Third Report - Q&A

November 17, 2024

1 Project Title

Real-time Anomaly Detection in Financial Transactions

Authors and Team - Author 1: Haozhen Guo

- Author 2: Yang Liu

Team name: Black Thym

2 Executive Summary

For this project, Real-Time Anomaly Detection in Financial Transactions, the major objective is to develop and optimize machine learning algorithms for anomaly detection, focusing on fraud detection in the financial market. By improving the performance of the anomaly detection method, we are trying to reduce financial losses for individuals, achieve better jobs in risk management, and enhance anti-money laundering systems.

2.1 Decisions to be impacted

Based on our research, anomaly detection algorithms play a crucial role in shaping business decisions across three key areas:

- **Personal Savings Protection**: By identifying abnormal transactions, banks and other financial institutions can proactively monitor and block suspicious or high-risk activities, thereby safeguarding individual savings and account security.
- **Risk Management**: Anomaly detection enables financial institutions to recognize unusual patterns within portfolio management, allowing for timely adjustments to investment strategies, and ultimately enhancing risk management and financial performance.
- **Anti-Money Laundering (AML)**: A major application of anomaly detection is in identifying irregular transaction behaviors that may indicate potential money laundering activities, helping institutions comply with regulations and prevent financial crime.

2.2 Business Value

Our research holds societal relevance in two primary areas:

- **Protection of Vulnerable Populations**: Across all of the people who suffer from financial fraud, vulnerable groups such as the elderly, low-income families, and individuals withliabilities are most likely to become the target of fraud, while at the same time, they do nothave enough ability to contend with it. Once the anomalies related to financial fraud becomeundetected, these populations are exposed to considerable financial risks. By developing advanced and effective detection methods, we wish to contribute more to preventing vulnerableindividuals from fraud and the risk of loss of personal property.

- **Enhancing Trust in the Digital Age**: As financial systems become more digitized, publicattitudes toward automated systems are becoming increasingly polarized and extreme. Somepeople fully trust algorithms without a comprehensive understanding of them, while othersbelieve nothing but rely solely on personal judgment. As both views are taken to extremes,the development of the financial market will face significant challenges. By establishing arobust anomaly detection system, we want to help restore and strengthen public trust in financial institutions.

2.3 Data Assets

Our fraud detection dataset was collected from Kaggle and provided from Vesta Corporation, a leader in e-commerce payment solutions. The dataset was split into two files: “train_identity.csv” and “train_transaction.csv”, both of which can be joined through the common and unique key TransactionID. Our goal is to establish machine learning methods to identify fraudulent transactions, which was labeled as isFraud in “train_transaction.csv”, using a wide range of features. Here are some detailed information about each dataset:

“train_transaction.csv”: Contains the majority of data related to transactions.

- TransactionDT: timedelta from a given reference datetime (not an actual timestamp)
- TransactionAMT: transaction payment amount in USD
- ProductCD: product code, the product for each transaction
- card1 - card6: payment card information, such as card type, card category, issue bank, country, etc.
- addr: address
- dist: distance
- P_ and (R___) emaildomain: purchaser and recipient email domain
- C1-C14: counting, such as how many addresses are found to be associated with the payment card, etc. The actual meaning is masked.
- D1-D15: timedelta, such as days between previous transaction, etc.
- M1-M9: match, such as names on card and address, etc.
- Vxxx: Vesta engineered rich features, including ranking, counting, and other entity relations.

“train_identity.csv”: Contains the identity information variables associated with transactions, such as network connection information (IP, ISP, etc) and digital signature (UA/os, etc). However, the field names were masked and pairwise dictionaries would not be provided for privacy protection and contract agreement.

For this project, we will benchmark machine learning models on a large-scale dataset about real-world e-commerce financial transactions. We wish to build up a better method to detect the fraud in transactions, and further achieve saving personal assets with higher accuracy and efficiency.

3 Questions and Answers

3.1 Xudong He

- Q:
  - In the process of anomaly detection, to address challenges such as data imbalance, feature collinearity, and missing values, as well as considering the potential impact of error types related to anomalies on financial systems, are there more suitable detection methods?
- A:
  - For this project, our goal is to investigate potential machine learning algorithms to address the problem effectively. Outlier detection using unsupervised learning is one of the possible approaches informed by our review of prior research. The workflow includes critical steps such as feature selection, data preprocessing, model training, and evaluation, all of which are essential to developing a machine learning model with strong performance. However, we believe that financial institutions with access to larger datasets and advanced resources might adopt alternative methods that enable more efficient fraud detection.

3.2 Bingyue Hu

- Q:
  - How do target encoding and one-hot encoding perform across different types of models?
- A:
  - For our project, we decided to explore the anomaly detection problem through both supervised and unsupervised learning methods. We believe that datasets with supervised encodings, such as target encoding, might perform better with supervised learning algorithms, while unsupervised encodings, such as one-hot encoding, might yield better results with unsupervised learning algorithms.

3.3 Charles Hang

- Q:
  - Do you all have plans to consider any other outlier detection methods? We found in our literature review that gradient boosting methods generally perform stronger than isolation forest.
- A:
  - Yes of course. In addition to unsupervised learning methods, for instance the Isolation Forest which was introduced in the presentation, we will also try some supervised learning methods for this question in the future. SVM with Gradient Descent Boosting and Random Forest have already been implemented in our project as well and we are looking forward to seeing which model performs best for this question.

3.4 Fatih Sogukpinar

- Q:
  - How can we differentiate outliers and actual fraud detections in this case? It feels like there could be a lot of overlap between them.

- A:
  - In our case, with unsupervised learning, identifying the actuarial fraud as outliers is considered as the ultimate goal of model training rather than merely a preprocessing step in other scenarios. Thus, we decided not to implement the multivariable outlier detection to filter the dataset before we started to detect the fraud.
  - However, we did try some single-variable outlier detection methods to handle the potential extreme values in specific columns. Using the IQR method returned unreasonably high proportions of outliers, which in most cases over 20% of the dataset. On the other hand, although the Z-score method did return more reasonable proportions of outliers, the data set after outlier dropping would contain a higher percentage of normal transactions, making it aparser for training. Consequently, we determined not to implement any outlier detection before training our models.

3.5 Kexin Zhang

- Q:
  - Have you tried any other model other than the isolation tree?
- A:
  - Yes of course. In addition to unsupervised learning methods, for instance the Isolation Forest which was introduced in the presentation, we will also try some supervised learning methods for this question in the future. SVM with Gradient Descent Boosting and Random Forest have already been implemented in our project as well and we are looking forward to seeing which model performs best for this question.

3.6 Naveen Asokan

- Q:
  - Could we conclude that outliers are basically fraud transactions?
- A:
  - Not exactly. Ideally, we wish that our unsupervised model can perfectly identify the fraud transactions as outliers from the original dataset. However, we observed that the model will also classify some normal transactions as outliers, resulting in false positives.
  - The result suggests that some fraud transactions are not statistically or computationally distinct from normal transactions, presenting a key challenge for further model optimization.

3.7 Miken Guo

- Q:
  - In your univariate outlier detection part, what caused these outliers, like mismeasurement or input error?
- A:
  - For both IQR and Z-score outlier detection methods, the outliers are determined by the models based on statistical significance. The result is highly dependent on the distribution of the original features. Our task focused more on analyzing whether dropping these outliers is reasonable.

3.8 David Xiong

- Q:
  - How do you detect outliers and will you consider eliminating the outliers? How did you apply the Isolation Forest and why did you pick it?
- A:
  - Let us answer your question in three parts.
  - Part 1: As mentioned in the presentation, we tried some single-variable outlier detection methods to handle the potential extreme values in specific columns. Using the IQR method returned unreasonably high proportions of outliers, which in most cases over 20% of the dataset. On the other hand, although the Z-score method did return more reasonable proportions of outliers, the data set after outlier dropping would contain a higher percentage of normal transactions, making it aparser for training. Consequently, we determined not to implement any outlier detection before training our models.
  - Part 2: We applied the Isolation Forest using the sklearn.ensemble packages. In detail, for this particular question, since we already know the proportion of outliers (fraud transactions) of the original dataset, we specifically set the hyperparameter, “contamination”, equals to 0.08 for our model training.
  - Part 3: We selected the Isolation Forest for model training for two reasons. First, Isolation Forest is regarded as a representative depth-based unsupervised learning for outlier detection, with robust performance through ensembling. Second, Isolation Forest is a tree-base model, which performs effectively on high dimensional dataset and is computationally efficient compared with OneSVM and DBSCAN. It is always worth a try.

3.9 Praneel Panchigar

- Q:
  - Why are unsupervised models more viable for this problem? Does that address the imbalance in the dataset?
- A:
  - Unsupervised models are more suitable for this problem for two reasons. First, unsupervised models are more suitable for an imbalanced data set, for our case, 92% of normal transitions vs. 8% of fraud, without requiring additional data engineering. Second, according to our research, fraud transactions are assumed to be more “extreme” compared with normal transitions in some ways. Therefore, we believe that unsupervised learning models might sensitively capture the abnormality from the data set and turn a better performance in the end.
  - Nevertheless, we plan to test supervised models for validation and comparison.

3.10 Junyuan Yao

- Q:
  - Did you use only recall to evaluate your model? With a high recall, it is also possible that the model is trying to predict all values as positive and this could also lead to a high recall, though this performance is not satisfying.
- A:
  - The evaluation metric is an important part to successfully evaluate the model performance. With our imbalance data set, we believe that the “Accuracy” is a bad model

evaluation metric. Since we care more about fraud transactions than the normal transactions, we trained the model using "Recall" as the primary evaluation metric at the beginning. However, as you state, a high recall paired with extremely low precision is also not satisfying enough for model training. In reality, there is always a tradeoff between identifying more fraud (high recall) and reducing the fraud-checking workload (high precision). For our model, we plan to explore other evaluation methods, such as the AUC-Precision-Recall curve and F-beta score, to find out a better evaluation approach.

3.11 Han Jiang

- Q:
  - Why do you use Isolation Forest?
- A:
  - We selected the Isolation Forest for model training for two reasons. First, Isolation Forest is regarded as a representative depth-based unsupervised learning for outlier detection, with robust performance through ensembling. Second, Isolation Forest is a tree-base model, which performs effectively on high dimensional dataset and is computationally efficient compared with OneSVM and DBSCAN. It is always worth a try.

3.12 Gino Kler (They)

- Q:
  - How are you adjusting to account for imbalance in your data?
- A:
  - For unsupervised models, we assumed that fraud transactions should be treated as outliers. In this way, we do not need to do a specific process for balancing the data set. However, for the model evaluation metric, we focus a lot on correctly detecting as many fraud transactions as possible, rather than overall "Accuracy". Hence, we evaluate our models with different evaluation metrics, such as Recall, AUC-Precision-Recall curve and F-beta score.
  - For supervised models, we will try different ways to balance the data set.
    - Oversampling or downsampling. Although this method is widely used in balancing data sets, we are worried about introducing additional bias through changes in the training data set.
    - Robust models: We plan to implement some algorithms which are naturally robust to imbalance datasets, such as Random Forest and XGBoost/LightGBM.
    - Alternative metrics: Similar to unsupervised models, we will explore metrics beyond accuracy, including AUC-PR and F-beta scores.

3.13 Wenjing Wang

- Q:
  - What would be the potential impact of false positives or false negatives in your model's predictions
- A:
  - False positives occur when normal transactions are wrongly classified by the model as outliers. The small amount of false positives are tolerated for this case since it just increases the workload for checking the fraud alerts.

  - False negatives, on the other hand, are undetected transactions. Unlike false positives, false negatives are far riskier since all fraud transactions can possibly result in substantial financial losses. Our model is designed for minimizing false negatives to capture as many fraudulent transactions as possible.

3.14 Shanke Wang

- Q:
  - Are there any methods you are considering to deal with the imbalanced dataset?
- A:
  - For unsupervised models, we assumed that fraud transactions should be treated as outliers. In this way, we do not need to do a specific process for balancing the data set. However, for the model evaluation metric, we focus a lot on correctly detecting as many fraud transactions as possible, rather than overall "Accuracy". Hence, we evaluate our models with different evaluation metrics, such as Recall, AUC-Precision-Recall curve and F-beta score.
  - For supervised models, we will try different ways to balance the data set.
    - Oversampling or downsampling. Although this method is widely used in balancing data sets, we are worried about introducing additional bias through changes in the training data set.
    - Robust models: We plan to implement some algorithms which are naturally robust to imbalance datasets, such as Random Forest and XGBoost/LightGBM.
    - Alternative metrics: Similar to unsupervised models, we will explore metrics beyond accuracy, including AUC-PR and F-beta scores.

3.15 Ferris Atassi

- Q:
  - Does your group think it would benefit from introducing some form of feature engineering? Our group is working on the same project and noticed a performance increase in models upon introducing new features created from already correlated fields in the dataset.
- A:
  - We haven't considered introducing some new features for our model training to avoid introducing personal bias in the dataset. However, we performed PCA for feature reductions after analyzing the feature correlations and it helped to improve the model performance in most cases.

3.16 Julia Tompkins

- Q:
  - It looks like you did a lot of work with feature reduction. After your feature reduction, did the features that remained seem reasonable? What types of features were they, and were they interpretable?
- A:
  - We believe that the features after the feature selection process are reasonable for model training. Our feature selection involves two major steps, dropping columns with two many null values and dropping the column with a single dominant value. We are confi-

dent that this process will not introduce personal bias into the dataset, making it reliable for further model training.
  - The remaining features include both numerical and categorical variables, related to the key information and identity through the transaction. While most categorical features are interpretable, the numerical features are less interpretable, as their exact meanings are anonymized in the original data set for privacy and security reasons. Although we can infer the general type of data these columns represent, the detailed information is unavailable in most cases.

3.17 Naveen Asokan

- Q:
  - What are common real-world constraints (e.g., data privacy, regulatory requirements) when implementing anomaly detection in financial systems?
- A:
  - There are three common real-world constraints for anomaly detection in real world financial systems.
    - ∗ Data privacy and security: Generally, trading-related data is not available to the public. However, I believe that most of the reputable financial institutions will publish documents in advance for their customers to review and sign, ensuring that most data is accessible and shareable in their internal systems.
    - ∗ Tradeoff of model performance: In our model, as you may have noticed, increasing the number of predicted outliers can help capture more fraudulent transactions but simultaneously decreases precision. In reality, there is always a tradeoff between identifying more fraud (high recall) and reducing the fraud-checking workload (high precision).
    - ∗ Anti-encryption for fraud transitions: In reality, fraud transactions are often closely associated with money laundering, which involves advanced data encryption. This encryption makes such transactions less suspicious and significantly harder to detect. Consequently, financial institutions often need to invest extra effort in decrypting these transactions before model training can be effective.

3.18 Anagha Mayasandra Vinaya Simha

- Q:
  - Does the time of the day or day of the week have any impact on anomaly detection?
- A:
  - That is a good approach to think about the anomaly detection in financial markets. Although the “TransactionDT” column is indeed included in our data set, due to data security reasons, the original transaction timestamps have been deliberately erased, leaving only sequential numbers to represent the order of transactions. Therefore, we cannot investigate this aspect with the current dataset.

3.19 Dylan Mack

- Q:
  - Why did you use PCA as opposed to picking a singular variable? If they are so extremely correlated, do you think it makes a significant difference to do dimensionality reduction?
- A:

    - I believe your question relates to our process of implementing the PCA before filling missing values with KNN. The key insight to use PCA here is to perform dimension reduction without introducing personal bias, even though the selected columns are extremely correlated. While the process may seem fussy, it ensures the unbiased selection for dimension reduction without consuming excessive time.

3.20 Zihang Shi

- Q:
    - It looks like your dataset has a lot of nan values, why is your method dealing with missing values reasonable?
- A:
    - We addressed the NaN values in two main steps:
        - Dropping columns with excessive missing values: After analyzing the proportion of missing values, we dropped all columns where more than 20% of the data was missing. This approach is reasonable since approximately 200 features remained for model training after this step.
        - Filling missing values using KNN: Instead of simply filling missing values with the mode (for classification) or mean (for regression), we used KNN classification/regression for imputation. We believe this method provides more accurate and reasonable results for filling in missing values.

[ ]: