Project Title

Machine Learning Modelling for Assessing Credit Worthiness & Control NPA in Banking Sector

College Advisor: Prof. Malkit Saini

Session: 2022-23

Submitted by: Ms. Neetu Kulkarni

Roll Number: 12BE2022

B.Tech. Computer Science

April 2022

Organization: YBI Foundation, Delhi

College Name: Akash College of Fundamental Sciences, Maharashtra

ABSTRACT

The role of machine learning is increasing day by day with the availability of data and computing power. These models help in prediction and also identify the anomalies. In real world there is challenge to identify the credit worthiness of bank customers. Formally bank are using their own assessment and CIBIL like scores. But even then there are incidence of default. To attend this problem and create a machine learning model which can help Indian banks and NBFC in better credit assessment. Reserve Bank of India is issuing guidelines for regulating NPA and overall banking ecosystem. Therefore, our project will help the financial institutes to upgrade their credit disbursement process with the addon of machine learning models.

Acknowledgement

I would like to thank my college faculty and advisor Prof. Malkit Saini for guiding and providing the opportunity to complete the internship. I would also like to than out Training and Placement In charge Mr. Raam Swaroop for arranging the internship at YBI Foundation.

I have been privileged to learn and upskill at YBI Foundation for two months period. The admin staff and academic team is great and provide in-depth knowledge with hands-on practice to master the concepts.

I would also like to thanks my parents for always support me in my life.

Table of Content

List of Figures

List of Abbreviations

AI: Artificial Intelligence

ANN: Artificial Neural Network

DL: Deep Learning

ML: Machine Learning

Chapter 1: Introduction

The common man of the country deposits his hard-earned money in banks. Banks give this accumulated capital as a loan to needy people and industrialists. While giving a loan, it is expected that the person or industrialist should return this money to banks through EMI. But when this loan does not come back to the bank, then it goes into the category of NPA. In simple words, it is a submerged debt of Banks. According to banking rules, when the EMI, Principal amount or interest of a loan does not come within 90 days of the due date, it is called NPA. That is, when a bank stops getting returns from a loan, then it becomes an NPA or bad loan for the bank.

When a bank offers a loan, it charges interest on the amount, which is why it is regarded as an asset to the bank. When the borrower stops paying the interest, or the principal, or both, the lender loses money. Such a loan then becomes a non-performing asset (NPA) for the bank. The banking industry in India is seriously affected by the NPA crisis with the rising number of defaulters.

As per the Reserve Bank of India (RBI), a loan is considered a “bad loan”, or an NPA when the interest due for any quarter is not fully paid within 90 days from the end of the quarter. However, this time period may vary based on the terms and conditions agreed upon by the bank and the borrower. A commonly accepted definition of NPA is: “An asset, including a leased asset, becomes non-performing when it ceases to generate income for the bank.”

Machine learning is a sub-domain of computer science which evolved from the study of pattern recognition in data, and also from the computational learning theory in artificial intelligence. It is the first-class ticket to most interesting careers in data analytics today. As data sources proliferate along with the computing power to process them, going straight to the data is one of the most straightforward ways to quickly gain insights and make predictions.

Machine Learning can be thought of as the study of a list of sub-problems, viz: decision making, clustering, classification, forecasting, deep-learning, inductive logic programming, support vector machines, reinforcement learning, similarity and metric learning, genetic algorithms, sparse dictionary learning, etc. Supervised learning, or classification is the machine learning task of inferring a function from a labelled data. In Supervised learning, we have a training set, and a test set. The training and test set consists of a set of examples consisting of input and output vectors, and the goal of the supervised learning algorithm is to infer a function that maps the input vector to the output vector with minimal error. In an optimal scenario, a model trained on a set of examples will classify an unseen example in a correct fashion, which requires the model to generalize from the training set in a reasonable way. In layman's terms, supervised learning can be termed as the process of concept learning, where a brain is exposed to a set of inputs and result vectors and the brain learns the concept that relates said inputs to outputs. A wide array of supervised machine learning algorithms are available to the machine learning enthusiast, for example Neural Networks, Decision Trees, Support Vector Machines, Random Forest, Naïve Bayes Classifier, Bayes Net, Majority Classifier etc., and they each have their own merits and demerits. There is no single algorithm that works for all cases, as merited by the No free lunch theorem. In this project, we predict based upon dataset the possible credit defaulters in bank.

Chapter 2: Data Source, Properties and Challenges

We have collected the data from open source where the bank customer information is masked to save guard their privacy. The dataset is rich source of information with many variables used by banks to track the transaction and demographic details of customers. The dataset is available in comma separated value (CSV).

Before attempting any machine learning modelling, one should clean, manipulate, and visualize data for robust modelling. This is also the opportunity for modeller to visualize data and observe pattern and any anomaly before attempting modelling.

Data is of 8 MB with one target variable as default status and twenty eight features. We will check our data for data type, missing values, outlier, encoding, distribution, correlation and summary statistics. If needed will pre-process the data for each of them.

We have taken care of missing values by using the mean in each column and one can also use K mean imputer.

Non scaled data is a big challenge in modelling as it will unnecessary influence or bias the results. We have done standard scaling of each numerical feature for unbiased modelling.

Further explain data

Chapter 3: Modelling Approach

There are various models available for attempting classification problem (as out target variable is categorical). In banking sector a white box model with easy explanation is preferred both for implementation and compliance. Therefore, we have not attempted any of the deep learning model in our project. The supervised machine learning models we have attempted are

a. Logistic Regression

b. Naïve Bayes

c. Decision Tree

d. Random Forest (to compare accuracy)

Lets, us give brief explanation of each of the used model in our project report

a. **Logistic Regression:** Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

b. **Naïve Bayes Classifier:** In machine learning, a Bayes classifier is a simple probabilistic classifier, which is based on applying Bayes' theorem. The feature model used by a naive Bayes classifier makes strong independence assumptions.

This means that the existence of a particular feature of a class is independent or unrelated to the existence of every other feature. Naive Bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method. The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class. To make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability.

c. **Decision Tree Classifier:** Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed based on features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, like a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. To build a tree, we use the CART algorithm, which stands for Classification and

Regression Tree algorithm. A decision tree simply asks a question and based on the answer (Yes/No), it further split the tree into subtrees.

d. **Random Forest Classifier:** The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. Feature randomness, also known as feature bagging. Random Forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Let's look at random forest in classification, since classification is sometimes considered the building block of machine learning.

In any modelling finding the model accuracy is fundamental for any use. We have used the below metrics for model comparison and accuracy

a. Accuracy: Classification accuracy is a metric that summarizes the performance of a classification model as the number of correct predictions divided by the total number of predictions. It is easy to calculate and intuitive to understand, making it the most common metric used for evaluating classifier models.
b. Sensitivity or Recall: Recall for Imbalanced Classification. Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made.

c. Pression: Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that were retrieved. Both precision and recall are therefore based on relevance.

d. Confusion Matrix: A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.

e. Support: Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing.

f. Specificity — Ratio of true negatives to total negatives in the data. Important when: you want to cover all true negatives. Used when: you don’t want to raise false alarms. For example, you’re running a drug test in which all people who test positive will immediately go to jail.

g. F1-Score — Considers both precision and recall. It’s the harmonic mean of the precision and recall. Important when, we have an uneven class distribution. Used when, the cost of false positives and false negatives are different. F1 score conveys the balance between the precision and the recall. It is higher if there is a balance between Precision and Recall. F1 Score isn’t so high if one of these measures, Precision or Recall, is improved at the expense of the other.

Chapter 4: Algorithm and Model Building

Share and explain your codes in this chapter

Chapter 5: Results and Conclusion

In this chapter share results and conclusion. Compare different models metrics in table format.

Reference

Mention any reference used or website