A Study on "Loan Predictions Using Fintech Decision Tree Analysis"

Srihari S, ITM Business School, Student, Chennai
S. Huxley, Visiting faculty, ITM Business School and AVP Indium Software, Chennai
Dr Ajitha Savarimuthu, Associate Professor, Acharya Bangalore Business School, Bangalore

Abstract: *In today's world, banking sector is crucial to the modern economy. As the primary supplier of credit, it provides money for people to buy cars and homes and for businesses to buy equipment, expand their operations, and meet their payrolls. The credit cards, debit cards, and checking accounts that banks make available facilitate all kinds of everyday transactions. They also help drive e-commerce, where cash is of little use. With banking products becoming increasingly commoditized, Analytics can help banks differentiate themselves and gain a competitive edge. Machine learning forecasting for banking enables more accurate reporting by automating credit risk testing for both banks and customers. By evaluating a consumer's financial history, recent transactions, and purchasing patterns, machine learning can make accurate forecasts of future spending and income. Predictive analytics helps banks distinguish between the various portfolio risks effectively, by optimizing the collections process. It helps banks segregate risky customers from the risk-free ones. This can help banks devise actions and strategies to achieve positive results. Predictive Analytics is a stream of advanced analytics which uses new as well as historical data to forecast activity, behaviour, and trends to predict the future. This involves data mining, modelling, employing statistical analysis techniques, and automated machine learning algorithms to make the predictions. It helps organizations discover business issues in real time and address them at the right time to get the best outcomes*

Keywords: *Banking Sector, Modern Economy, Machine learning forecasting, credit risk, Predictive analytics, Positive Results, Data Mining, Statistical Analysis.*

1. Introduction: All loans are treated via way of means of a financial institution. They may be discovered in all varieties of urban, semi-urban, and rural settings. Customer first applies for a residence mortgage, and then the corporation verifies the client's mortgage eligibility. The corporation desires to automate the mortgage eligibility procedure (in actual time) primarily based totally at the statistics supplied via way of means of the client while filling out the net utility shape. Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and different records are included. To automate this procedure, they created a trouble to pick out the purchaser segments which are certified for a mortgage amount, letting them goal those purchasers individually.

The idea of borrowing has existed because the sunrise of humanity, however now it has many aspects. This extends to non-public credit score exchanges for non-public performance-primarily based totally repayments, with credit score as earnings from unsecured every day contributions, besides with inside the banking zone, which calls for collateral for public lending Will be granted. The uniform prevalence of defaults and the resultant defaults is contemplated with inside the diploma of financial institution closures and the diploma of sovereign defaults that the creditors have absolutely experienced. A financial institution mortgage chance evaluation desires to recognize the motive for this chance. In addition, the range of exchanges with inside the economic zone is developing rapidly, and the quantity of statistics approximately client conduct and lending chance is expanding. The

cause of this paper is to pick out the sort or info of the client making use for the mortgage.

The time period mortgage refers to a kind of credit score car wherein a amount of money is lent to every other celebration in change for destiny compensation of the fee or foremost amount. In many cases, the lender additionally provides hobby and/or finance expenses to the foremost fee which the borrower have to pay off further to the foremost balance. Loans can be for a precise, one-time amount, or they'll be to be had as an open-ended line of credit score as much as a distinctive limit. Loans are available many distinctive bureaucracies together with secured, unsecured, commercial, and private loans. A mortgage is a shape of debt incurred via way of means of an person or different entity. The lender—normally a corporation, economic institution, or government—advances a amount of money to the borrower. In return, the borrower has the same opinion to a sure set of phrases together with any finance expenses, hobby, compensation date, and different conditions. In a few cases, the lender might also additionally require collateral to stable the mortgage and make certain compensation. Loans may additionally take the shape of bonds and certificate of deposit (CDs). It is likewise viable to take a mortgage from a 401(k) account.

2. Review of Literature: R. Ghatge, P.P(2006) developed the artificial neural network model to predict the credit risk of a bank. The Feed- forward back propagation neural network issued to forecast the credit default. 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.

Alaraj, M, Abbod, M (2008) Introduced a credit risk 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. The results show that the heterogeneous classifiers ensemble gave improved performance and accurateness as compared to homogeneous classifiers ensemble.

Dr. A. Chitra, S. Uma (2009) introduces a two-level ensemble model for prediction of time series based on radial bias function network (RBF), k nearest neighbour (KNN) and self-organizing map (SOP). The aim is to increase 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.

Board Diversity And Its Effects On Bank Performance(2010) This study analyses the effect of board diversity (gender and nationality) on performance in banks. By making use of a sample of 159 banks in nine countries during the period 2004–2010, our empirical evidence shows that gender diversity increases bank performance, while national diversity inhibits it. Complementarily, according to their institutional characteristics, we also show the moderating effect of investor protection and bank regulatory regime on this previous relationship, analysing their substitution or complementary roles. Our results also suggest that these institutional factors play a significant role in these effects.

Problem Loans and Cost Efficiency in Commercial Banks (2012). This paper addresses a little examined intersection between the problem loan literature and the bank efficiency literature. We employ Granger-causality techniques to test four hypotheses regarding the relationships among loan quality, cost efficiency, and bank capital. The data suggest that problem loans precede reductions in measured cost efficiency; that measured cost efficiency precedes reductions in problem loans; and the reductions in capital at thinly capitalized banks precede increases in problem loans. Hence, cost efficiency may be an important indicator of future problem loans and problem banks. Our results are ambiguous concerning whether or not researchers should control for problem loans in efficiency estimation.

Income Contingent Loans. For Higher Education International Reforms (2014) it is well known that higher education financing involves uncertainty and risk with respect to students' future economic fortunes, and an unwillingness of banks to provide loans because of the absence of collateral. It follows that without government intervention there will be both socially sub-optimal and regressive outcomes with respect to the provision of higher education. The historically most common response to this market failure – a government guarantee to repay student loans to banks in the event of default – is associated with significant problems.

Income contingent loans offer a possible solution. Since the late 1980s ICLs have been adopted in, or recommended for, a significant and growing number of countries, and it is this important international policy reform that has motivated the chapter.

An icl provides students with finance for tuition and/ or income support, its critical and defining characteristic being that the collection of the debt depends on the borrowers' future capacity to pay. ICL have two major insurance advantages for borrowers over more typical arrangements: default protection and consumption smoothing.

A Comparative Study of Machine Learning Approaches for Non-Performing Loan Prediction (2015) Credit risk estimation and the risk evaluation of credit portfolios are crucial to financial institutions which provide loans to businesses and individuals. Non-performing loan (NPL) is a loan type in which the customer has a delinquency; because they have not made the scheduled payments for a time period. NPL prediction has been widely studied in both finance and data science. In addition, most banks and financial institutions are empowering their business models with the advancements of machine learning algorithms and analytical big data technologies. In this paper, we studied on several machine learning algorithms to solve this problem and we propose a comparative study of some of the mostly used non-performing loan models on a customer portfolio dataset in a private bank in Turkey. we also deal with a class imbalance problem using class weights. a dataset, composed by 181.276 samples, has been used to perform the analysis considering different performance metrics (i.e., precision, recall, f1 score, imbalance accuracy).

3. Research Methodology:

3.1 Data Description: The primary data for this research will be to get the real time opinions but as there are several researches conducted out this has a lot of secondary data available online for the research purposes in which the model has been created which notifies but also has alerted the user using many machine learning and deep learning algorithms. Due to the researchers conducted, a lot of data is available online. So, here we are going to use a dataset from Analytics Vaidya and we are mainly going to use a deep learning algorithm called Decision Tree.

4.1 Pre-processing: Initially the Attributes which are critical to make a Loan Credibility Prediction is identified with Information Gain as the attribute-evaluator and Ranker as the search-method. Manual pre-processing is also performed.

2. Final dataset after pre-processing is divided in such a way that there is 80 % training set and 20 % test set. Test set is used to validate the final result of the classifier.
3. Fitting a Decision-Tree algorithm to the Training set.
4. Predicting the test result.
5. Test accuracy of the result (Creation of Confusion matrix)
6. Visualizing the test set result.

4.1.1 Train Test Split Ups: The dataset has to be split into training data and testing data and it plays a vital role in increasing the accuracy for the results in the model. One of the mores in machine learning is that to divide the data into train data and test data into 80:20 format, it is the commonly used format but does not necessarily be the same. The table below shows how much data were used to train and test from each class. We must use same number of samples from each

class to train and then to test. That's one way to check how accurate the model is performing. The model and remaining 100 samples will be utilized to test the model performance. So, on the whole we have 3600 samples to train the model and 400 samples to test the model accuracy.

4.1.2 Data Mining: Data mining is the process of analysing data from different perspectives and extracting useful knowledge from it. This is the core of the knowledge discovery process. Various steps for extracting knowledge from raw data, Various data mining techniques include classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression, and more. Classification is the most commonly applied data mining technique that uses a pre-sorted set of samples to develop a model that can classify the entire population of a dataset. Fraud detection and credit risk applications are particularly well suited for classification techniques.

4.2 Important Terminologies in Decision Tree:

4.2.1 Information Gain: Information Gain is used to determine which feature/attribute gives us the maximum information about a class. Information Gain is based on the concept of entropy, which is the degree of uncertainty, impurity or disorder. Information Gain aims to reduce the level of entropy starting from the root node to the leave nodes. It is calculated by subtracting the sum of squared probabilities of each class from one. It favours larger partitions and easy to implement whereas information gain favours smaller partitions with distinct values. A feature with a lower Gini index is chosen for a split. The classic CART algorithm uses. Information is a measure of a reduction of uncertainty. It represents the expected amount of information that would be needed to place a new instance in a particular class. These informativeness measures form the base for any decision tree algorithms. When we use Information Gain that uses Entropy as the base calculation, we have a wider range of results whereas the Gini Index caps at one.

4.2.2 Gini Index: Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen. To find the best feature which serves as a root node in terms of information gain, we first use each descriptive feature and split the dataset along the values of these descriptive features and then calculate the entropy of the dataset. This gives us the remaining entropy once we have split the dataset along the feature values. Then, we subtract this value from the originally calculated entropy of the dataset to see how much this feature splitting reduces the original entropy which gives the information gain of a feature

The feature with the largest information gain should be used as the root node to start building the decision tree.ID3 algorithm uses information gain for constructing the decision tree.

4.2.3 Entropy: It is used to measure the impurity or randomness of a dataset. Imagine choosing a yellow ball from a box of just yellow balls (say 100 yellow balls). Then this box is said to have 0 entropy which implies 0 impurity or total purity.

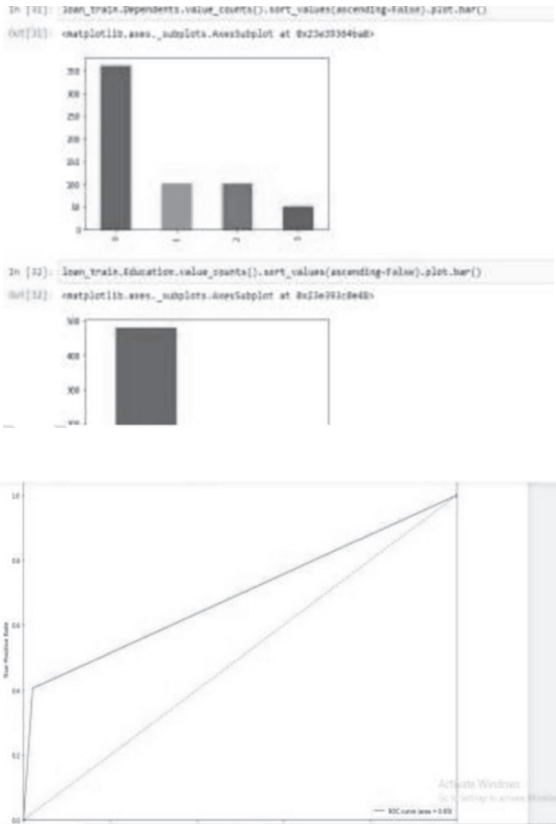
Now, let's say 30 of these balls are replaced by red and 20 by blue. If we now draw another ball from the box, the probability of drawing a yellow ball will drop from 1.0 to 0.5. Since the impurity has increased, entropy has also increased while purity has decreased. Shannon's entropy model uses the logarithm function with base 2 (log2(P(x)) to measure the entropy because as the probability P(x) of randomly drawing a yellow ball increase. When a target feature contains more than one type of element (balls of different colours in a box), it is useful to sum up the entropies of each possible target value and weigh it by the probability of getting these values assuming a random draw. This finally leads us to the formal definition of Shannon's entropy which serves as the baseline for the information gain calculation:

Where P(x=k) is the probability that a target feature takes a specific value, k.

4.3.5 Decision Tree: A Decision Tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root nodal. Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

5.Application Result: The below table shows the accuracy and other criteria when the model is developed.

```
In [10]: loan_ID=loan_train.Loan_ID

In [11]: loan_train.drop(["Loan_ID"],axis=1,inplace=True)

In [12]: loan_train['Gender'].value_counts()

Out[12]: Male      489
         Female    112
         Name: Gender, dtype: int64

In [13]: loan_train["Gender"]=loan_train.Gender.astype('str').transform(lambda x: x.replace('nan','Male'))

In [14]: loan_train['Married'].value_counts()

Out[14]: Yes    398
         No     213
         Name: Married, dtype: int64

In [15]: loan_train["Married"]=loan_train.Married.astype('str').transform(lambda x: x.replace('nan','Yes'))

In [16]: loan_train.Dependents.value_counts()

Out[16]: 0     345
         1     102
         2     101
         3+     51
         Name: Dependents, dtype: int64
```



From a proper analysis of positive points and constraints on the component, it can be safely concluded that the product is a highly efficient component. This application is working properly and meeting to all Banker requirements. This component can be easily plugged in many other systems. There have been numbers cases of computer glitches, errors in content and most important weight of features are fixed in automated prediction system, so in the near future the called software could be made more secure, reliable and dynamic weight adjustment. In near future this module of prediction can be integrate with the module of automated processing system. The system is trained on old training dataset in future software can be made such that new testing date should also take part in training data after some fix time.

References:

- *Accurate loan approval prediction based on machine learning approach J. Tejaswini1 ,T. Mohana Kavya2 , R. Devi Naga Ramya3 , P. Sai Triveni4 Venkata Rao Maddumala*
- *International journal of engineering research & technology (ijert)*
- *Dileep B Desai, dr. r.v.kulkarni a review: Application of data mining tools in CRM for Selected Banks, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2)*
- *Rob Gerritsen, Loan Risks: A Data Mining Case Study.*
- *Dr. Madan Lal Bhasin, Data Mining: A Competitive Tool in the Banking and Retail Industries,*
- *S. Kotsiantis, D. Kanellopoulos, P. Pintelas, "Data Pre-processing for Supervised Leaning", International Journal of Computer Science, 2006, Vol 1 N. 2, pp 111117.*
- *Bharati M. Ramageri, DATA MINING TECHNIQUES AND APPLICATIONS, Indian Journal of Computer Science and Engineering Vol. 1 No. 4*
- *Vivek Bhambri Application of Data Mining in Banking Sector, International Journal of Computer Science and Technology Vol. 2,*
- *P.Sundari, Dr.K.Thangadurai An Empirical Study on Data Mining Applications, Global Journal*
- *Kazi Imran Moin, Dr. Qazi Baseer Ahmed Use of Data Mining in Banking*
- *Rajanish Dass, "Data Mining in Banking and Finance: A Note for Bankers",*
- *Loan Applicants based on Risk profile*
- *Hamid Eslami Nosratabadi, Sanaz Pourdarab and Ahmad Nadali, A New Approach*