A Case Study in Financial Fraud Detection using Big Data Analytics

W. P. A. Boteju, I. U. Hewapathirana

Software Engineering Teaching Unit, Faculty of Science, University of Kelaniya, Sri Lanka

The financial industry is currently undergoing digital transformations across products, services and business models. This digitization is aimed at automating most of the manual financial transactions and other relevant services. Therefore, spotting fraud in financial transactions has become an important priority for all financial institutes. With the advances in modern technology and global communication, fraud has increased significantly, causing great damages. The focus of this paper is to experiment different approaches for detecting fraudulent activities in a real-world dataset of financial payment transactions. The dataset is obtained from Kaggle and consists of 6 million transaction records and 10 features with the transaction label as 'fraudulent' or 'non-fraudulent'. These features are investigated using exploratory data analysis and only 6 are retained for the experiment such as payment-type, account-balance, transaction-amount etc. Two supervised machine learning algorithms, the random forest and the support vector classifier are employed for detecting fraudulent transactions. The dataset is large and requires high computational power to process and train machine learning algorithms. Furthermore, another challenge is the highly imbalanced distribution between fraudulent (0.1%) and the non-fraudulent (99.9%) classes. The goal of this research is to solve both these issues. In order to handle class imbalance, the effect of oversampling the minority class data using the synthetic minority oversampling technique (SMOTE), and undersampling the majority class using random undersampling are investigated. Computational efficiency is achieved through the Apache Spark implementation, which provides distributed processing for big data workloads. The best performance is obtained using the random forest algorithm on the oversampled dataset with an accuracy of 99.95%, F1-score of 0.9994, recall of 0.9994, Geometric mean of 99.94% and a model training time of 13.9 minutes. This paper provides valuable insights on dealing with large scaled highly imbalanced big datasets for predicting financial frauds and generating alerts.

Keywords: Financial Fraud Detection, Big Data Analytics, Apache Spark, SMOTE, Ensemble Learning Methods