# 

# **ESOF 0151 - Large Scale Data Analytics**

# **Date Submitted: September 19, 2020.**

# **Project Proposal**

# **Contributing Members:**

# **David Adese, Lucas Dillistone, Kevin Genereux**

**Title:** Detecting Fraudulent Online Transactions using Gradient Boosting Techniques and an Artificial Neural Network

**Executive summary:** The rapid growth of the e-commerce industry has contributed to the exponential increase in online transactions and consequently the surge in fraudulent online activity [1]. Fraudulent online transactions are described as being transactions unauthorized by the cardholder or the intentional misrepresentation of some information [2]. Machine learning algorithms allow us to extract useful patterns from large datasets that are not easily discernible by humans. This project will explore an automated system for detecting fraudulent online transactions using unsupervised learning classifiers: Artificial Neural Network (ANN), AdaBoost, XGBoost, LightGBM, and a Majority Voting Ensemble Classifier.

**Problem definition and motivation:** Identifying fraudulent online transactions is a major issue in the e-commerce industry due to the large volumes of transactional data that require processing. Fraudsters are becoming more adept at exploiting authentication processes and are constantly adapting their behaviours to masquerade as regular customers. As a result, differentiating between fraudulent and genuine transactions becomes a very complex task [3]. An automated fraud detection system is capable of identifying fraudulent transactions with much greater efficiency and accuracy than manual approaches. Large e-commerce companies are investing significant amounts of resources into detecting fraudulent online activity to prevent a decrease in revenues, legal repercussions, and a loss of reputation. [3]

**Proposed solution, methods, and objectives:** The proposed system presents an automated approach for detecting fraudulent online transactions. Five different machine learning algorithms will be developed, namely: Artificial Neural Network (ANN), AdaBoost, XGBoost, LightGBM, and an Ensemble classifier that combines the results of these models. Gradient boosting techniques offer unsupervised learning approaches that detect anomalous patterns from new online transactions. The main objectives of the proposed system are to attain high accuracy, precision, sensitivity, and specificity in detecting fraudulent online transactions. We also aim to develop an efficient system that is capable of quickly detecting fraudulent online transactions without inconveniencing genuine customers.

**Data source**: The dataset has been acquired from Vesta’s real e-commerce transactions and was used for a previous Kaggle competition [4]. The dataset consists of two tables: a transaction table (394 features) and identity table (41 features). The transaction table contains information about the transaction time and date, product code, transaction amount, and the address of the associated user. The identity table contains information about the associated user to each transaction, such as network information, browser, and operating system. Vesta has already partitioned the transaction dataset into training (590,541 rows of data) and testing (506,692 rows of data) files. While most features contain no missing values, some features have up to 96% of their data missing. The dataset consists of mostly numerical data but also has some categorical features. The dataset is highly imbalanced as the positive class (fraudulent transactions) accounts for only 3.5% of the dataset.

**Theoretical and practical contributions**: Many existing online fraudulent detection systems rely on supervised learning approaches to study labelled data and create predictions based on historical data. However, fraudsters are constantly adapting new online behaviours to circumvent detection systems. As such, our project will focus on developing unsupervised learning approaches that acquire information from new online transactions and extract anomalous patterns. Our project will explore how Gradient Boosting techniques and an Artificial Neural Network (ANN) can be applied to achieve highly accurate results. Dimensionality reduction techniques will be used to extract important features, evaluate the importance of each feature and reduce the complexity of our models. A prototype may be developed that will provide a fraud detection analyst with an interactive user interface to upload the dataset and obtain predictions on whether transactions are fraudulent or not.

**Measure of success:** Each algorithm will analyze the training data and infer a predictive model that is capable of predicting whether or not an online transaction is fraudulent. The effectiveness of each machine learning algorithm will be quantified using evaluation metrics of accuracy, precision, specificity, sensitivity, and F1 score. The performance tradeoff between sensitivity and specificity for each algorithm will be visualized using an ROC curve. ROC curves are an effective metric for measuring performance when handling a severe class imbalance, as in this case.

**Project plan:** Missing numerical values in the dataset will be imputed by using mean, median, and mode values for that column. MinMax scaling will be used as a form of feature engineering to ensure that features are given appropriate weights to reflect their importance. We will explore using various dimensionality reduction techniques such as Principal Component Analysis to reduce the complexity of our models and improve their efficiency. The imbalanced dataset will be handled through the SMOTE and ADASYN algorithms to prevent the algorithms from favouring the biased class (non-fraudulent). Four different machine learning models will be developed, namely: Artificial Neural Network (ANN), AdaBoost, XGBoost, and LightGBM. An ensemble classifier will be used to combine the results of these models and deliver final predictions using majority voting. The effectiveness of each model will be computed and compared on the test dataset by using the previously mentioned metrics in the *Measures of Success* section. If time permits, a prototype may be developed to allow analysts to upload a csv file containing online transactions and obtain predictions on the classification of the transaction.

**Deliverables:** Our implementation code will be thoroughly documented and uploaded to a folder on Google Drive for review. A final report will be written throughout the duration of the project that will discuss our research findings. A presentation will be conducted showing our methodology and results. A prototype may be developed that will provide an analyst with a simple user interface to upload a file of transactions and obtain predictions on which of those transactions are fraudulent.

**Application areas:** The proposed systemcould be applied to a wide variety of companies that leverage e-commerce systems and allow customers to make online transactions. In particular, banking systems and electronic payment services, such as PayPal, are developing very sophisticated fraud detection algorithms to adapt to the constantly-evolving patterns of fraudsters. These large companies jeopardize losing their established reputations and billions of dollars in revenue if fraudulent customer activity remains undetected. The proposed fraudulent detection system may flag a customer for suspicious activity so that a data analyst may further investigate and prevent the fraudulent activity from continuing.

**Impact and significance**: According to the 2019 Identity Fraud Study from Javelin Strategy and Research, 14.4 million consumers were victims of identity fraud. In 2018, $24.26 billion was lost due to payment fraud worldwide [5].Globally, automated fraud detection systems are popularly known and implemented because of their ability to help monitor multiple transactions in real time and prevent revenue loss [6]. Automated systems are very robust and scalable which allows companies to quickly adapt to new data that is generated. Businesses can adapt quickly to new fraud techniques by updating their data and applying new rules [7]. Combining real-time transactional data with historical analysis of customer behavior is an effective strategy to identify the anomalous patterns of fraudulent customers.

**References Cited:**

[1] S. D. Kavila, “Machine Learning For Credit Card Fraud Detection System,” *International Journal of Applied Engineering Research*, vol. 13, pp. 16819–16824. ISSN 0973-4562.

[2] “What is Online Fraud & Why Does It Take Occur?,” *Signifyd*. [Online]. Available: <https://www.signifyd.com/resources/fraud-101/why-does-online-fraud-occur/>.

[3] A. Pumsirirat and L. Yan, “Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine,” *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 1, 2018.

[4] IEEE Computational Intelligence Society, “IEEE-CIS Fraud Detection”. Kaggle: IEEE Computational Intelligence Society, 2019. [Dataset]. Available: <https://www.kaggle.com/c/ieee-fraud-detection/data>.

[5] “Facts Statistics: Identity theft and cybercrime,” *Insurance Information Institute*. [Online]. Available: https://www.iii.org/fact-statistic/facts-statistics-identity-theft-and-cybercrime#:~:text=Identity theft and fraud complaints,-The Consumer Sentinel&text=Of the 1.7 million fraud,in these cases was $320.

[6] Patil, Dipti D., et al. "Efficient scalable multi-level classification scheme for credit card fraud detection." IJCSNS International Journal of Computer Science and Network Security 10.8 (2010): 123-130.

[7] “Fraud Prevention and Detection,” *Software AG*. [Online]. Available: https://www.softwareag.com/resources/Fraud-prevention-and-detection.