Loan Risk Prediction Using Transaction Information

Yashaswi Raj Suresh Kumar, Brijesh Kumar Singh
Department of Information Technology, JECRC Jaipur

Abstract

In today’s world there are many risks involved in bank loans, so as to reduce their capital loss; banks should perform the risk and assessment analysis of the individual before sanctioning loan. In the absence of this process there are many chances that this loan may turn in to bad loan in near future. Banks hold huge volumes of customer behavior related data from which they are unable to arrive at a decision point i.e. if an applicant can be defaulter or not. This can be achieved using the data mining techniques. Data analysis can be done using the data mining techniques. Here customers data sets compared with the trained data sets and depend on that comparison final prediction can be done. Data Mining is a promising area of data analysis which aims to extract useful knowledge from tremendous amount of complex data sets. We are going to implement a model for the bankers that help them predict the credible customers who have applied for loan. This model can be used by the organizations in making the right decision to approve or reject the loan request of the customers. Bank plays a vital role in market economy. The success or failure of organization largely depends on the industry’s ability to evaluate credit risk. Before giving the credit loan to borrowers, bank decides whether the borrower is bad (defaulter) or good (non defaulter). The prediction of borrower status i.e. in future borrower will be defaulter or non defaulter is a challenging task for any organization or bank.基本上 loan defaulter prediction is a binary classification problem. Loan amount, costumer’s history governs his credit ability for receiving loan. The problem is to classify borrower as defaulter or nor defaulter. However developing such a model is a very challenging task due to increasing in demands for loans. Prototypes of the model which can be used by the organizations for making the correct or right decision for approve or reject the request for loan of the customers. This work includes different machine learning models. Based on our demands and requirements, we may need to choose different methods.

Keywords

loan risk, Loan defaulter prediction, binary classification problem, predicting future of loan by machine learning

I. Introduction

There are various areas in which data mining can be used in financial sectors like customer segmentation and profitability, high risk loan applicants, predicting payment default, marketing, credit analysis, ranking investments, fraudulent transactions, optimizing stock portfolios, cash management and forecasting operations, most profitable Credit Card Customers and Cross Selling. There are many different types of loans you have to take into account when you’re looking to borrow money and it’s important to know your options. Loan categorization refers to the process of evaluation loan collections and assigning loans to groups or grade based on the perceived danger and other related loans properties. The process of monitoring the quality of the loan portfolios and to take action to counter fall in the credit quality of the portfolios.
It is required for banks to use more complicated internal classification schemes than the more standardized schemes that bank managers need for reporting reasons and that are intended to make easy observing and inter bank evaluation.

There are many types of loans such as: Open-ended loans are loans that you can have a loan of more and more. Credit cards and lines of credit are the famous types of open-ended loans. You have a credit limit that you can buy with both of these two types of loans. In any time you can purchase automatically your available credit will decreases. since you make expenditure, you’re on hand increases permitting you to use the credit more and more. Closed-ended loans, this type of loans cannot be on loan once they’ve been repaid. while you make expenditure on close dended loans, the balance of the loan became downward. though, you don’t have any existing credit you can employ on closed-ended loans. As an option, if you want to lend more money, you’d have to make application for other loan. widespread types of closed-ended loans involve auto loans, mortgage loans, and student loans.[2]

Credit risk analysis can be thought of as an extension of the credit allocation process. After an individual or business applies to a bank or financial institution for a loan, the bank or financial institution analyzes the potential benefits and costs associated with the loan. Credit risk or credit default risk is a type of risk faced by lenders. Credit risk arises because a debtor can always renege on their debt payments.

In the lead-up to the 2008 Great Recession, commercial banks, investment banks, and other financial markets participants underestimated both the default probability and the loss rate and consequently underestimated the credit risk they were facing.

II. Literature Review
Since the global financial crisis, risk management in banks has gained more prominence, and there has been a constant focus on how risks are being detected, measured, reported and managed. Considerable research (Van Liebergen 2017; Deloitte University Press 2017; Helbekkmo et al. 2013; MetricStream 2018; Oliver Wyman 2017), both in academia and industry, has focused on the developments in banking and risk management and the current and emerging challenges. In tandem, there has been a growing influence of machine learning in business applications, with many solutions already implemented and many more being explored. McKinsey & Co highlighted that risk functions in banks, by 2025, would need to be fundamentally different from what they are today. In [2] the author introduces a framework to effectively identify the Probability of Default of a Bank Loan applicant. New products, services and risk management techniques are being enabled through the application of evolving technologies and advanced analytics. Machine learning, identified as one of the technologies with important implications for risk management, can enable the building of more accurate risk models by identifying complex, non linear patterns within large datasets. The predictive power of these models can grow with every bit of information added, thus enhancing predictive power over time. It is expected that machine learning will be applied across multiple areas within a bank’s risk organisation. Machine learning has also been recommended as an initiative that could help in the transformation of the risk management function at banks.

2.1 Machine Learning
Machine learning has been explained as lying at the intersection of computer science, engineering and statistics. It has been highlighted as a tool that can be applied to various problems, especially in fields that require data to be interpreted and acted upon (Awad and Khanna 2015). The work in [4] develops many credit scoring models that are based on the multilayer approach. The work proves its performance than the other models that uses logistic regression techniques. The results show that the neural network model
performs better than the other three techniques. The work in [5] compares support vector machine based models for credit-scoring developed using the various default definitions. The work concluded that the broad definition models are better than the narrow definition models in their performance. Financial data analysis is done in [6] using the techniques such as Decision Tree, Random forest, Boosting, Bayes classification, Bagging algorithm and others. Support Vector Machine, Decision Tree, Logistic Regression, Neural Network, Perceptron model, all these techniques are combined in this model. The effectiveness of applying the above techniques on credit scoring is studied. The analysis results show the performance is outstanding based on accuracy.

Machine learning tools that are driving the advances in search engines and self-driving cars can be adopted and applied to the financial sector. A variety of technological developments have contributed to the financial sector being able to explore and mine a voluminous data infrastructure that includes diverse sets of unstructured forms of financial data about markets and consumers. Economists are increasingly adopting machine learning, in conjunction with other tools and expertise to evaluate complex relationships, despite machine learning’s limitations in being able to determine causality. The adoption of machine learning has been motivated by the potential opportunities for cost reduction, improved productivity and improved risk management. New regulations have also pushed the banks to automate with the need to have efficient regulatory compliance (Financial Stability Board 2017). Data driven and computational-based, machine learning algorithms rely less on assumptions about the data, including about the distribution. While they are considered more robust and better at addressing complex non-linear relationships, they also are seen as being difficult to interpret (Galindo and Tamayo 2000). Recent years have seen a surge in the amount of data gathered within financial institutions (FI). A big push towards the digitalisation of services and increased regulatory reporting requirements has resulted in a large amount of unstructured data being created and/or collected at a high frequency. This data comes from various sources, including consumer apps, client interactions, metadata and other external data sources. The desire to enhance their analytical capabilities and automate across business lines, including risk management, by managing and mining these increased volumes and a variety of data has led financial institutions to explore powerful and analytical solutions, a consequence of which is the rise in interest and the popularity of machine learning and artificial intelligence within the FI community (Van Liebergen 2017). Machine learning is widely seen in the financial services sector as having the potential to deliver the analytical capability that FIs desire. Machine learning is capable of impacting every aspect of the FI’s business model—improving insight into client preferences, risk management, fraud detection, conduct monitoring, client support automation and even automated identity verification when coupled with biometrics.

III. Problem Statement

People often save their money in the banks which offer security but with lower interest rates. Lending Club operates an online lending platform that enables borrowers to obtain a loan, and investors to purchase notes backed by payments made on loans. It is transforming the banking system to make credit more affordable and investing more rewarding. But this comes with a high risk of borrowers defaulting the loans. Hence there is a need to classify each borrower as defaulter or not using the data collected when the loan has been given. The objective of proposed work is to predict loan credit risk and determine the probability of non-payment of bank financial services.e.g. whether a person will pay back a loan or not. The other objective of the project is to study the ability of
Machine learning algorithms to handle the problem of predicting credit default that measures the creditworthiness of the loan application over a time period. However, there are many risks related to bank loans, for the bank and for those who get the loans.

Risk prediction and monitoring is critical for the success of the business model. Credit risk is the probability that a customer won’t be able to make a required payment, causing a loss for the bank or financial institute that provided the loan.

To classify if the borrower will default the loan using borrower’s finance history. That means, given a set of new predictor variables, we need to predict the target variable as 1 - Defaulter or 0 - Non-Defaulter. The metric we use to choose the best model is ‘False Negative Rate’. (predictor and target variables explained later)

IV. Methodology

The proposed model focuses on predicting the credibility of customers for loan repayment by analyzing their behavior. The input to the model is the customer behavior collected. On the output from the classifier, decision on whether to approve or reject the customer request can be made. 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. To extract patterns from a common loan approved dataset, and then build a model based on these extracted patterns. The training data set is now supplied to machine learning model; on the basis of this data set the model is trained. Every new applicant details filled at the time of application form acts as a test data set. After the operation of testing, model predict whether the new applicant is a fit case for approval of the loan or not based upon the inference it conclude on the basis of the training data sets. To extract important information and predict if a customer would be able to repay his loan or not.[1] we are taking the raw data from our source and will apply some data cleaning methods to make our data smooth. Then the most important step is feature extraction and selection will be applied to select best features out of available which are influencing the result more. Our problem can be expressed as Supervised learning problem where we have target variable specified. We have used supervised learning algorithms like K-nearest neighbors, support vector machines and logistic regression etc.

We had 140,000 observations of data, which took a lot of time for training. By implementing learning curve for our data, we realized that our models do not learn after 7000 observations. So, we downsized it. Our data has been preprocessed using techniques like scaling, filling null values.

Before beginning to train models we should transform our data in a way that can be fed into a Machine Learning model. The most common techniques are:

1. Dealing with missing data

Missing values are typically represented with the “NaN” or “Null” indicators. The problem is that most algorithms can’t handle those missing values so we need to take care of them before feeding data to our models. Once they are identified, there are several ways to deal with them: Eliminating the samples or features with missing values. Imputing the missing values, with some pre-built estimators such as the Imputer class from scikit learn.

We’ll fit our data and then transform it to estimate them. One common approach is to set the missing values as the mean value of the rest of the samples.

2. Feature Scaling

This is a crucial step in the preprocessing phase as the majority of machine learning algorithms perform much better when dealing with features that are on the same
scale. The most common techniques are:

Normalization: it refers to rescaling the features to a range of [0,1], which is a special case of min-max scaling. To normalize our data we’ll simply need to apply the min-max scaling method to each feature column.

Standardization: it consists in centering the feature columns at mean 0 with standard deviation 1 so that the feature columns have the same parameters as a standard normal distribution (zero mean and unit variance). This makes much more easier for the learning algorithms to learn the weights of the parameters. In addition, it keeps useful information about outliers and makes the algorithms less sensitive to them.

3 Learning Curve

This learning curve clearly shows that our models are not learning anything after ~11,000 samples. So we randomly sampled our dataset and used only 11,000 samples

Note: In the plot, only ~9000 samples are shown because it is a plot for training set.

The AUC – ROC curve is a performance measurement for classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells you to what extent the model is capable of distinguishing between classes.

4.1 Machine Learning Algorithms

1. Random forest

(RF) is an advanced form of decision trees (DT) which is also a supervised learning model. RF consists of large number of decision trees working individually to predict an outcome of a class where the final prediction is based on a class that received majority votes. The error rate is low in random forest as compared to other models, due to low correlation among trees.

We have a feature ‘last_pymnt_amnt’ which has an importance of more than 30%. Random forests select a subset of features in each of its decision trees thereby reducing the bias (because of high importance of single feature) of the model. The final output will be the mode of the outputs of all its decision trees which has better results than decision trees (which can possibly overfit). Hence, we chose to start our classification with random forests. Random forest when implemented with randomized search we got the best accuracies and minimum false negatives(predicting borrower will not default even though he will. This might impact on the credibility of the company). We used the randomized search to find the best hyper parameters for the model.

2. Multi Layer Perceptron

It utilizes backpropagation for training. Its multiple layers and non-linear activation function help us distinguish data that is not linearly separable. A multi-layered perceptron (MLP) is one of the most common neural network models used in the field of deep learning. Often referred to as a “vanilla” neural network, an MLP is simpler than the complex models of today’s era. However, the techniques it introduced have paved the way for further advanced neural networks. The multilayer perceptron (MLP) is used for a variety of tasks, such as stock analysis, image identification, spam detection, and election voting predictions.
Here, $b^{(1)}$ and $b^{(2)}$ are the bias vectors, $w^{(1)}$ and $w^{(2)}$ are the weight matrices, and $g$ and $s$ are the activation functions. In our case, the activation function is ReLU and the Adam solver, with 3 hidden layer.

3. Support Vector Machine

The objective of the support vector machine algorithm is to find a hyperplane in an $N$-dimensional space ($N$ — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Support vector machine (SVM) is another model for binary classification problem and is available in various kernels functions. The objective of an SVM model is to estimate a hyperplane (or decision boundary) on the basis of feature set to classify data points. The dimension of hyperplane varies according to the number of features. As there could be multiple possibilities for a hyperplane to exist in an $N$ dimensional space, the task is to identify the plane that separates the data points of two classes with maximum margin.

5. Logistic Regression

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical.

For example,
- To predict whether an email is spam (1) or (0)
- Whether the tumor is malignant (1) or not (0)

Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequence in real time.

From this example, it can be inferred that linear regression is not suitable for classification problem. Linear regression is unbounded, and this brings logistic regression into picture. Their value strictly ranges from 0 to 1.

With logistic regression, outputs have a nice probabilistic interpretation, and the algorithm can be regularized to avoid overfitting. Hence, we choose to build logistic regression classifier.

6. Bagging Ensemble Classifiers

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into
its construction procedure and then making an ensemble out of it.

Each base classifier is trained in parallel with a training set which is generated by randomly drawing, with replacement, N examples (or data) from the original training dataset – where N is the size of the original training set. Training set for each of the base classifiers is independent of each other. Many of the original data may be repeated in the resulting training set while others may be left out.

V. RESULTS & DISCUSSIONS

After applying many machine learning algorithms like Logistic Regression, Support Vector Machine, K Nearest Neighbor, Multi layer Perceptron etc and building models we need to evaluate their performances and find the best algorithm for our model.

For our Loan default prediction project, False Negatives Rate is the best metric to evaluate the model. Lower the number of false negatives, better the model is. In this project, False negative is when model predicting “a borrower will not default a loan even though he will “. Our model cannot afford having higher False Negatives as it leads to negative impact on the investors and the credibility of the company. So, we evaluated our models using the number of False negatives and accuracies.

1. Model Accuracies:
   Random Forest with Randomized search CV ----- 82.09
   Logistic Regression with Grid search CV ----- 83.18
   Support Vector Machine with Grid search CV----- 82.50
   K Nearest Neighbors with Grid search CV ----- 77.40
   Bagging with Base estimator as Random Forest ----- 84.10
   Bagging with Base estimator as Logistic Regression ----- 83.10
   MultiLayer Perceptron Classifier 83.40

2. Although all the algorithms except KNN have almost same accuracy, their False Negative rates differ which is Evaluation Metric

our main evaluation metric. From the confusion matrices, we can infer that Logistic Regression model had the least False Negative rate (FNR).

VI. CONCLUSIONS

This Model can help banks in predicting the future of loan and its status and depends on that they can take action in initial days of loan. Using this model banks can reduce the number of bad loans and from incurring severe losses. Several python functions and packages were used to prepare the data and to build the classification model. Python libraries help in successful data analysis and feature selection. Using this methodology bank can easily identify the required information from huge amount of data sets and helps in successful loan prediction to reduce the number of bad loan problems. Data Mining techniques are very useful to the banking sector for better targeting and acquiring new customers, most valuable customer retention, automatic credit approval which is used for fraud prevention, fraud detection in real time, providing segment based products, analysis of the customers, transaction patterns over time for better retention and relationship, risk management and
marketing Machine Learning can help banks in predicting the future of loan and its status and depends on that they can act in initial days of loan. Using Machine Learning banks can reduce the number of bad loans and from incurring sever losses. We tried to build a model which had least False negative Rate.

VI. REFERENCES


