Applied Analytics Project - Week 1 Assignment
Identify the Problem Statement and Dataset

Brady Levenson
ADAN 8888 - Applied Analytics Project
Professor Nurtekin Savas
September 8, 2024

Problem Statement

It is known that fraudulent financial transactions occur on a regular basis, but it is difficult to distinguish between them and legitimate financial transactions simply by viewing the transactions. The inability to correctly classify fraudulent financial transactions can promote uncontrolled and untraceable levels of financial losses for banks, investment firms, corporations, and numerous other entities. We are given a dataset of financial transactions and their unique properties. We will analyze this dataset using various machine learning techniques to determine effective measures to identify historical fraudulent transactions and prevent their occurrence in the future.

Articulation of Value

The value in conducting this experimentation process lies in preventing large financial losses incurred from an accumulation of many undetected fraudulent transactions. We have access to 5 months of financial transactions for this project, outlining some of the most relevant details of the respective transactions. We will create models that account for each of the variables present in a training set that we will extract from the original dataset and use the insights gained from this process to provide a holistic view of the dataset. We will then apply the constructed models to the remainder of the available data (e.g. the test set) to identify which financial transactions are fraudulent with a high rate of accuracy. Once we have established a good basis for identifying fraudulent financial transactions, we can apply this basis in the form of anomaly detection programs that will flag the financial transactions as they are occurring and stop their successful processing.

Total Potential Economic Value: $6 Billion Annually

For the purpose of establishing the potential economic profit of creating a model to identify which financial transactions are fraudulent we will hypothesize some fairly reasonable values for certain characteristics of the dataset. Upon further analyses of the dataset and through our modeling process we should be able to produce more accurate metrics for this project's economic profit potential.

- According to 2023 data on financial service industry transactions, it is suspected that about 4.3% of attempted transactions in this industry are fraudulent.
- According to LexisNexis, a financial risk solutions corporation, financial institutions incur a $4.41 additional cost for each dollar that is lost to fraud from customer relation reconciliation, financial recovery efforts, and other efforts, and this metric increases end over end by about 9% each year.

- From a quick calculation of the average transaction value of our dataset, we find that the average financial transaction that we are dealing with has a monetary value of about $50,090.

Calculation

216,960 transactions times 4.3% that are suspected of being fraudulent equals 9,329 suspected fraudulent transactions. If the average value of these transactions is the same as the average value of the dataset as a whole ($50,090), the total monetary value of the suspected fraudulent transactions would be $50,090 per transaction times 9,329 transactions, which equals $467,304,000. Then, accounting for the additional $4.41 of financial loss per one dollar of fraudulent transactions, we have $467,304,000 multiplied by ($1.00 + $4.41), which equals a total of $2,528,115,000 ($2.5 billion). Since these 216,960 transactions are from 5 months of data we can multiply the total by 12/5 to obtain the potential economic value that this project has for one full year of implementation. [1]

$2,528,115,000 times 12/5 equals $6,067,475,000 ($6 billion).

In practice, this figure may be smaller given the hasty estimates we have made of the percentage of fraudulent transactions out of all of the transactions and their average value, but this calculation does exemplify the significant potential economic value that this modeling project has the potential of providing.

[1] Assumptions: The 4.3% of financial transactions that are suspected of being fraudulent are not being flagged currently, the average value of fraudulent financial transactions is roughly equal to the average value of all transactions in the dataset, and There are no losses associated with the implementation of the model.

Assumption Descriptions

Assumption 1: *The 4.3% of financial transactions that are suspected of being fraudulent are not being flagged currently*. Realistically there will already be models in place for conducting this task in the financial industry and our model will be used to improve the outcomes that are already being realized in practice.

Assumption 2: *The average value of fraudulent financial transactions is roughly equal to the average value of all transactions in the dataset*. This assumption is being used so that I have a placeholder for estimation purposes as I have not conducted enough data analysis to determine a better estimate at this stage of the project.

Assumption 3: *There are no losses associated with the implementation of the model.* It remains to be seen how the precision and recall rates of our model will perform, although it is possible that it may classify valid transactions as fraudulent mistakenly. This would detract from the total potential economic value of this model's implementation, and we will see exactly to what extent this factor has an impact on the model's potential once we do some rigorous modeling within the experimentation process.

Project Plan

Here is my 13 week plan adapted from the course syllabus to meet my project's specific needs:

Week 1: Identify the problem statement and dataset

I found the Financial Anomaly Dataset on Kaggle and developed a problem statement and potential solution paths around it. I was originally interested in doing a project pertaining to the field of anomaly detection and I thought fraudulent transaction classification would make for a robust and versatile project that can be extended to other related industries as well. Fraud Detection is a very large field with significant economic potential, so any modeling task that can improve the accuracy of fraud detection techniques strikes me as a worthwhile project to execute. This stage of the project also has allowed me time to ensure my respective environments are set up to compile code as well as organize my sequential progress in an organized manner.

Week 2: Ingest and explore the dataset

Now that the problem statement and dataset are identified, I will explore the dataset by inputting it into an IDE, identifying exactly what I am trying to predict and which variables may make for appropriate predictors to accomplish this task.

Week 3: Perform exploratory data analysis

At this stage I will split my dataset into training, validation, and test datasets. Exploratory data analysis will be conducted at this stage to identify possible patterns in the datasets that could ultimately lead to a model producing significant outcomes. In week 3, I will identify problems that will need to be addressed, opportunities that could be taken, and pre-processing measures that should be implemented to obtain the most practical and accurate results possible in this project. This may include cleaning the dataset, removing NA values if applicable, assigning numeric values to categorical variables, or other preprocessing functions.

Week 4: Make the data model ready

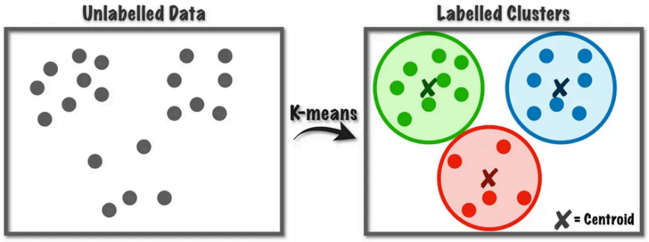
In this week I will execute preprocessing tactics on the datasets as recommended by the findings of the preceding week to ensure that the dataset

Week 5: Engineer features

This week will be used to create and engineer new features and essentially construct final datasets that will be used for the modeling and evaluation processes. Dimensionality reduction may be done in this week to ensure that only variables that are correlated with data points being categorized as anomalies are included in the final datasets.

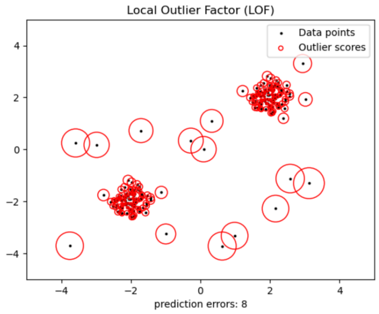
Week 6: Develop first modeling approach (simple, the baseline)

The modeling approaches taken from Week 6 to Week 8 are tentative and subject to change based on the discoveries made within the dataset from Week 2 to Week 5 that may influence my decision, but currently I am considering using an unsupervised K-means clustering model. Since my dataset is unlabeled, it would make sense to consider unsupervised learning approaches. The K-means clustering method groups data points into clusters and can flag potential anomalies that are furthest away from the mean of that cluster.

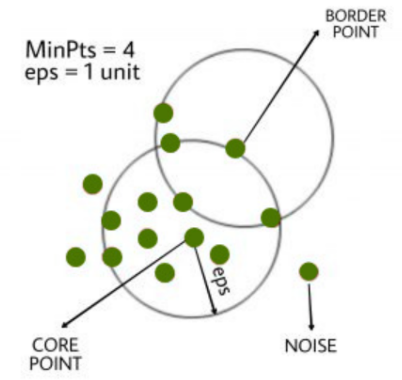


Week 7: Develop second modeling approach (more complex)

The second modeling approach I am considering is Local Outlier Factor. This modeling approach, taking a step up in complexity from the K-means clustering approach, measures the extent to which a data point's density deviates from that of nearby data points. This modeling approach may be more robust than K-means clustering as it factors in many variable weights that could reduce the potential for misclassification as compared with the less dynamic K-means approach.



Week 8: Develop third modeling approach (even more complex)
The third and comparatively most complex approach I am planning to take in this project is the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) unsupervised modeling technique. This approach searches for heaviest concentration areas within high density regions and expands clusters from them. I will determine in the first few weeks of the modeling project if this approach is appropriate for this specific dataset and move forward or adjust to a different modeling technique thereafter.



Week 9: Select the winning model
In Week 9 I will evaluate each of the models I have constructed at this point and determine which model performs the best on the test dataset. This step will help to determine the extent that the model can be used as a viable tool for detecting anomalies in financial transaction data that could be deemed as fraudulent.

Week 10: Data Centric AI
This stage of the modeling process has an objective of improving the winning model by improving the data. By improving the quality and quantity of data used to train this model we will be able to extend the reach and applicability of the model for its confident utilization real-world practice.

Week 11: Explain the model, analyze risk, bias and ethical considerations
This week will be used to rigorously apply practical business meaning to the model's results and conclusions. We will consider relevant business perspectives of the implementation of a model like this one and ultimately determine to what extent it should justly be applied in the real-world.

Week 12: Save and package your model for deployment. Build your model monitoring plan

This stage will be dedicated to saving a fully executable version of the model and ensuring that both the model and the input data's quality is up to standard.

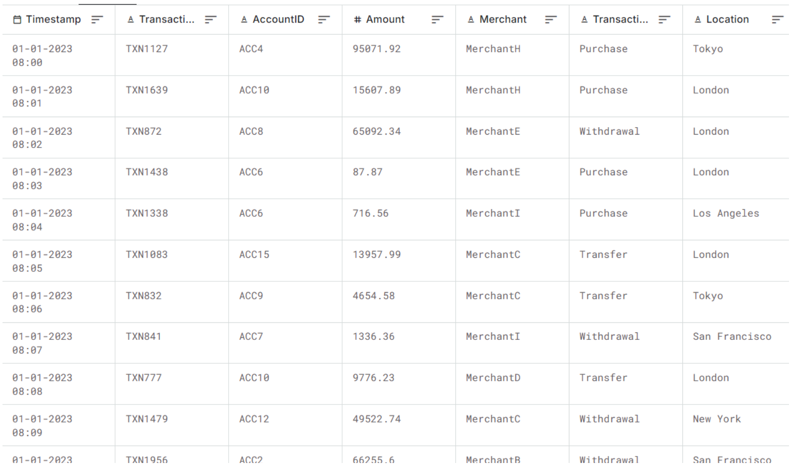
Week 13: Bring it all together

This week will be used to compile the finalized modeling project containing all of the components from the previous 12 weeks of modeling work. At the conclusion of this week I will have completed and prepared a comprehensive modeling project that is ready for presentation to business executives in the financial services industry.

Discussion of Dataset

The dataset being used for this project is the Financial Anomaly Data, found on Kaggle.

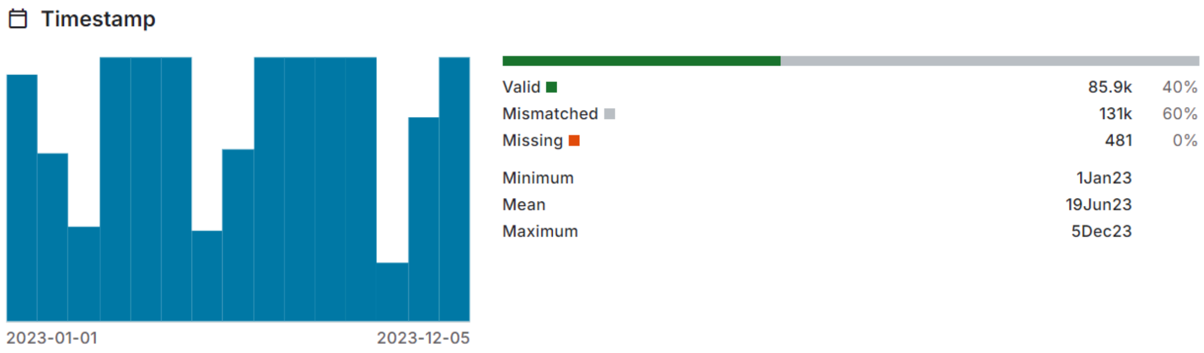
Here is a sample of some rows of the dataset with their corresponding columns:



Timestamp	Transacti...	AccountID	# Amount	Merchant	Transacti...	Location
01-01-2023 08:00	TXN1127	ACC4	95071.92	MerchantH	Purchase	Tokyo
01-01-2023 08:01	TXN1639	ACC10	15607.89	MerchantH	Purchase	London
01-01-2023 08:02	TXN872	ACC8	65092.34	MerchantE	Withdrawal	London
01-01-2023 08:03	TXN1438	ACC6	87.87	MerchantE	Purchase	London
01-01-2023 08:04	TXN1338	ACC6	716.56	MerchantI	Purchase	Los Angeles
01-01-2023 08:05	TXN1083	ACC15	13957.99	MerchantC	Transfer	London
01-01-2023 08:06	TXN832	ACC9	4654.58	MerchantC	Transfer	Tokyo
01-01-2023 08:07	TXN841	ACC7	1336.36	MerchantI	Withdrawal	San Francisco
01-01-2023 08:08	TXN777	ACC10	9776.23	MerchantD	Transfer	London
01-01-2023 08:09	TXN1479	ACC12	49522.74	MerchantC	Withdrawal	New York
01-01-2023	TXN1956	ACC2	66255.6	MerchantB	Withdrawal	San Francisco

This dataset contains 7 feature variables per financial transaction. They are as follows:

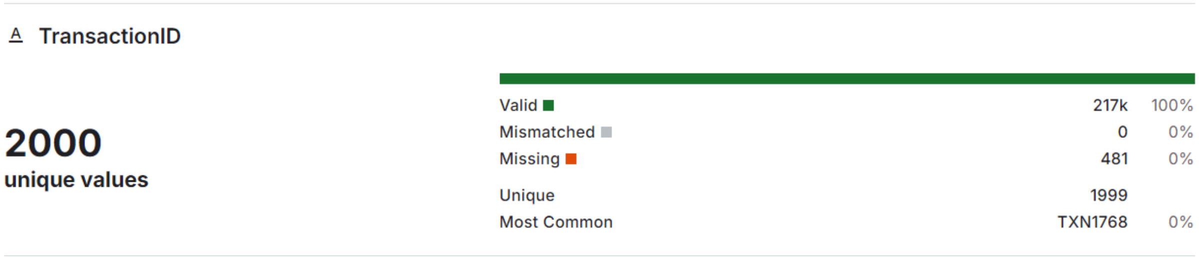
Timestamp: This variable represents the exact date and time that a specific financial transaction occurred. This column of the dataset may be telling of time series patterns that may lead to the detection of anomalies.



Timestamp

Valid 🟩	85.9k	40%
Mismatched ⬜	131k	60%
Missing 🟥	481	0%
Minimum	1Jan23	
Mean	19Jun23	
Maximum	5Dec23	

TransactionID: This column of the dataset can be used to identify certain types of transactions that may contain trends that lead to their identification as anomalies.

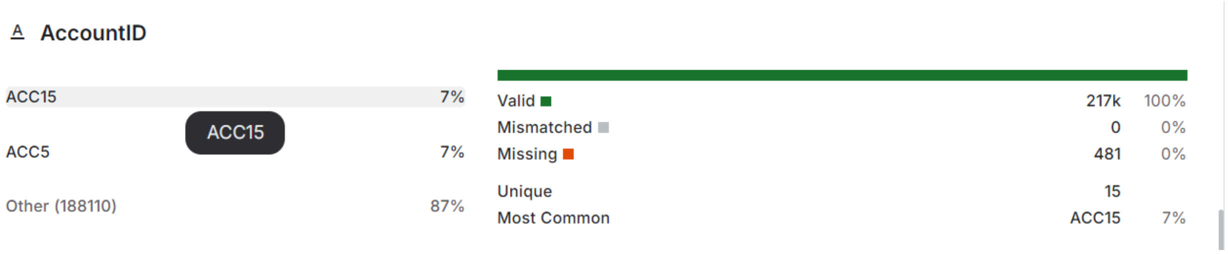


TransactionID

2000
unique values

Valid 🟩	217k	100%
Mismatched ⬜	0	0%
Missing 🟥	481	0%
Unique	1999	
Most Common	TXN1768	0%

AccountID: In a similar light to TransactionID, we will use the AccountID variable to see how the account a transaction occurred within impacts those transactions detection as anomalies.

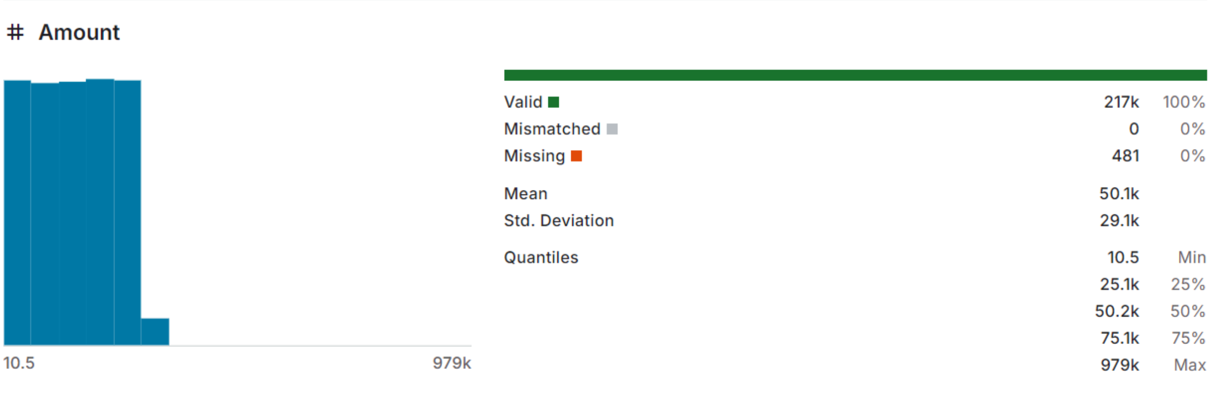


AccountID

ACC15	7%
ACC5	7%
Other (188110)	87%

Valid 🟩	217k	100%
Mismatched ⬜	0	0%
Missing 🟥	481	0%
Unique	15	
Most Common	ACC15	7%

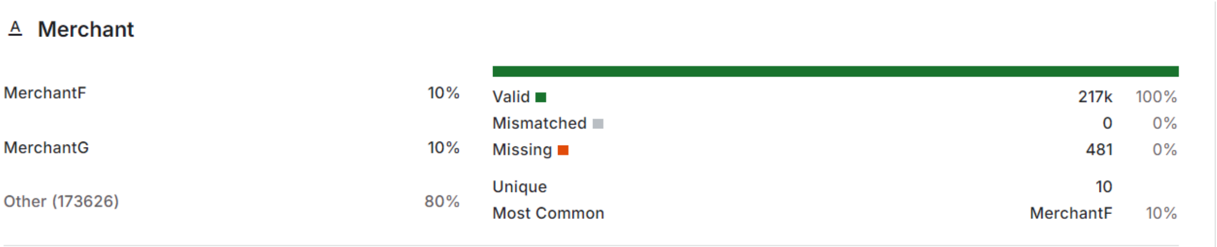
Amount: This column gives the dollar value of the specified transaction. We may ultimately notice certain transaction amounts may create curiosity of them not being valid, honest transactions.



Amount

Valid 🟩	217k	100%
Mismatched ▨	0	0%
Missing 🟧	481	0%
Mean	50.1k	
Std. Deviation	29.1k	
Quantiles	10.5	Min
	25.1k	25%
	50.2k	50%
	75.1k	75%
	979k	Max

Merchant: This column can aid us in identifying the specific industries that the transactions occurred within. We may be able to identify trends of specific industries containing higher volumes of fraudulent transactions than others.

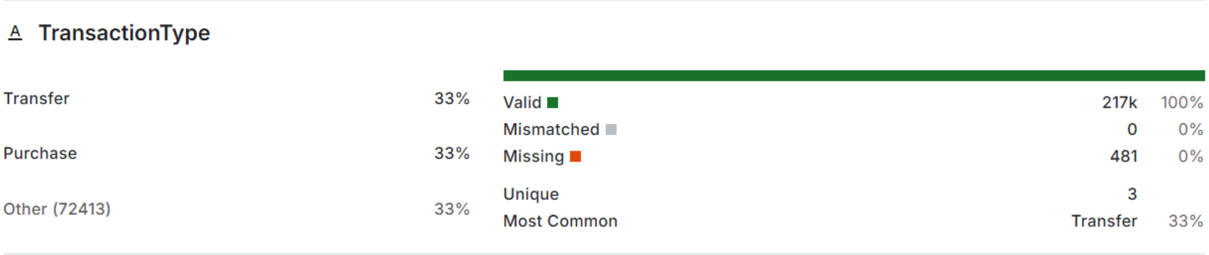


A Merchant

MerchantF	10%
MerchantG	10%
Other (173626)	80%

Valid 🟩	217k	100%
Mismatched ▨	0	0%
Missing 🟧	481	0%
Unique	10	
Most Common	MerchantF	10%

TransactionType: The type of transaction (withdrawal, deposit, transfer, etc.) can possibly be related to whether or not those transactions are deemed to be fraudulent. We will explore this variable more in-depth as the experimentation process ensues.

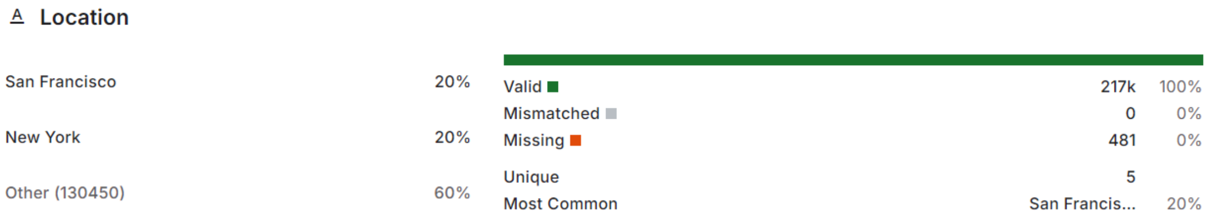


A TransactionType

Transfer	33%
Purchase	33%
Other (72413)	33%

Valid 🟩	217k	100%
Mismatched ▨	0	0%
Missing 🟧	481	0%
Unique	3	
Most Common	Transfer	33%

Location: This variable indicates the city or country that a specific transaction occurred in. We can incorporate it in our models to see if certain geographic locations are producing higher levels of suspicious transactions than others.



Location

San Francisco	20%
New York	20%
Other (130450)	60%

Valid 🟩	217k	100%
Mismatched ▨	0	0%
Missing 🟥	481	0%
Unique	5	
Most Common	San Francis...	20%

(The above feature descriptions were recycled from my week 1 discussion post, for clarification)

Altogether, I will locate connections between combinations of predictor variables from this dataset to produce a model that can identify which financial transactions are the most likely to be fraudulent from the set. Once a well-performing model has been established, it will have the capability of being deployed to resolve the problem of fraudulent financial transactions slipping between the cracks and creating substantial financial losses.

Type of modeling that will solve the problem

Due to the fact that the dataset I have chosen does not contain labels identifying certain transactions as either fraudulent or non-fraudulent, I decided unsupervised classification approaches would be most appropriate for analyzing this dataset. The unsupervised approaches I have tentatively chosen are K-means clustering, Local Outlier Factor, and Density-Based Spatial Clustering of Applications with Noise. These three unsupervised classification techniques each have differing levels of complexity, and depending on how my final dataset ends up looking could have varying levels of effectivity as well.

Bibliography

Caporal, Jack. “Identity Theft and Credit Card Fraud Statistics for 2024.” The Motley Fool, August 26, 2024. https://www.fool.com/the-ascent/research/identity-theft-credit-card-fraud-statistics/.

Choudhary, Mukesh Kumar. “Financial Anomaly Data.” Kaggle, December 17, 2023. https://www.kaggle.com/datasets/devondev/financial-anomaly-data.

“DBSCAN Clustering in ML: Density Based Clustering.” GeeksforGeeks, May 23, 2023. https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/.

Dwirany Puspitasari Jennifer. “K-Means Clustering.” Medium, May 5, 2023. https://medium.com/@JenniferPuspita/k-means-clustering-c0a645715231.

“Every Dollar Lost to a Fraudster Costs North America’s Financial Institutions $4.41 According to LexisNexis True Cost of Fraud Study from LexisNexis Risk Solutions.” LexisNexis Risk Solutions. Accessed September 8, 2024. https://risk.lexisnexis.com/about-us/press-room/press-release/20240424-tcof-financial-services-lending#:~:text=Press%20Releases-,Every%20Dollar%20Lost%20to%20a%20Fraudster%20Costs%20North%20America’s%20Financial,Study%20from%20LexisNexis%20Risk%20Solutions.

“Outlier Detection with Local Outlier Factor (LOF).” Scikit Learn. Accessed September 8, 2024. https://scikit-learn.org/dev/auto_examples/neighbors/plot_lof_outlier_detection.html.