Illicit transactions detection in the Bitcoin blockchain using supervised and unsupervised learning techniques

Pat Kosakanchit
pathik@uw.edu

Lakshmi Venkatasubramanian
lvenkat@uw.edu

Megan Morrison
mmtree@uw.edu

Abstract

Bitcoin has become a preferred method for transferring money for illicit economic activity such as ransomware payments and money laundering. Robust detection of these illicit transactions in the Bitcoin blockchain is crucial for anti-money laundering operations. Major challenges for illicit transaction detection include the lack of transaction labels, the massive size of the bitcoin network, and the fact the bad actors work to mask their activity and avoid detection. Previous researchers have proposed anomaly detection methods, supervised learning, and active learning to classify illicit transactions. Supervised learning is necessary to classify illicit transactions as illicit transactions do not correlate with anomalies in the dataset. Unsupervised learning methods are useful when there's scarcity of labels. Active learning is a useful modification of supervised learning for illicit transaction classification as real world bitcoin datasets have very few licit/illicit labels. We compare the performance of these methods using the Elliptic dataset.

Preprint. Under review.

1 Introduction

Bitcoin has become a popular means of transferring money for illicit economic activity due to its pseudo-anonymity for users and difficulty to regulate [3, 4]. Detecting illicit transactions in the Bitcoin network is an active area of research and requires machine learning methods that can classify data with few labels, unbalanced groups, and potentially adversarial behavior by the bad actors attempting to avoid detection. We compare the illicit transaction detection methods employed be previous researchers and find that supervised learning methods are superior to anomaly detection methods. We employ a modified version of supervised learning, called Active Learning (AL), to classify transactions with the minimum amount of labeled datapoints, reproducing the results found in previous work [5]. Active learning is a useful supervised learning variant for illicit transaction detection as most transactions in the Bitcoin blockchain are unlabeled. We propose a modification to the AL method that uses network edges to select transactions that are connected to illicit transactions to label. This method may be useful as illicit transactions are infrequent in the blockchain but transactions with edge connections to illicit transaction may be more likely to be illicit, improving our selection process for data points to label and hopefully improving the performance of the resulting model.

Detecting illicit transactions is crucial for anti-money laundering operations and comes with major challenges. The Bitcoin blockchain is massive in size (400 GB) and contains a larger number of users (100 M) requiring big data techniques to find patterns in the data efficiently and leverage the large number of features associated with each transaction. Although the Bitcoin blockchain is publicly available, until recently, very few labels existed as to whether transactions were licit versus illicit. Because of this, researchers relied primarily on unsupervised learning methods, predominantly anomaly detection, to find suspicious transactions [7, 6]. Datasets with a much larger quantity of licit/illicit labels have recently been curated which has allowed for substantial improvements in illicit activity detection and improved quality in evaluation metrics for unsupervised methods [11, 9, 5]. In practice, however, labels for transactions are difficult to obtain. Illicit transaction detection methods that do not require large amounts of labeled data are necessary in order to be useful for anti-money laundering operations [5]. Another major challenge in illicit activity detection is that criminals often work to disguise their activity as normal in order to avoid detection [5, 2]. This may lower the performance of anomaly detection methods or lower the performance of supervised learning methods over time as criminals modify their transactions to avoid having them classified as illicit.

2 Related Work

2.1 Anomaly detection

Initial illicit transaction detection methods made the assumption that illicit transactions were rare and would display atypical features compared to the majority of transactions in the network. With this assumption, researchers explored anomaly detection techniques to find outliers in the data that would, in theory, correspond to suspicious activity [7, 6]. Pham and Lee [6] apply K-means clustering to bitcoin transaction data followed by Multivariate Gaussian distribution anomaly detection using Mahalanobis distance metric and Unsupervised SVM to detect anomalies users and transactions that in theory correspond to suspicious activity. In Pham and Lee's Mahalanobis distance based method [6], they first cluster all bitcoin transactions using k-means clustering on network features. They reason that transactions involving illegal activity will have features which are outliers to normal, legal transactions. Since such transactions are rare compared to legal transactions, they conclude that anomaly detection methods are well suited to find suspicious transactions and to this end aim to find outliers to the k-means clusters. While such techniques were able to reliably find outliers in the data, it was uncertain whether such outliers were, in fact, illicit transactions due to a lack of labels.

2.2 Supervised learning

With the recent introduction of labeled licit/illicit transactions in the Elliptic dataset [11], researchers were able to employ supervised learning techniques to classify transactions as well as evaluate the reliability of anomaly detection methods. Weber *et al.* [9] used to Elliptic dataset to classify transactions as either "licit" or "illicit" using binary classification supervised learning methods. Four machine learning models were applied: Logistic Regression, Random Forest, Multilayer Perceptrons,

and Graph Convolutional Networks. Random Forest was the best model, achieving the highest F_1 score of 0.796 despite the fact that Graph Convolutional Networks capture relational information among transactions instead of simply using individual transaction features for classification like the rest of the models. These techniques are typically applied to datasets with large samples in each category and meet with greater success under these conditions. We hypothesize that the supervised learning performance can be increased by balancing the datasets. Their models performance also varied over time. The classification performance, while initially high, dropped precipitously after the dark market shutdown, indicating that the model is not reliable after unexpected events or changes to the network. Further research must be done to obtain a more robust model.

Another limitation that occurs with illicit transaction detection is the scarsity of labels in typical Bitcoin transaction datasets in comparison to the labels found in [9]. Researchers have addressed this by modifying supervised learning algorithms to identify illicit transactions with a much smaller sample of labels [5]. Using Active Learning (AL), Lorenz *et al.* [5] were able to obtain the same performance as standard supervised learning methods while using only 5% of the labels. While this method is able to successfully classify transactions, it also has the advantage of working with few illicit labels which is more typical in real world networks.

Active Learning is implemented by iteratively querying subsets of the data for labels and then re-training the model with the updated set of labels [5]. The model is incrementally improved with a larger and larger set of labels until a satisfactory level or performance is reached. An important step in this process is the querying strategy which, instead of randomly selecting data points to label, strategically chooses points that are likely to have the most impact on the model. Lorenz *et al.* [5] uses two supervised querying strategies — uncertainty sampling and expected model change — as well as two unsupervised querying strategies — elliptic envelope and isolation forest. They find that the supervised learning querying strategies outperform the unsupervised learning strategies. They reason that the unsupervised learning strategies perform poorly due to these strategies selecting outliers for classification which do not correlate well with illicit transactions. We will test the AL method for illicit transaction detection using the same querying strategies as [5] and compare our results to the supervised methods trained on the entire test set.

3 Data Collection

We use the Elliptic Data Set which contains 200K Bitcoin transactions, a subset of which are tagged as belonging to either licit or illicit categories [9, 11]. This dataset is a time series graph where Bitcoin transactions are nodes in the network and directed payment flows are edges. Along with licit or illicit tags, the nodes are also tagged with 166 additional node features. 94 of the features are local information about the transaction, such as the time step, transaction fee, and output volume. 72 features are aggregated features which are aggregated information from the nodes one-hop backward and forward. The data is divided evenly into 49 time steps. In each time step, the transactions are connected as a single component. There is an important event occurs at time step 43 which is a sudden closure of a dark market. This causes the number of illicit data to decrease significantly after the shut down.

We use this dataset to evaluate the effectiveness of anomaly detection methods to classify illicit transactions, reproduce supervised learning classification method employed in previous studies, including AL, and evaluate the performance of a modification to the AL method that uses the network graph to aid in selecting datapoints.

4 Active Learning

Active learning is a supervised learning method that is ideal to use when labels are difficult to procure, as is the case with the Bitcoin blockchain. Although the Elliptic dataset has thousands of labels, this quantity of labels typically does not exist for most Bitcoin datasets. We apply active learning to the bitcoin dataset and compare our models performance with the performance of a supervised learning model using all 30000 training labels as well as models trained with similar quantities of data where points are instead randomly selected.

Active learning consists of two stages that are applied iteratively — the supervised learning step and the querying step [5, 8]. In the supervised learning step we use logistic regression (LR) to train

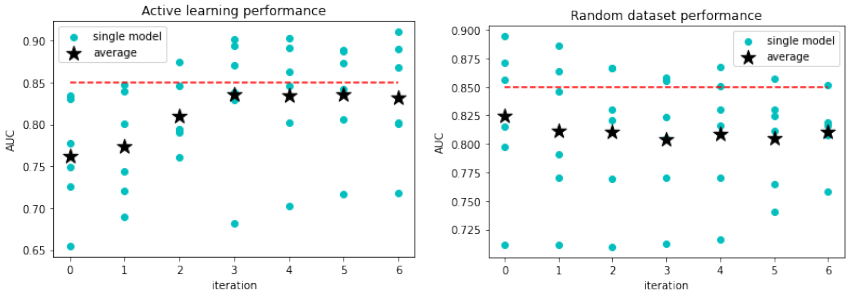

Figure 1: Left: Performance of active learning model as more data is added to the training dataset at each iteration. Red dashed line is performance using entire training dataset. Right: Performance of model trained with random data using the same dataset size as the active learning condition at each iteration.

a classifier based on the labeled pool of data. In the querying step we update the labeled pool by selecting a set of unlabeled data to label and add to the labeled pool. We use uncertainty sampling, a successful querying method employed by [5] to select data to label. In uncertainty sampling, unlabeled points that have the highest uncertainty, that is the closest classification to 0.5 for a binary classification task, are chosen. We use an initial dataset of about 39 randomly selected datapoints and then select 10 additional datapoint to add to the training data at each querying step based on the querying strategy.

Figure 1 shows the active learning performance compared to the performance of models trained on similarly sized training datasets where additional datapoints are selected randomly instead of based on a querying criteria. The average active learning performance increases over the initial iterations and approaches the performance of the model trained on the entire set of 30000 labeled data points. This is promising for practical applications because it shows that a similar performance can be achieved with a strategically selected sample of 100 datapoints as the performance using thousands of datapoints. Because labels can be difficult to obtain, using this method to select new data to label could be useful.

Illicit transactions are far less frequent in the dataset, resulting in an imbalanced dataset when training on the entire training dataset as well as randomly chosen data (Fig. 2). Active learning, however, oversamples illicit transactions, resulting in a more balance dataset when active learning is used to choose data to label and perhaps contributing to a higher model performance as a result of balancing the training data (Fig. 2).

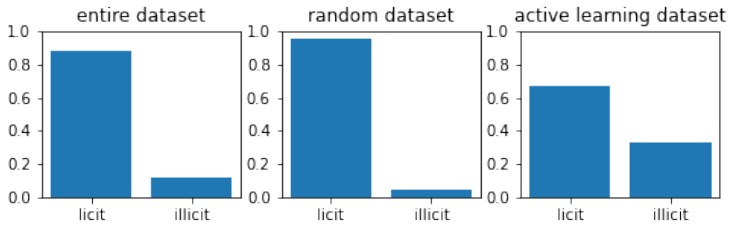

Figure 2: Distribution of training data using different data selection methods.

5 Supervised Learning

5.1 Traditional supervised learning

Our first attempt is to use supervised learning methods to classify illicit transactions. We apply logistic regression, random forests, and XGBoost to a training set selected from the labeled Elliptic data. We choose the training data to be labeled transactions from the time points 1-34 and the test data to be the labeled transactions from the time points 35-49. Figure 3 shows the F1 scores from the test data. All three methods perform well during the initial time points with random forests performing

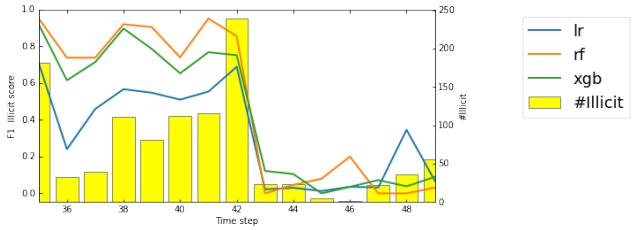

Figure 3: F1 scores for illicit transaction detection using supervised learning techniques: logistic regression (blue), random forests (orange) and XGBoost (green). The number of illicit transactions at each time point is shown in yellow.

the best. The scores drop precipitously after the Silk Road dark market shutdown at time point 43, indicating that as the Bitcoin network undergoes substantial changes, models trained on data collected too far in the past may loose their validity. Our results are consistent with those found in previous studies [9, 5].

5.2 Handling Imbalanced Data

The Elliptic Dataset contains 203,769 node transactions and 234,355 edge payments. However, only two percent of the nodes are labelled illicit, and twenty-one percent are labelled licit. This imbalance between the majority and the minority class led us to the question whether the performance of the models could be improved if the dataset is balanced. Therefore, we have explored techniques to handle imbalanced data.

To oversample the minority class, we applied SMOTE (Synthetic Minority Oversampling Technique) which creates extra training data by generating synthetic examples [1], and SMOTE+ENN (Edited Nearest Neighbors) to remove noisy example and clean the data [10]. Then, the balanced data is used for training logistic regression, random forest and XGBoost models.

Table 1: Evaluation Metrics of the supervised learning models

Method	Accuracy	F1-score	Precision	Recall
Logistic Regression	0.89	0.45	0.33	0.71
Random Forest	0.97	0.78	0.85	0.72
XGBoost	0.93	0.61	0.51	0.76
Logistric Regression (SMOTE)	0.96	0.75	0.79	0.71
Random Forest (SMOTE)	0.94	0.61	0.55	0.68
XGBoost (SMOTE)	0.77	0.33	0.21	0.86
Logistic Regression (SMOTE+ENN)	0.95	0.65	0.60	0.74
Random Forest (SMOTE+ENN)	0.90	0.43	0.35	0.56
XGBoost (SMOTE+ENN)	0.75	0.31	0.19	0.87

Table 1 shows the F1 score of all the models. Among all models trained with SMOTE and and SMOTE+ENN, logistic regression trained with SMOTE has the best F1 score of 0.75, which is close to the best F-1 score of 0.78 by the random forest model trained with the original imbalanced data.

Figure 4 compares the F1 scores at each time step of the baseline supervised learning models and two models trained with SMOTE data: logistic regression and XGBoost. After time step 42, the models that are trained with SMOTE and SMOTE+ENN data perform better in most of the time steps. We believe that the synthesized data generated by SMOTE and SMOTE+ENN help reduce the overfitting.

5.3 New Train-Test Split

Though oversampling techniques help increase F1 score after the market shutdown, the result is inferior compared to the time steps prior to the shutdown. Therefore, we started to question: given we have some data points after the shutdown as training data, would the model has a better performance? So we ran another experiment where the training data is not solely from the time step 1 - 34 but from all time steps. The split between training dataset and test dataset is 70:30.

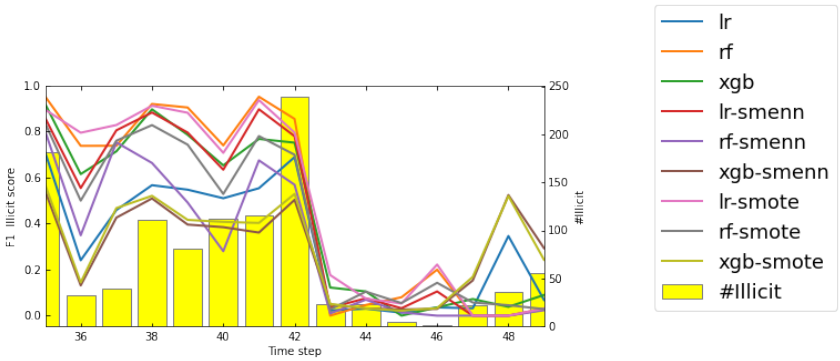

Figure 4: F1 scores for illicit transaction detection using the baseline models and the models trained with SMOTE and SMOTE+ENN. The number of illicit transactions at each time point is shown in yellow.

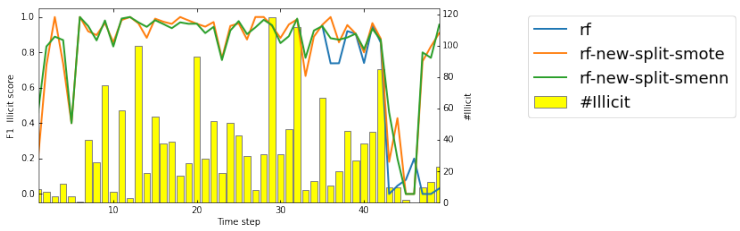

Figure 5: F1 scores for illicit transaction detection using random forest (blue), random forest with SMOTE (orange), and random forest with SMOTE+ENN (green). The models are trained using the training data from every time step. The number of illicit transactions in the test data at each time point is shown in yellow.

Figure 5 shows the F1 scores of three of the best models trained using the new train and test dataset split. The models are the baseline random forest, random forest trained with SMOTE, and random forest trained with SMOTE+ENN. Random Forest trained with SMOTE has the highest overall f1 score: 0.93. At time step where the number of illicit transactions are low, especially time step 45 and 46, the model did not perform well. But at time step 44, 47, 48, and 49, the model has a much better performance compared to the baseline models which are trained using only the data from the time step 1-34 in Figure 3. This illustrates that F1 score would improve if there are more data points from the period where there are unusual market activities.

6 Unsupervised Learning

6.1 Outliers

Using the licit/illicit labels provided in the Elliptic dataset [11], we evaluate the extent to which illicit transactions correspond to outlier values in the transaction features. We find that outlier transactions do not correspond to illicit transactions and that the illicit transactions are distributed within the feature clusters, making anomaly detection not the best method for classifying illicit activity. Figure 6 shows K-means clusters in PCA space with outliers to the clusters colored blue. The true transaction anomalies (illicit transactions), colored in green, fall within the second cluster and do not correspond to outliers in the K-means clusters. Our results agree with the findings of other studies that anomalies in the Bitcoin blockchain in the Elliptic dataset do not correspond to illicit transactions [5].

Figure 7 shows the training and test confusion matrices for K-means in PCA space with all features included and it can be seen that no illicit transactions are predicted correctly.

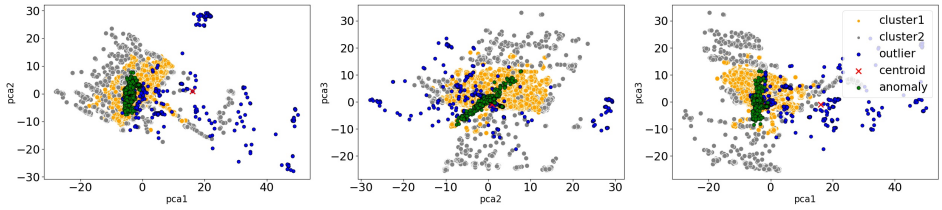

Figure 6: Transactions in PCA space after K-means clustering and outlier detection applied on all features. Outliers to the k-means cluster (blue) do no correspond to the true anomalies/illicit transactions (green).

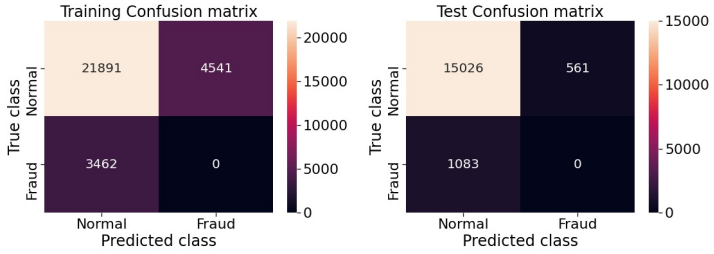

Figure 7: Confusion matrices for K-means with PCA including all the 165 features

6.2 Feature Extraction

We manually identified a subset of features from the PCA distribution that are indicative of anomalies. These 17 features have deviant mean and variance for anomalous transactions as compared to normal transactions. Then applied K-means in PCA space again to see if there's any improvement. We observed that the algorithm is now able to predict a few anomalies correctly. Figure 8 shows K-means clusters in PCA space with the subset of 17 features identifying some of the fraudulent transactions correctly.

6.3 Kernel PCA

Kernel PCA is a non-linear form of PCA and is useful if the fraud transactions are not linearly separable from the non-fraud transactions. Dimensionality reduction algorithms reduce the dimensionality of data while attempting to minimize the reconstruction error. However, these algorithms cannot capture all the information of the original features as they move to a lower dimensional space; therefore, there will be some error as these algorithms reconstruct the reduced feature set back to the original number of dimensions.

In the context of our Bitcoin transactions dataset, the algorithms will have the largest reconstruction error on those transactions that are most anomalous. The reconstruction error for each transaction is the sum of the squared differences between the original feature matrix and the reconstructed matrix using the dimensionality reduction algorithm. We scale the sum of the squared differences by the max-min range of the sum of the squared differences for the training dataset, so that all the reconstruction errors are within a zero to one range. The anomalous transactions that have the largest sum of squared differences will have an error close to one, while the normal transactions that have the smallest sum of squared differences will have an error close to zero.

Using Kernel PCA, we calculated the reconstruction error on the test data for each of these 16670 transactions. If we sort these transactions by highest reconstruction error (also referred to as anomaly score) in descending order and extract the top 1083 transactions from the list, we can see that 82 of these transactions are fraudulent.

The downside of Kernel PCA is that it cannot be trained on large datasets as it consumes extremely large memory. We could only train on a subset of data points from training set. Figure 9 shows that

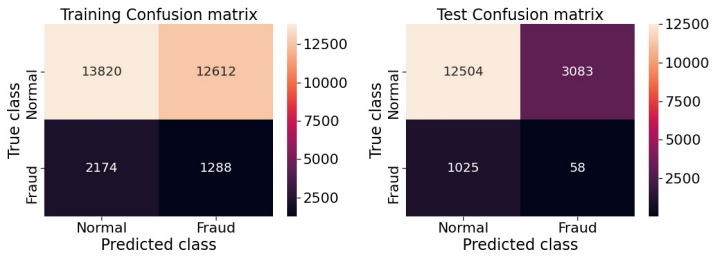

Figure 8: Confusion matrices for K-means with PCA for the subset of 17 features

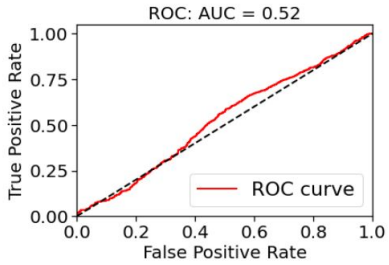

Figure 9: Area Under the Curve from ROC graph for Kernel PCA on test data

ROC curve area under the curve is relatively small on test data. So, we continue to explore other unsupervised learning methods in the next section using Iterative Modelling.

6.4 Iterative Modelling

In iterative modelling, we randomly sample a fixed size sample from training data in each iteration, balance the classes by oversampling minority class using SMOTE method and then apply 5 unsupervised methods namely Local Factor Outlier(LOF), Isolation Forest(IF), Extended Isolation Forest(EIF), DBSCAN and One Class SVM(OCSVM). Although we will not use the fraud labels to build the unsupervised fraud detection solutions, we will use the labels to evaluate the unsupervised solutions we develop. The labels will help us understand just how well these solutions are at catching known patterns of fraud.

LOF starts by computing the distance of each instance to its k nearest-neighbour. LOF uses the distance to compute the instance's density, and if the density is substantially lower than the average density of its k nearest-neighbours, the instance is declared anomalous.

IF isolates anomalies by performing recursive random splits on attribute values. Based on the resulting tree structure, anomalies are instances that are easy to isolate, i.e., have shorter paths. Decision boundaries in traditional isolation forests are one-dimensional (either vertical or horizontal), parallel to the axes. There are regions that contain many branch cuts and only a single or few observations, which may mask some anomalies as normal

EIF mitigates the decision boundary issue by using hyperplanes for splitting the data with random slopes. Hyperplanes have dimensions one less than the dimension of the data. Variance between scores tends to be lower with extended isolation forests compared to traditional ones, making it easier to detect anomalies.

DBSCAN is a clustering algorithm (an alternative to K-Means) that clusters points together and identifies any points not belonging to a cluster as outliers. It's similar to K-means except that the number of clusters need not be specified in advance.

Lastly, OCSVM defines anomalies as observations that deviate from normal behaviour. It detects anomalous instances that lay outside of the decision boundary (OCSVM) learned around them.

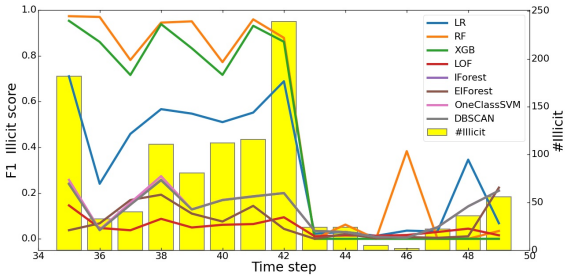

Figure 10: F1 Illicit Score per Time step for Supervised vs Unsupervised methods

6.5 Evaluation

The fraud labels and the evaluation metrics will help us assess just how good the unsupervised fraud detection systems are at catching known patterns of fraud that we have caught in the past and have labels for. However, we will not be able to assess how good the unsupervised fraud detection systems are at catching unknown patterns of fraud.

After 5 iterations, the evaluation metrics for unsupervised learning methods on test data computed as average of all the 5 iteration values are provided in the Table 2. Also, Figure 9 shows that unsupervised methods perform worse as compared to supervised top performing methods with respect to F1 Illicit scores. But around 49^{th} step, F1 scores are relatively higher for a few unsupervised methods when compared to supervised counterparts.

Table 2: Evalution Metrics using Iterative Modelling for 5 unsupervised methods

Method	Accuracy	Precision	Recall	F1-score	ROC
LOF	0.82	0.04	0.08	0.06	0.50
IF	0.07	0.07	1.0	0.12	0.62
EIF	0.82	0.04	0.08	0.05	0.49
DBSCAN	0.06	0.06	1.0	0.12	0.45
OCSVM	0.09	0.07	0.99	0.12	0.54

7 Future Work

If we have more time, we would like to explore semi-supervised learning methods more extensively to take advantage of the unlabeled data points. Label propagation is one of the semi-supervised learning method that we have tried but did not get a good result. The method is computationally expensive, but we could potentially use fewer data points by using those chosen by active learning method.

8 Conclusion

There are various techniques that can be used to detect illicit transactions in the Bitcoin blockchain each with their advantages and disadvantages depending on the amount of labeled data available. Active learning allows a supervised learning model to be trained with very little data and naturally balances the imbalanced dataset due to its tendency to oversample illicit transactions. For supervised learning models, we found that oversampling methods like SMOTE and SMOTE+ENN help improve the F1 score after the market shutdown but only to a certain level. A diverse set of data points from usual and unusual market time would be useful for the training of a more accurate model.

Overall, unsupervised methods performed worse than supervised methods. This could be partially attributed to feature anonymity, feature extraction done manually from PCA distribution, and the features indicating anomalies may not be completely accurate, and the fact that illicit cases are indeed not outliers which we would normally observe in some of the synthetically generated datasets. With better feature extraction methods and hyperparameter tuning, there is a possibility that the unsupervised methods scores could improve.

Individual contributions

Megan Morrison: Implemented active learning algorithm. Worked on writing report.

Pat Kosakanchit: Implemented oversampling techniques, new train-test split, label propagation. Worked on writing report.

Lakshmi Venkatasubramanian: Implemented supervised baselines and unsupervised techniques. Worked on writing the report.

References

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, Jun 2002.

[2] Shubhomoy Das, Weng-Keen Wong, Thomas Dietterich, Alan Fern, and Andrew Emmott. Incorporating Expert Feedback into Active Anomaly Discovery. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 853–858, December 2016. ISSN: 2374-8486.

[3] Sean Foley, Jonathan R. Karlsen, and Tālis J. Putniņš. Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed Through Cryptocurrencies? SSRN Scholarly Paper ID 3102645, Social Science Research Network, Rochester, NY, December 2018.

[4] D. Y. Huang, M. M. Aliapoulios, V. G. Li, L. Invernizzi, E. Bursztein, K. McRoberts, J. Levin, K. Levchenko, A. C. Snoeren, and D. McCoy. Tracking Ransomware End-to-end. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 618–631, May 2018. ISSN: 2375-1207.

[5] Joana Lorenz, Maria Inês Silva, David Aparício, João Tiago Ascensão, and Pedro Bizarro. Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity. *arXiv:2005.14635 [cs, stat]*, May 2020. arXiv: 2005.14635.

[6] Thai Pham and Steven Lee. Anomaly Detection in Bitcoin Network Using Unsupervised Learning Methods. *arXiv:1611.03941 [cs]*, February 2017. arXiv: 1611.03941.

[7] Thai Pham and Steven Lee. Anomaly Detection in the Bitcoin System - A Network Perspective. *arXiv:1611.03942 [cs]*, February 2017. arXiv: 1611.03942.

[8] Burr Settles. Active Learning Literature Survey. page 47.

[9] Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I. Weidele, Claudio Bellei, Tom Robinson, and Charles E. Leiserson. Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics. *arXiv:1908.02591 [cs, q-fin]*, July 2019. arXiv: 1908.02591.

[10] Dennis L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2(3):408–421, 1972.

[11] www.elliptic.co. Elliptic Data Set.