big Data is the hottest trend in the tech industry at the moment, machine learning is incredibly powerful to make predictions or calculated suggestions based on large amounts of data. Some of the most common examples of machine learning are Netflix's algorithms to make movie suggestions based on movies you have watched in the past or Amazon's algorithms that recommend books based on books you have bought before.

IV. MACHINE LEARNING ALGORITHMS

1. Decision Trees: A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance-event outcomes, resource costs, and utility. Take a look at the image to get a sense of how it looks like. From a business decision point of view, a decision tree is the minimum number of yes/no questions that one has to ask, to assess the probability of making a correct decision, most of the time. As a method, it allows you to approach the problem in a structured and systematic way to arrive at a logical conclusion.
2. Naive Bayes Classification: Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. The featured image is the equation — with P(A|B) is posterior probability, P(B|A) is likelihood, P(A) is class prior probability, and P(B) is predictor prior probability.
   Some of real world examples are:

   - To mark an email as spam or not spam

- Classify a news article about technology, politics, or sports
- Check a piece of text expressing positive emotions, or negative emotions?
- Used for face recognition software.

3. Ordinary Least Squares Regression: If you know statistics, you probably have heard of linear regression before. Least squares is a method for performing linear regression. You can think of linear regression as the task of fitting a straight line through a set of points. There are multiple possible strategies to do this, and "ordinary least squares" strategy go like this—You can draw a line, and then for each of the data points, measure the vertical distance between the point and the line, and add these up; the fitted line would be the one where this sum of distances is as small as possible.
4. Logistic Regression: Logistic regression is a powerful statistical way of modeling a binomial outcome with one or more explanatory variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. In general, regressions can be used in real-world applications such as:
    a. Credit Scoring
    b. Measuring the success rates of marketing campaigns
    c. Predicting the revenues of a certain product
    d. Is there going to be an earthquake on a particular day.
5. Support Vector Machines: SVM is binary classification algorithm. Given a set of points of 2 types in N dimensional place, SVM generates a (N—1) dimensional hyperplane to separate those points into 2 groups. Say you have some points of 2 types in a paper which are linearly separable. SVM will find a straight line which separates those points into 2 types and situated as far as possible from all those points. In terms of scale, some of the biggest problems that have been solved using SVMs (with suitably modified implementations) are display advertising, human splice site recognition, image-based gender detection, large-scale image classification.
6. Clustering Algorithms: - Clustering, like regression, describes the class of problem and the class of methods. Clustering methods are typically organized by the modeling approaches such as centroid-based and hierarchal. All methods are concerned with using the inherent structures in the data to best organize the data into groups of maximum commonalities.
    a. The most popular clustering algorithms are:
    b. k-Means
    c. k-Medians
    d. Expectation Maximization (EM)
    e. Hierarchical Clustering
7. Artificial Neural Network Algorithms: Artificial Neural Networks are models that are inspired by the structure and/or function of biological neural networks. They are a class of pattern matching that are commonly used for regression and classification problems but are really an enormous subfield comprised of hundreds of algorithms and variations for all manner of problem types. Note that I have separated out Deep Learning from neural networks because of the massive growth and popularity in the field. Here we are concerned with the more classical methods. The most popular artificial neural network algorithms are:
    a. Perceptron
    b. Multilayer perceptions (MLP)
    c. Back-Propagation
    d. Stochastic Gradient Descent
    e. Hopfield Network
    f. Radial Basis Function Network (RBFN)

V. PROPOSED SYSTEM

Decision tree algorithm in machine learning methods which efficiently performs both classification and regression tasks[2]. It creates decision trees. Decision trees are widely used in the banking industry due to their high accuracy and ability to formulate a statistical model in plain language. In Decision tree each node represents a feature (attribute), each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continues value). Using different data analytics tools loan prediction and there severity can be forecasted. In this process it is required to train the data using different algorithms and then compare user data with trained data to predict the nature of loan. Several R functions and packages were used to prepare the data and to build the classification model. The work proves that the R package is an efficient visualizing tool that applies data mining techniques. Using R Package, customer's data analysis can be done and depends on that bank can sanction or reject the loan. In real time customers data sets may have many missing and imputed data which needs to be replaced with valid data generated by making use of the available completed data. The dataset has many attributes that define the credibility of the customers seeking for several types of loan. The values for these attributes can have outliers that do not fit into the regular range of data.

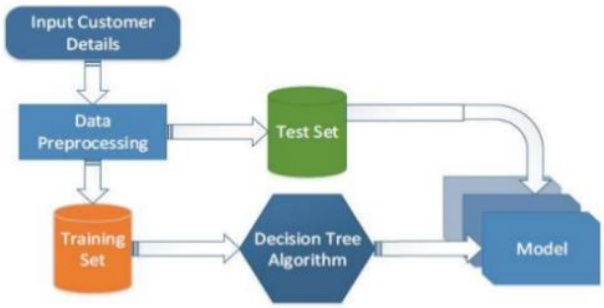

Fig. System Architecture

VI. ALGORITHM USED

DT is a supervised learning algorithm used to solve classification and regression problems too. Here, DT uses tree representation to solve the prediction problem, i.e., external node and leaf node in a tree represents attribute and
class labels respectively. The pseudo code for DT model is depicted in the following section:
Step 1: Best attribute is chosen as the tree's root.
Step 2: Training set is divided into subsets, such that, each subset comprises similar value for an attribute.
Step 3: Step 1 and Step 2 are repeated for all subsets until all
the leaf nodes are traversed in a tree.

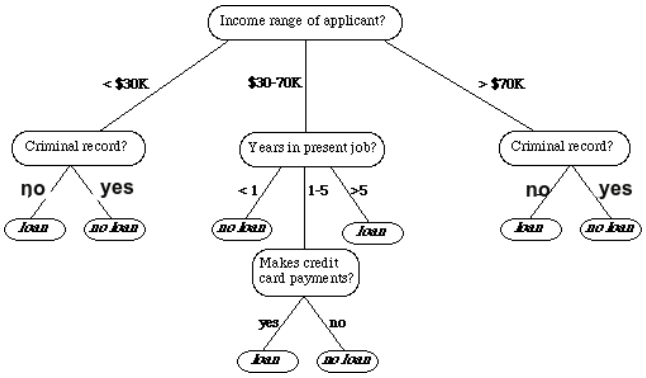

Fig. Decision Tree Algorithm

VII. CONCLUSION

The analytical process started from data cleaning and processing, Missing value imputation with mice package, then exploratory analysis and finally model building and evaluation. The best accuracy on public test set is 0.811. 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.

VIII. REFERENCES

[1] Toby Segaran, "Programming Collective Intelligence: Building Smart Web 2.0 Applications." O'Reilly Media.

[2] Drew Conway and John Myles White," Machine Learning for Hackers: Case Studies and Algorithms to Get you Started," O'Reilly Media.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman,"The Elements of Statistical Lear

[4] PhilHyo Jin Do ,Ho-Jin Choi, "Sentiment analysis of real-life situations using loca- tion, people and time as contextual features," International Conference on Big Data and Smart Computing (BIGCOMP), pp. 39–42. IEEE, 2015.

[5] Bing Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, May 2012.

[6] Bing Liu, "Sentiment Analysis: Mining Opinions, Sentiments, and Emotions," Cambridge University Press, ISBN:978-1-107-01789-4.

[7] Shiyang Liao, Junbo Wang, Ruiyun Yu, Koichi Sato, and Zixue Cheng, "CNN for situations understanding based on sentiment analysis of twitter data," Procedia computer science, 111:376–381, 2017.CrossRef.

[8] K I Rahmani, M.A. Ansari, Amit Kumar Goel, "An Efficient Indexing Algorithm for CBIR,"IEEE- International Conference on Computational Intelligence & Communication Technology ,13-14 Feb 2015.

[9] Gurlove Singh, Amit Kumar Goel ,"Face Detection and Recognition System using Digital Image Processing" , 2nd International conference on Innovative Mechanism for Industry Application ICMIA 2020, 5-7 March 2020, IEEE Publisher. ning: Data Mining, Inference, and Prediction," Springer ,Kindle

[10] Amit Kumar Goel, Kalpana Batra, Poonam Phogat," Manage big data using optical networks", Journal of Statistics and Management Systems "Volume 23, 2020, Issue 2, Taylors & Francis.

[11] Raj, J. S., & Ananthi, J. V., "Recurrent neural networks and nonlinear prediction in support vector machine" Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40, 2019.

[12] Aakanksha Saha, Tamara Denning, VivekSrikumar, Sneha Kumar Kasera. "Secrets inSource Code: Reducing False Positives usingMachine Learning", 2020 InternationalConference on Communication Systems &Networks (COMSNETS), 2020.

[13] X.Frencis Jensy, V.P.Sumathi,Janani Shiva Shri, "An exploratory Data Analysis for Loan Prediction based on nature of clients", International Journal of Recent Technology and Engineering (IJRTE),Volume-7 Issue-4S, November 2018.

[14] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma,Namburi Vimala Kumari, k Vikash,"Loan Prediction by using Machine Learning Models", International Journal of Engineering and Techniques.Volume 5 Issue 2, Mar-Apr 2019