Credit Score: Credit Risk Valuation Using An Efficient Machine Learning Algorithm

Abstract. The process of automation is likely to improve the efficiency of the detection process, and it may also provide higher detection accuracy by removing the internal subjective human factors in the process. If machine learning can automatically identify bad customers, it will provide considerable benefits to the banking and financial system. The goal is to calculate the credit score and categorize customers into good or bad. Algorithms of machine learning library is used to classify the data sets of finance sectors. A large volume of multi-structured customer data is generated. When the quality of this data is incomplete the exactness of study is reduced. In the proposed system, we provide machine learning algorithms for effective prediction of various occurrences in societies. We experiment the altered estimate models over real-life bank data collected. Compared to several typical estimate algorithms, the calculation exactness of our proposed algorithm is high.

Keywords: : Machine Learning, Credit Scoring, Logistic Regression, Random forest, CRISP DM Framework.

1 Introduction

Hundreds of banks in the United States alone suffer from non-payment or late-payment of loans. Predicting such customers earlier facilitates preventive banking interventions, which in turn can lead to tremendous cost savings and improved outcomes. Algorithms are developed for predicting customer behavior by drawing from ideas and techniques in the field of machine learning. Standard classification methods are explored such as logistic regression and random forest, as well as more sophisticated sequence models, including recurrent neural networks. We focus especially on the use of banking code data for customer behavior prediction and explore different ways for representing such data in our prediction algorithms. A problematic information assortment mechanism is intended and therefore the correlation analysis of this collected knowledge is performed. A stochastic prediction model is designed to foresee the future condition of the most correlated customers based on their current account status. In banking and finance communities, a large volume of multi-structured customer data is generated from the transactions, account statements and online purchases.

Imagine a system where banks can quickly comb through millions of anonymized customer records to find people with good credit scores and bank experiences. Through this massive, searchable database, banks could determine how best to offer a loan, based on what has worked effectively for others with similar behavior and characteristics.

1.1 Precision Banking

What makes precision banking unique is that it goes beyond predicting for existing customers and conditions to predicting and preventing debts from new customers before they manifest. It stands at the intersection of finance, technology and big data, offering new ways to keep banks profitable. Precision banking is a way of translating data into information that can lead us to

prevent losses for banks in a way that we might not have done before. We are poised to have a whole new level of precision in maintaining banks.

In shifting through this data, researchers can better predict individual credit score, develop approaches to early detection and prevention, with information to help them make real-time decisions about the best way to offer loans for customers.

Large-scale data analysis also is enabling researchers to develop more targeted and cost-effective methods for early prediction of credit score before transactions are made.

The publicly available banking customer data is used to identify specific attributes associated with the defaulter condition, laying the groundwork for a simple test for defaulters.

1.2 Data Access

Banking has long been a data-rich field. With so many moving parts, banking providers and financers have no shortage of variables to measure. The captured data have many important uses. They keep tabs on credits and debits. They track the activity of transactions and savings. Crucially, the data record the states of people at a microscopic and macroscopic level. It is hard to overstate the importance of data in banking, especially when it comes to improving banking systems. Although in this work we focus on using banking data in credit score prediction, there are many other facets of banking sector that can be enhanced and even revolutionized through intelligent use of data. Obtaining access to banking data is often a fraught endeavor. Privacy laws and business concerns set a host of hurdles that must be cleared before data can be shared. Unfortunately, this can halt the progress of researchers unaffiliated with finance companies or banking systems. Despite these difficulties, the potential rewards of better understanding and utilizing banking data to improve banking sector far outweigh the frustrations of data access.

1.3 Data Capture

There are various systems in place for capturing banking data. Modern banking systems use tools for systematically and digitally storing a wide range of data, including customer demographics and account history, purchases, transactions, deposits, and more. Systems also facilitate data access and visualization, allowing bankers and customers to better inform themselves. Although these records only capture the activities that occur within a particular facility or set of facilities, they provide a vivid account of an individual’s state. Insurance claims form another rich repository of banking data. Claims data 3 center on individuals enrolled in the insurance policy and their interactions with the banking system. These records typically include basic demographic information about customers, along with purchases, transactions, deposits, savings and associated taxes. Because insurance claims data are so customer-centric and capture customer banking activity across a variety of banks, and financial organizations, they paint a rather comprehensive picture of an individual’s account history and current account state

1.4 Simple Variables

Much of banking data consists of simple numerical and categorical variables. These include

demographic variables such as age, sex, and ethnicity. Employment variables such as salary, job type, designation, work experience, and many others are also straightforward. These types of simple data are suited for standard analytical and statistical methods (such as linear or logistic regression). To stop with just the simple variables, however, would be to miss out on potentially valuable insight provided by more complex sources of data.

2.1 Statistical Observation:

Based on the data sets, a statistical approach is framed from the data. There is a large number of missing data due to human error. Thus, we need to fill the structured data. Before data imputation, First identify uncertain or incomplete medical data and then modify or delete them to improve the data quality. Using linear regression model to find the missing data. We use data integration for data pre-processing. To explain the variables in terms of the latent variables.

2.2 Training Dataset:

From all the training dataset, we apply the random forest algorithm in finding all the possible decision trees and predict the results. Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees. The decision tree algorithm is also implemented. Creating more number of decision trees. The calculation of nodes selection will be same for the same dataset. The same algorithm can use for both regression and classification problem.

2.3 Algorithms for Logistic regression and random forest:

- Step 1: Train the data on minimum dataset.
- Step 2: Check the results.
- Step 3: Apply on the complete dataset.
- Step 4: Generate the results from the model.

Form the above results generate the reports in terms of behavior of the data. Missing values problem can be resolved using machine learning algorithms.

3. Conclusion

A random forest model based multimodal credit score prediction algorithm using structured data from banking data set. Analyzing the factors based on the banking sector data. Missing values problem can be resolved using machine learning algorithms. Default customer prediction can be done based on the data and type of, region and risk level of the customer’s account status by the availability of the data.