Efficient methodology to identify the outliers in financial big data for fraud detection

Girija Attigeri

Title of the Thesis: Efficient methodology to identify the outliers in financial big data for fraud detection

Name of the Researcher: Girija Attigeri

Name of the Guide: Manohara Pai M M and Radhika M Pai,

Completed Year: 2020

Name of the Department: Manipal Institute of Technology

Abstract

The frauds in the public financial institutions like banks, Non-Banking Financial Companies (NBFCs) during the last 5 years have resulted in loss to the tune of Rs. 612 billion to the government. In today's digital world fraudulent cases are increasing day by day as there is no automated system for early fraud detection. An automated fraud detection system is very essential to strengthen the financial system whereby frauds can be predicted or detected early. According to the Economic Times, 18 public sector banks have faced a loss of Rs. 32000 crores in the first quarter of 2019 and most of these are due to bad loans. Bad loans are the Non Performing Assets (NPAs) which are used as one of the measures for assessing the financial health of a bank. The banking and financial system in the last 10 years have seen huge number of NPAs due to which the government has focused on changing the policies for controlling the same. Hence, it is very important to control NPAs, specifically happening due to wilful defaults. Wilful defaults occur because of the ill-intentions of the borrowers and hence constitute fraudulent activities. It is essential to build an automatic fraud detection system for banking to evaluate and detect fraudulent behaviour by the borrower.

It is observed from the literature that, though there is a significant research towards fraud detection, there is no effort towards building a complete framework for fraud detection system, which can be integrated with the existing automated banking system. Existing banking systems do not have mechanisms for early detection of NPA or wilful default. Banks follow the rules to declare a loan as NPA and then initiate actions. The fraud detection system will help banks to get alerts about loans, which are likely to become NPA/Wilful default.

The proposed framework for fraud detection system uses outlier detection approach for fraud detection, as the occurrence of fraud cases are less compared with legitimate cases. It begins with preprocessing of the data, building unsupervised and supervised models for outlier detection, performing Social Outlook Analysis (SOA) for wilful default identification, and subsequently building a web-based framework for fraud detection system. This fraud detection system can be deployed independently or can be integrated with the existing banking system.

The data preparation process is significant as it decides the effectiveness of the analysis. Hence, care is taken to identify the right parameters for data consideration by defining the parameterization process. Parameterization process provides a list of the parameters essential for predicting whether the loan will become NPA, and thereby wilful default. This process necessitates understanding of the existing system to check NPAs and to identify the critical parameters. The collected data having these parameters is cleansed and important features having strong correlation with NPA status are selected using proposed Correlation-based

Feature Selection using Submodular Optimization (CFS-SO). The prepared data is further used for outlier analysis. Outlier analysis is performed using two approaches unsupervised and supervised.

Unsupervised outlier detection uses clustering techniques to find fraudulent activities in corporate lending of banking system. All loans of banks are grouped into loans of small amount, large amount, their payment structure, etc. based on pattern similarity. Since the number of legitimate loans are large in numbers, the groups formed are of legitimate loans. Samples which do not belong to these groups are the outliers and hence, are identified as potential frauds. Subsequently, the loan data is labelled as Performing and NPA. In order to find outliers, various clustering algorithms such as K-Means, DBSCAN, kNN and Rule Based are implemented and evaluated. For the data in consideration K-Means shows the best performance and it is chosen to be used in the framework. Supervised models are built for detecting NPA/Wilful defaults for labelled loan data. The loan data is divided into training and test data. The training data is used to build various prediction models. The prediction models considered are Logistic Regression, Naïve Bayes, Neural Network, Support Vector Machine and Random Forest. These models are used for predicting NPA/Wilful default on the test data. The performance of these models is evaluated. The Neural Network performs best for the data in consideration. Hence, it is chosen to be used in the framework. The proposed Outlier analysis predicts whether the loan is NPA or not. In order to classify it as a loan with wilful default behaviour, SOA is done for the loans identified as NPA. If the loan is marked as NPA and SOA indicates positive social outlook, then the chances of loan being wilful default are very high.

The web based framework for fraud detection integrates all the phases, pre-processing, unsupervised outlier detection, supervised outlier detection, and SOA. This helps in classifying loan as genuine NPA or wilful default. The research also focuses on loan management system, where the loan monitoring and recovery phase is very crucial to avoid NPAs and further wilful defaults. The developed system is easy to deploy as a web application in any bank or financial firm for identification of fraud activities in the loans. The application aims at helping bank systems to become strong and ultimately to improve the financial health of the country. Research work contributes, methodologies and automated means to detect and prevent frauds in the banking system.