Approaches for Credit Risk Evaluation

Youtube link for the video presentation: https://youtu.be/9uky8oz1ZKo

By

Chinta Manikanta Mahesh Reddy - 18NA10006

Nanisetty Sai Shailesh - 18IM30013

Akash Nitin Dusane - 18QM30002

Contents:

Abstract

In the Finance Sector, Credit Score play an important role in assessing and determining a subject's creditworthiness. The higher a customer's credit score, the higher will be its creditworthiness, and accordingly, lower will be the risk of credit default and more likely to be eligible for a credit facility.

Although Small and Medium-Sized Entities (SMEs) and to a lesser extent, large entities, can also be assigned a credit score, it is more widely entrenched in the consumer market — where a consumer's credit score is the de-facto measure of its lending eligibility.

Credit scores are extensively utilized by the financial services industry as a standard mechanism to qualify potential borrowers, classify them according to the internal risk criteria, and price a credit facility accordingly.

Given the recent technological advancements, credit scores are now usually developed using big data analytics and efficient machine learning algorithms.

Introduction to Credit Scores

The data features/factors that are used to carry out statistical analysis and calculate credit scores:

- Repayment history: whether the amount due was settled in time or not
- The total amount owed: total outstanding balance, usually as a percentage of the total credit limit
- Length of credit history: when was the first credit line opened. A general criterion is that the longer an individual has had credit, the better will be his credit score
- Types of credit lines: does the consumer has only one kind of credit line (e.g., only mortgage) or a good mix of retail loans, credit cards, mortgage, and vehicle finance
- New credit lines: number of new credit lines opened in a short amount of time is generally looked unfavorable
- Available credit: unutilized balance on open credit lines

The features used mainly contain customer details, transaction and defaulter history

Approaches to building a Credit Risk Model

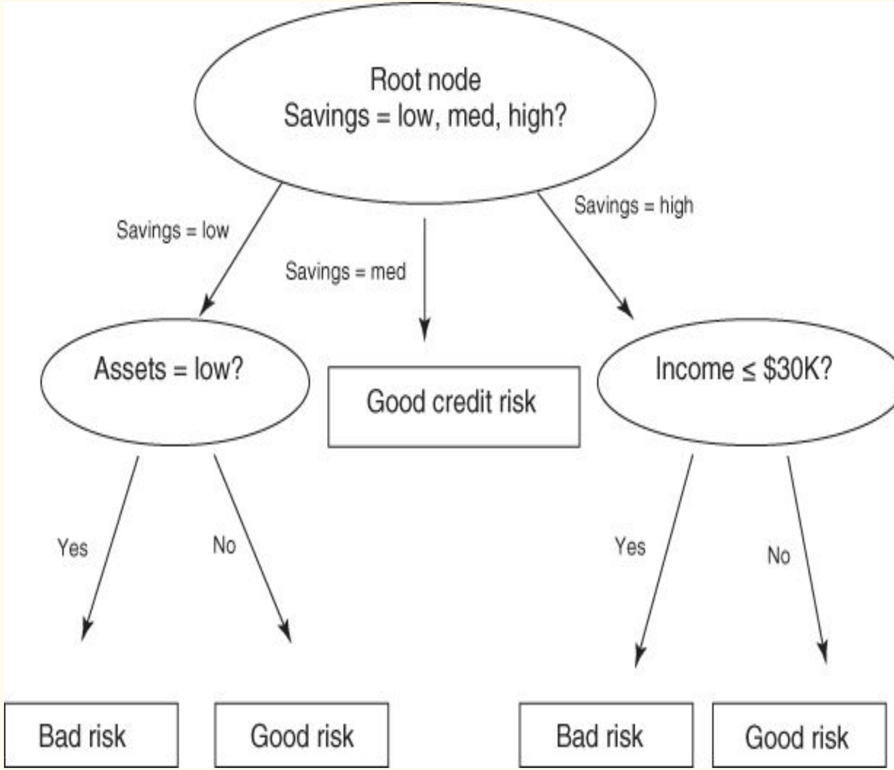
We can use various approaches to build a model predicting the credit scores.

❖ Decision Tree Classifier
❖ Gradient Boosting
❖ Deep Learning Approaches

Before building a model, the data is cleaned, preprocessed and important features are selected accordingly after carrying out ANOVA test. The data is usually highly imbalanced and thus require oversampling to balance the data.

Model 1: Decision Tree Classifier

- ❖ It is a type of non-parametric supervised learning method used for classifying the person as a defaulter or not.
- ❖ This algorithm uses the tree representation to solve the problem, internal nodes in the tree to represent an attribute and leaf nodes to represent class labels.
- ❖ We start from the root of the tree, separate the sample into groups of homogeneous sets according to most essential splitters of input variables.
- ❖ This process is repeated until leaf nodes in all the branches of the tree are found.



ADVANTAGES:

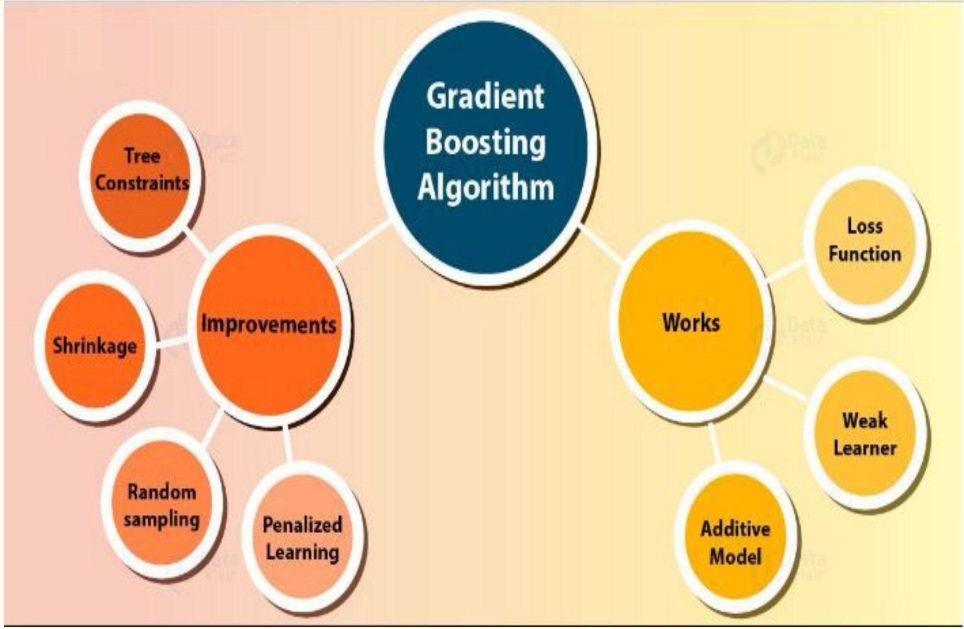
- This algorithm is easy to visualize, understand, and interpret.
- This method is helpful in variable screening and feature selection especially when we want to select important financial variables which impacts the selection of credit defaulter meaningfully.
- Also, Handling high dimensional data such as the large volumes of historical credit information becomes easier requiring little user's data preparation with this approach.
- Furthermore, this method can deal with both quantitative and qualitative data, and this property is superior to other methods because other clas
require single attribute of the data.

Drawbacks & Future Scope

- ❖ The main drawback with this model is overfitting. Because of this, It can give wrong results and hence increases the probability of credit default by classifying bad customers as good and vice versa.
- ❖ The reason the decision tree is likely to have the problem of overfitting when we don't set a limitation of maximum depth is because the tree can keep growing until it has exactly one leaf node for every single observation.
- ❖ Also, This algorithm doesn't have stability because even small variation in the datasets might lead to a completely different decision tree being generated.
- ❖ So, To mitigate the issues w.r.t the decision tree, we further experiment in our next method using an ensemble approach.

Model 2: Gradient Boosting

- Boosting is an important ensemble approach for improving prediction of a decision tree wherein the trees grow sequentially,
- The boosting model is built by an iterative and sequential process where a base model is built and tested with all of the training data, and based on the outcome the next base model is developed.
- We'll focus on gradient boosting which builds an ensemble of trees one-by-one and it generalizes them by allowing optimization of an arbitrary differentiable loss function.
- It is composed of 3 parameters: Number of trees (N), Interaction depth (Maximum nodes per tree) and the Learning rate (penalizes the importance of each consecutive iteration and reduces the size of incremental steps).



Advantages

- The boosting process concentrates on the training records that are hard to classify and over-represents them in the training set for the next iteration.
- In general, Gradient boosting can improve upon random forests and is easier to interpret because of the smaller tree structure.
- With this method, there has been a significant improvement in overall accuracy with increase in recall and precision.
- Also, the model was able to identify debt to income ratio as the most important parameter.

Drawbacks & Future scope

- Despite having robust models that extensively assess an individual’s financial credibility, banks are still exposed to the risk of potential loan default.
- .With the help of advanced data analytics and contemporary prediction techniques, banks are exploring ways in which credit scoring models can be made more robust with higher accuracy level.
- Historically, Ensemble methods have been good and have worked better than individual models with higher accuracy exhibiting more apparent benefits in credit risk prediction.
- However, With explosion in volumes of financial data and different kinds of transactions on various platforms; we need to have more efficient & robust models which can be achieving by deep learning.

Deep learning ensemble credit risk evaluation model

Base learner for credit score calculation : LSTM network

Algorithms : Adaboot

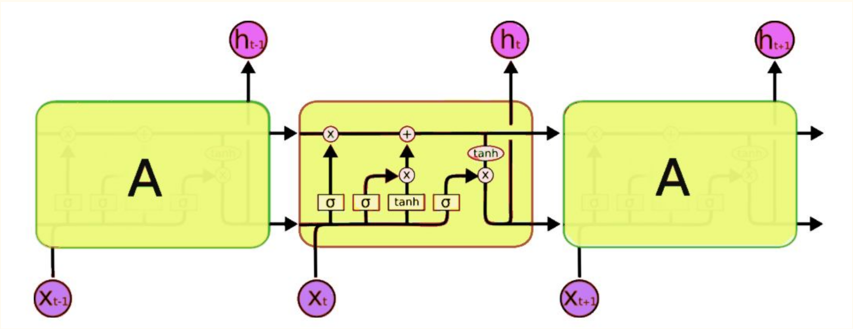
Pre-processing for imbalance data : improved synthetic minority oversampling technique

MODEL : First, an improved synthetic minority oversampling technique (SMOTE) method was developed to overcome known SMOTE shortcomings, after which a new deep learning ensemble classification method combined with the long-short-term-memory (LSTM) network and the adaptive boosting (AdaBoost) algorithm was developed to train and learn the processed credit data.

LONG SHORT TERM MEMORY NETWORK

Long Short Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies.

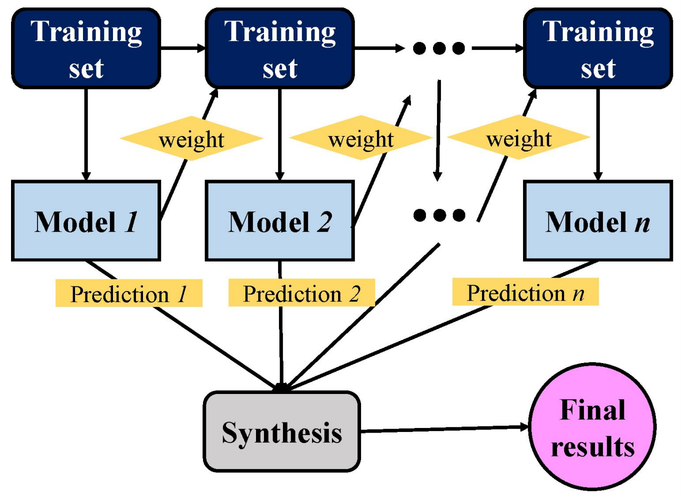
LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn!



Adaboost

AdaBoost, short for Adaptive Boosting

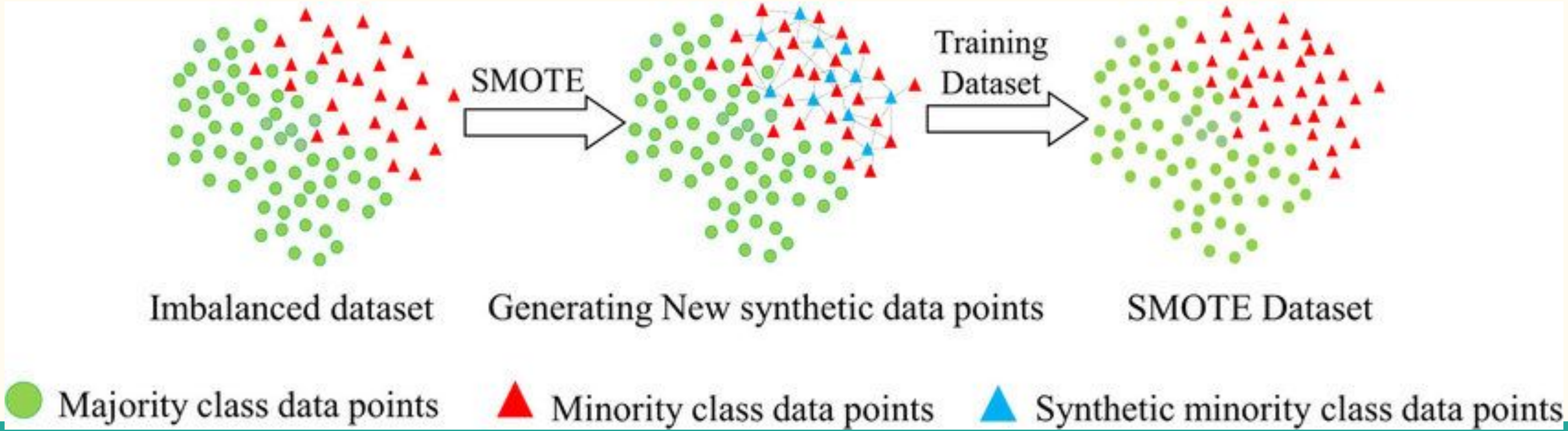
- ❖ The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier.
- ❖ AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers.
- ❖ In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner.



Synthetic Minority Oversampling Technique (SMOTE)

Most real world prediction problems have imbalanced data. Here an imbalance in dataset is represented by a mismatch in class representation. An application of classification algorithms on imbalanced data is biased in favor of the majority class and gets further biased in case of high-dimensional data. This class imbalance problem can be reduced by under-sampling of majority data or by oversampling of minority data.

Synthetic Minority Oversampling Technique (SMOTE) is one such popular technique



CONCLUSION

Statistical methods

↓

Machine Learning

↓

Deep learning

Reference

https://www.sciencedirect.com/science/article/pii/S1877050920315830 - Single Machine Learning Model

https://www.researchgate.net/publication/324573074_Application_of_Ensemble_Models_in_Credit_Scoring_Models - Ensemble Machine Learning Model

https://www.sciencedirect.com/science/article/pii/S1568494620307900?casa_token=SKnxPz_o6oIAAAAA:pGr9t7uwaxULTmFoJQj5uNdKy032dR6s5XbKGayCLw4v2moEMgeHGamHSZUVwoWNlZ4gTBaPyAw - Deep learning ensemble technique

https://towardsdatascience.com/how-to-develop-a-credit-risk-model-and-scorecard-91335fc01f03 - Credit Risk Model

https://medium.com/swlh/how-to-quantify-risk-and-creditworthiness-c76725bc2380 - Finance point of view background over credit scores and its evaluation

Thankyou.