CM3203 – Final Year Project

Detecting fraud from customer transactions

Initial Plan

Author:
Stanislav Kataev

Supervisor:
Yuhua Li

Project Description

The e-commerce sector is one of largest markets of exchange of goods and services. With the growing popularity of the Internet, a huge number of companies moved to sell their goods and services to the digital space pursuing customers to start shopping online. The number of internet users is increasing every year causing an increase in transactions on e-commerce. With the rapidly increasing number of business starting to sell online, fraud detection and prevention became a very important topic due to its big impact on user's security and success of operating businesses.

This project looks into fraud detection in e-commerce transactions. E-commerce frauds are closely linked to payments using credit cards. One of the most typical type of fraud is identity theft which entails a scammer breaching a user account and stealing their personal information. The number of frauds related to e-commerce has been increasing every year since 1993 and in 2019 the total sum of revenue lost due to fraud has reached 70 trillion dollars [2]. The prevention of fraud in e-commerce is mainly solved by a behavioural analysis.

The common approach to fraud detection is a rule-based approach which uses algorithms that perform several fraud prevention tasks created by fraud analysts. The disadvantages of this approach are the problem with processing real-time data streams with enough efficiency. The approach does not detect fraudulent scenarios if they do not match rules manually written by a fraud analyst and are not efficient at detecting implicit correlations [3].

Thus, machine learning based fraud detection is currently the mostly used approach. The advantages of the approach are faster data processing and identification of correlations between user behaviours and likelihood of fraud. Advanced fraud detection systems use supervised or unsupervised machine learning either independently or combined together. Unsupervised machine learning is mainly used for clustering unlabelled data and detects hidden correlations. Supervised models are used to detect fraudulent transactions using labelled datasets. The common supervised learning models used for fraud detection are Random Forests, Support Vector Machine Model, K-Nearest Neighbours and Neural Networks. The goal of the project is to find the most appropriate model for the given dataset. The datasets for this project are taken from Kaggle and are provided by Vesta Corporation [4].

The data given for the project is highly imbalanced and various balancing approaches will be tested to establish the most effective one. After the dataset is balanced, the built supervised models will be trained using the achieved datasets and tested against testing datasets. Using testing methods such as AUROC, the models will be tested against each other and their performances will be evaluated. The final goal of the project is to find the most accurate model to use for fraud detection.

Project Aims and Objectives

The aims of the project are focused on labelling imbalanced data and comparing machine learning models in order to evaluate the most efficient model for fraud detection. Due to the time constraints, the use of pre-existing machine leaning libraries such as Skikit-learn, Keras and Tensorflow is necessary for completion of the project before the final deadline.

Aim

The aim of the project is to develop a method that detects fraudulent e-commerce transactions so that customers are protected from financial fraud.

Objectives

- Research ways of dealing with imbalanced data and choose the appropriate approach. This will be focused on SMOTE (Synthetic Minority Oversampling Technique), RandomUnderSampler and combined class methods (SMOTE + ENN), where ENN is edited-nearest neighbours.

- Build unsupervised learning model to segment data into clusters.

- Research supervised learning models suitable for fraud detection: Random forests, Naïve Bayes, K-nearest neighbours, Support Vector Machine and Neural Networks.

- Build the researched models using Python3 machine learning libraries.

- Train models using the training datasets.

- Run tests with models using the test datasets.

- Evaluate the effectiveness of each model by using methods such as AUROC (Area Under the Receiver- Operating Characteristic) by testing individual supervised learning models against unsupervised learning models, considering the trade-off between precision and recall.

- Evaluate the most effective model by testing it against the current antifraud standards:

    - Detection of fraud in real-time
    - Improvement of data credibility
    - Analysis of user behaviour
    - Uncovering of hidden correlations

Work plan

Deliverables

I plan to submit the following:

- Final report
- Source code
- Supporting documents containing research and analysis
- Data visualisation of performances of tested models

Supervisor meetings

My supervisor, Yuhua Li, has scheduled weekly meetings on Wednesdays. These meetings will be used to share my progress, discuss issues with the code or models, and seek help if the project progress comes to a halt. The meetings will happen every week to ensure an efficient workflow.

Milestones

- **Week 1**: Discuss the project with my supervisor and create the initial report. Start the research of machine learning.
- **Week 2**: Submit the initial plan and start coding with Tensorflow, Keras and Skikit-learn.
- **Week 3-4**: Research and compare approaches to deal with imbalanced data. Evaluate the approach most suitable for the given task and dataset.
- **Week 5-6**: Research supervised learning models and evaluate their strength and weaknesses when using with the given dataset.
- **Week 7-8**: Build researched models using python libraries and test them against the balanced training datasets.
- **Week 9-10**: Test supervised models with different balancing approaches to evaluate the most efficient ones, taking into consideration the trade-off between precision and recall.
- **Week 11-12**: Test the models using AUROC and other testing methods. Experiment with using multiple models together to try to achieve higher accuracy.
- **Week 12-13**: Conduct the final tests of all models and evaluate the most accurate and efficient model. Check that the model follows all modern standards. If time permits, work on the increase of the speed of training the chosen model.

- **Week 14**: Write the report discussing the findings, including detailed results, data visualisations and analysis.

- **Week 15**: Final retouches to the report and code. Debugging and refactoring of the code. Submit the final report.

References

[1] AltexSoft. (2020). Fraud Detection: How Machine Learning Systems Help Reveal Scams in Fintech, Healthcare, and eCommerce. [online] Available at: https://www.altexsoft.com/whitepapers/fraud-detection-how-machine-learning-systems-help-reveal-scams-in-fintech-healthcare-and-ecommerce/ [Accessed 1 Feb. 2020].

[2] Renjith, S. Detection of Fraudulent Sellers in Online Marketplaces using Support Vector Machine Approach. International Journal of Engineering Trends and Technology (2018).

[3] AltexSoft. (2020). Fraud Detection: How Machine Learning Systems Help Reveal Scams in Fintech, Healthcare, and eCommerce. [online] Available at: https://www.altexsoft.com/whitepapers/fraud-detection-how-machine-learning-systems-help-reveal-scams-in-fintech-healthcare-and-ecommerce/ [Accessed 1 Feb. 2020].

[4] Kaggle.com. (2020). IEEE-CIS Fraud Detection | Kaggle. [online] Available at: https://www.kaggle.com/c/ieee-fraud-detection/data [Accessed 30 Jan. 2020].