Applied Analytics Project - Week 8 Assignment
Develop Third Modeling Approach

Brady Levenson
ADAN 8888 - Applied Analytics Project
Professor Nurtekin Savas
October 27, 2024

Revisitation of Problem Statement for Motivation to Perform Modeling

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.

Dataset Structure Update

My dataset's dimensionality has been reduced all the way down to 16 unique columns, namely .Timestamp, Amount (log normalized), Date, Day of the Week (7 columns, one-hot encoded), Deviation From Mean (normalized), Account (embeddings), Merchant (embeddings), Transaction Type (embeddings), Location (embeddings), and Hour Group (embeddings). I am happy with the current state of my dataset and believe all of the transformations I have performed on it up to this point will result in each of my models producing viable results that could ultimately be applied in industry.

Revisitation of First Model Approach

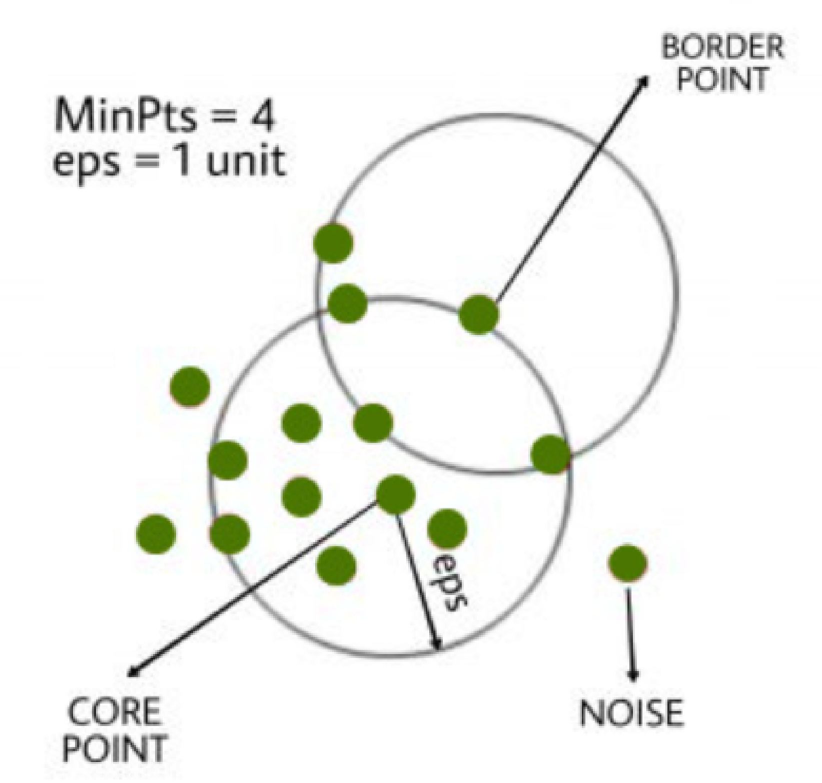
My first model approach for this project was an unsupervised K-Means Clustering Model. There are a couple reasons why I decided this modeling approach may be appropriate for this task. Firstly, the structure of my chosen dataset indicates that this problem should be solved in an unsupervised manner. While the target variable of this project is the binary classification of yes, a financial transaction is an anomaly, or no, a financial transaction is not an anomaly, we do not have a clearly defined labeled training set that our models can learn from directly. Thus, we needed to exercise a more elaborate approach. The K-Means Clustering technique is great for anomaly detection tasks as it allows for comparisons to be made between financial transactions within the same dataset, specifically aiding in the determination of which groups of transactions reside farthest away from the rest. This can help us to identify patterns of data that could potentially represent fraudulent transactions, which can then be isolated for further evaluation.

Revisitation of Second Model Approach

My second model for this project was an unsupervised Local Outlier Factor (LOF) Model. The primary benefit I see with the LOF model over my K-means model is its ability to capture the local densities of points within the context of the financial transaction dataset. This is a factor that my K-means model would have more trouble considering as this technique mainly knows that a point is located within a certain cluster, but does not capture the relative volume of points within specifically compacted spaces. By considering the local density of data points in our dataset, we can determine which points are more isolated relative to others, which may result in those most isolated points being classified as outliers in the dataset. I believe the complexity introduced by this model will present a more trustworthy view of which financial transactions in my dataset may represent outliers, whereas my K-means model may suggest transactions with a relatively lower level of confidence.

Third Model Approach (this week)

For my third and most complex modeling approach of this financial anomaly detection project, I have produced a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) unsupervised model. This model combines elements from my first modeling approach that clustered nearby data points together and my second modeling approach that applied distance metrics to those data points. By defining the relative distance between financial transactions and groups of financial transactions, we can determine what types of points we can consider high density (and to what extent they are high density) as well as what types of points we can consider low density. By evaluating the extent that a point or a region of points exhibits low density, we can prospect the relative probability that those data points truly represent financial fraud and should be flagged.



Sample Diagram of DBSCAN Classification

Complexity of Modeling Approach

Following my projected upward progression of modeling approach complexity, this model is the most complex of the three I have planned, and I would say it has a moderate to high level of complexity. A portion of the complexity of this model comes from the feature engineering that I conducted in the past few weeks, though the model standing alone still possesses complexity that is important to be addressed. This model requires determining which data points in the dataset are within a prespecified distance of one another or have enough nearby data points to classify them as border points or core points respectively while simultaneously isolating points that are not surrounded by many other points and should be factored into anomaly consideration. Additionally, adjusting hyperparameter values can also increase or decrease the base complexity of the model, as we can perform hyperparameter tuning to discover setups that work the best for the specific feature set at hand.

Hyperparameters Evaluated

epsilon (eps)- radius of neighborhood around a point

- This metric is significant as it applies differing perspectives on how we view the relationship between individual points from the dataset. For instance, if epsilon is set very low, then only data points that are spatially very close to one another will be considered in a certain cluster, which consequently would result in many tightly knit neighborhoods that may introduce suboptimal noise to our analysis. A very high epsilon setting leads to more data points being considered in each neighborhood, which may result in true fraudulent transactions mistakenly being grouped with legitimate transactions due to the wide span of the respective neighborhoods. Selecting an optimal value of epsilon is a priority for producing a versatile anomaly detection model.

minimum samples (min_samples) - minimum number of points required to form a dense region

- This metric defines how particular the model is when defining dense regions. A small value of min_samples may result in more dense regions being formed while a large value of min_samples may result in a smaller number of dense regions being formed. These cases result in true outliers going undetected or normal points being classified as outliers respectively, so it is essential to find a balancing point for this metric that optimizes the performance of the model.

metric- method for distance calculation

- This model uses the Euclidean Distance Metric for performance metric evaluation..
    - **Euclidean Distance**- straight line distance between two points in euclidean space.

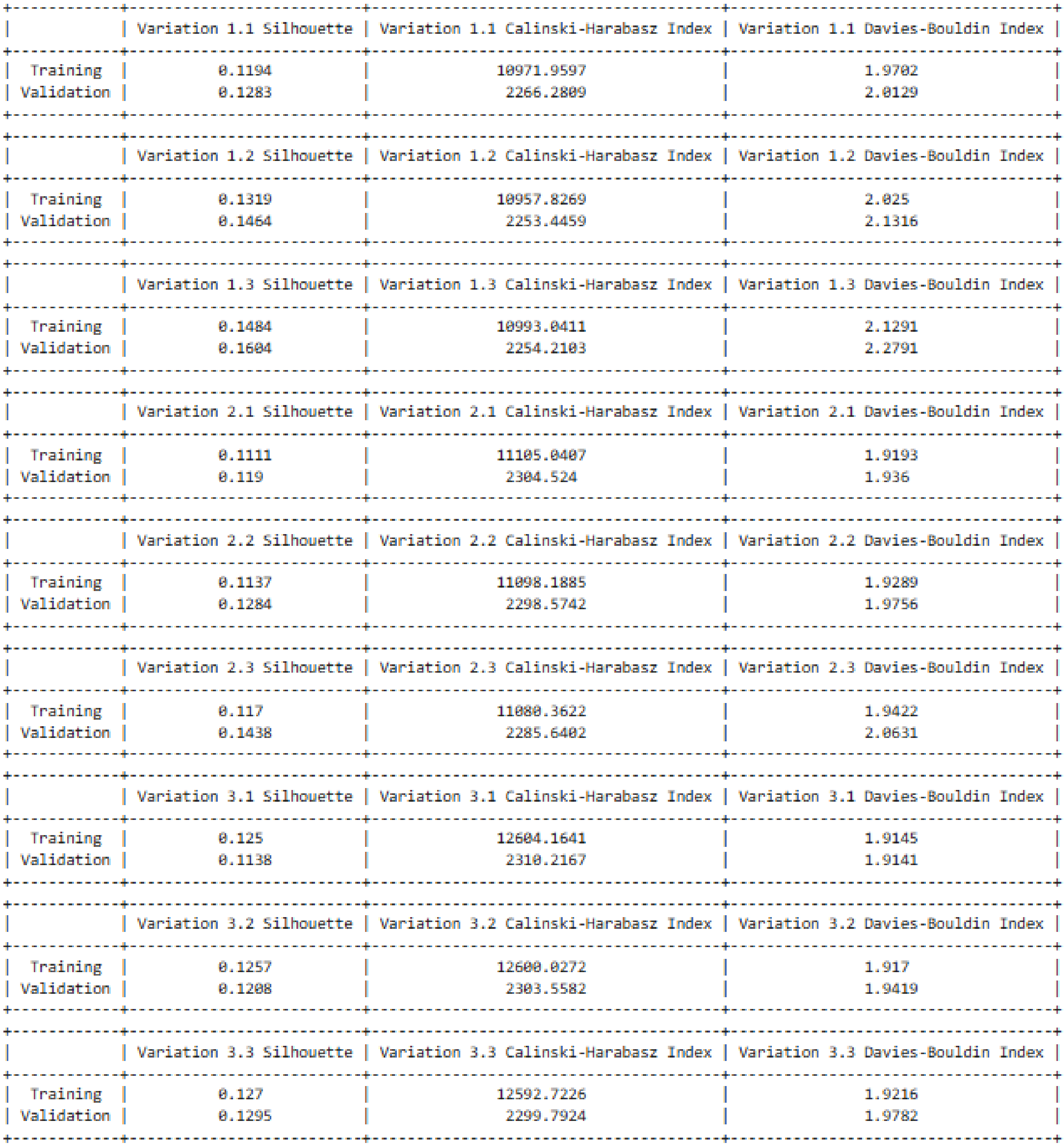
Model Performance Metrics

Silhouette Score: Measures a transaction’s similarity to its own neighborhood versus others. This metric can be leveraged indirectly to determine if certain points are poorly clustered and may be more likely to be anomalies. Lower scores indicate poorer clustering, which indicates greater anomaly potential. Higher scores indicate better clustering, which may make it easier to distinguish anomaly transactions from the well-clustered normal transactions.

Calinski-Harabasz Index- Ratio between the ratio of the sum of between-neighborhood dispersion to within neighborhood dispersion. This index can be used to determine how well-separated our neighborhoods of data are given the modeling approach and hyperparameter specifications in effect. Higher scores indicate well-separated neighborhoods that make it easier to distinguish anomaly transactions from normal transactions given how well normal transactions are clustered together in this scenario.

Davies-Bouldin Index- Average similarity ratio between each cluster and its most similar neighborhood. Higher scores indicate neighborhoods that are very similar to one another and difficult to differentiate, while lower scores point to neighborhoods that are more well-separated and obviously-defined. We are aiming for lower scores for this index to identify good clustering that will give us more confidence to classify data points that stray away from the characteristics of well-clustered data as anomalies.

Metric Calculations Results Table



```
+------------+--------------------------+---------------------------------------+------------------------------------+
|            | Variation 1.1 Silhouette | Variation 1.1 Calinski-Harabasz Index | Variation 1.1 Davies-Bouldin Index |
+------------+--------------------------+---------------------------------------+------------------------------------+
|  Training  |          0.1194          |              10971.9597               |               1.9702               |
| Validation |          0.1283          |               2266.2809               |               2.0129               |
+------------+--------------------------+---------------------------------------+------------------------------------+
+------------+--------------------------+---------------------------------------+------------------------------------+
|            | Variation 1.2 Silhouette | Variation 1.2 Calinski-Harabasz Index | Variation 1.2 Davies-Bouldin Index |
+------------+--------------------------+---------------------------------------+------------------------------------+
|  Training  |          0.1319          |              10957.8269               |               2.025                |
| Validation |          0.1464          |               2253.4459               |               2.1316               |
+------------+--------------------------+---------------------------------------+------------------------------------+
+------------+--------------------------+---------------------------------------+------------------------------------+
|            | Variation 1.3 Silhouette | Variation 1.3 Calinski-Harabasz Index | Variation 1.3 Davies-Bouldin Index |
+------------+--------------------------+---------------------------------------+------------------------------------+
|  Training  |          0.1484          |              10993.0411               |               2.1291               |
| Validation |          0.1604          |               2254.2103               |               2.2791               |
+------------+--------------------------+---------------------------------------+------------------------------------+
+------------+--------------------------+---------------------------------------+------------------------------------+
|            | Variation 2.1 Silhouette | Variation 2.1 Calinski-Harabasz Index | Variation 2.1 Davies-Bouldin Index |
+------------+--------------------------+---------------------------------------+------------------------------------+
|  Training  |          0.1111          |              11105.0407               |               1.9193               |
| Validation |          0.119           |               2304.524                |               1.936                |
+------------+--------------------------+---------------------------------------+------------------------------------+
+------------+--------------------------+---------------------------------------+------------------------------------+
|            | Variation 2.2 Silhouette | Variation 2.2 Calinski-Harabasz Index | Variation 2.2 Davies-Bouldin Index |
+------------+--------------------------+---------------------------------------+------------------------------------+
|  Training  |          0.1137          |              11098.1885               |               1.9289               |
| Validation |          0.1284          |               2298.5742               |               1.9756               |
+------------+--------------------------+---------------------------------------+------------------------------------+
+------------+--------------------------+---------------------------------------+------------------------------------+
|            | Variation 2.3 Silhouette | Variation 2.3 Calinski-Harabasz Index | Variation 2.3 Davies-Bouldin Index |
+------------+--------------------------+---------------------------------------+------------------------------------+
|  Training  |          0.117           |              11080.3622               |               1.9422               |
| Validation |          0.1438          |               2285.6402               |               2.0631               |
+------------+--------------------------+---------------------------------------+------------------------------------+
+------------+--------------------------+---------------------------------------+------------------------------------+
|            | Variation 3.1 Silhouette | Variation 3.1 Calinski-Harabasz Index | Variation 3.1 Davies-Bouldin Index |
+------------+--------------------------+---------------------------------------+------------------------------------+
|  Training  |          0.125           |              12604.1641               |               1.9145               |
| Validation |          0.1138          |               2310.2167               |               1.9141               |
+------------+--------------------------+---------------------------------------+------------------------------------+
+------------+--------------------------+---------------------------------------+------------------------------------+
|            | Variation 3.2 Silhouette | Variation 3.2 Calinski-Harabasz Index | Variation 3.2 Davies-Bouldin Index |
+------------+--------------------------+---------------------------------------+------------------------------------+
|  Training  |          0.1257          |              12600.0272               |               1.917                |
| Validation |          0.1208          |               2303.5582               |               1.9419               |
+------------+--------------------------+---------------------------------------+------------------------------------+
+------------+--------------------------+---------------------------------------+------------------------------------+
|            | Variation 3.3 Silhouette | Variation 3.3 Calinski-Harabasz Index | Variation 3.3 Davies-Bouldin Index |
+------------+--------------------------+---------------------------------------+------------------------------------+
|  Training  |          0.127           |              12592.7226               |               1.9216               |
| Validation |          0.1295          |               2299.7924               |               1.9782               |
+------------+--------------------------+---------------------------------------+------------------------------------+
```

Variation Specifications

Variation 1.1: eps=0.3, min_samples=3
Variation 1.2: eps=0.3, min_samples=5
Variation 1.3: eps=0.3, min_samples=10

Variation 2.1: eps=0.5, min_samples=3
Variation 2.2: eps=0.5, min_samples=5
Variation 2.3: eps=0.5, min_samples=10

Variation 3.1: eps=0.7, min_samples=3
Variation 3.2: eps=0.7, min_samples=5
Variation 3.3: eps=0.7, min_samples=10

Considering Factors for Hyperparameter Tuning
I conducted my hyperparameter tuning to gain perspectives of varying levels of local density specifications this dataset would benefit best from. I created models to consider 0.3, 0.5, and 0.7 epsilon maximum distance values respectively to capture this factor. For each of those three epsilon values, I adjusted the minimum number of samples required for cluster consideration metric to 3, 5, and 10 respectively to determine which density combinations would result in the most optimally performing model.

Top 3 Train Set Performance Metric Rankings (best to worst):
Silhouette Score: Variation 1.3, Variation 1.2, Variation 3.3
Calinski-Harabasz Index: Variation 1.3, Variation 1.1, Variation 1.2
Davies-Bouldin Index: Variation 3.1, Variation 3.2, Variation 2.1

Validation Dataset Performance Rankings:
Silhouette Score: Variation 1.3, Variation 1.2, Variation 3.3
Calinski-Harabasz Index: Variation 3.1, Variation 2.1, Variation 3.2
Davies-Bouldin Index: Variation 3.1, Variation 2.1, Variation 3.2

My Variation 1.3 Model exhibited high Silhouette and Calinski Harabasz Scores when faced with train set data, indicating that it may be the best equipped of the variations produced here today at identifying potentially fraudulent financial transactions due to their anomaly status within the dataset. After applying the same models to validation set data, my Variation 3.1 model proved to perform best in the categories of Calinski-Harabasz Index and Davies Bouldin Index, indicating that it may generalize the best on unseen data for classifying anomaly financial transactions within the dataset.

Week 8 Winning Model

When considering all variations across both training and validation data, I select Variation 3.1 as my winning model of the week. In 3 out of 6 performance categories, this model placed first amongst the six unique variations, namely train set Davies-Bouldin Index, validation set Calinski-Harabasz Index, and validation set Davies-Bouldin Index. Based on the statistics calculated in this portion of the project, I would expect Variation 3.1's hyperparameter combination of epsilon=0.7 and min_samples=3 to perform the best on an unseen test set. This indicates that a larger maximum distance allowed between points and a smaller number of samples required to classify regions as core points allows the model to identify more anomaly transactions than what may be possible with the other hyperparameter tuning combinations applied here today. It is important to consider that it is possible that this model is classifying some normal transactions as anomalies due to its less stringent classification requirements than its other variation counterparts, so we will factor this observation into the selection of the overall winning model from the past three weeks.

Next Steps

Next week, I will be applying my winning models of each of the past three weeks to an unseen test dataset to determine which model is the overall winner of this project that will be given the greatest amount of consideration for real world financial anomaly detection tasks.