Applied Analytics Project - Week 9 Assignment
Select the Winning Model

Brady Levenson
ADAN 8888 - Applied Analytics Project
Professor Nurtekin Savas
November 3, 2024

Revisitation of Problem Statement for Motivation to Select an Appropriate Model

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.

Purpose of Selecting a Winning Model

By selecting a winning model, we can isolate the best performing model of the many variations we have constructed throughout the duration of this project. We can then refine our data to further optimize this selected model and determine if it presents a valid real world solution to the significant problem of identifying fraudulent financial transactions as they are occurring in real time and prevent their destructive effects..

Validation Error For All Models

K-Means Clustering Model Variations

	Variation	Init	n_init	k	Silhouette Score	max_iter	tol
0	km_v1	random	1	5	0.249626	NaN	NaN
1	km_v2	k-means++	1	5	0.247696	100.0	0.010
2	km_v3	k-means++	1	10	0.490904	200.0	0.001

Local Outlier Factor Model Variations

	Variation	n_neighbors	metric	Silhouette Score (Val)
0	lof_v1	15	euclidean	0.337440
1	lof_v2	10	euclidean	0.325215
2	lof_v3	15	manhattan	0.236254

Density-Based Spatial Clustering of Applications with Noise

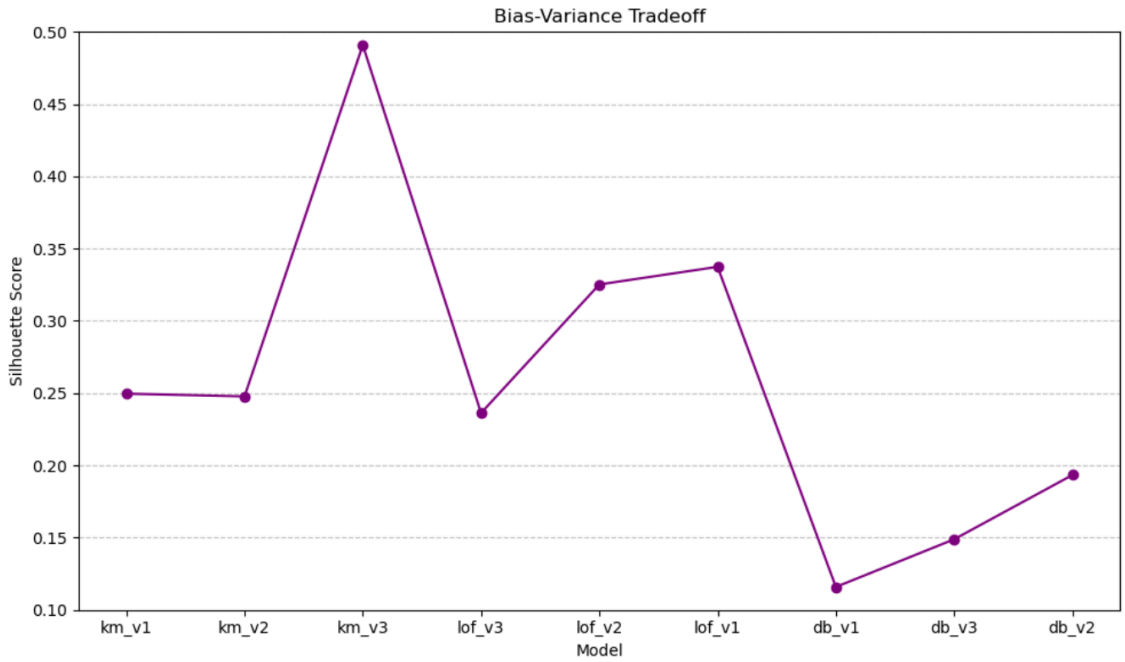
	Variation	eps	min_samples	Silhouette Score (Val)
0	db_v1	0.7	3	0.115870
1	db_v2	0.3	10	0.193271
2	db_v3	0.3	5	0.148751

Final Winning Model Based on Validation Performance: K-Means Clustering Variation 3

Reasoning For This Result

Throughout the course of this project, we have optimized our Financial Anomaly Dataset to capture complex relationships held between seemingly simple data fields to successfully complete an anomaly detection task. Through the implementation of one-hot encoding, feature engineering, time interval bucketization, and feature embedding, this dataset has been tailored to be input to a simple model and still attain robust outputs. I originally had prospected that one of my Local Outlier Factor models would be my top performing model for this project given its consideration of both bias and variance reduction, though it turned out that a simple, yet strategically hyperparameter-tuned model outperformed them (hence the winning model selection). The third variation of my K-Means Clustering Model represented the simplest modeling approach explored through this experimentation process, but with the most hyperparameter tuning of those K-Means models. This model construction proved to be ideal for capturing trends from a complex dataset in a "simple" manner due to the preemptive feature adjustments done on the dataset in the opening few weeks of this assignment. I suspect that the dataset taken in a less finely adjusted form may benefit from one of the more complex models we have created throughout this project due to their abilities to capture more nonlinear relationships.

Bias-Variance Tradeoff Graph



What Does the Bias_Variance Tradeoff Graph Tell Us?

The bias-variance tradeoff graph depicts a few very notable trends. Firstly, the graph is constructed with the x-axis indicating a specific model (ordered from simplest to most complex, left to right) and the y-axis indicating the Silhouette Score for that model (higher scores indicate data points being most similar to data in their own cluster than neighboring clusters, indicating more well-defined separation of clusters and "easier" anomaly detection). The two primary trends of the graph lie within each of the three model types (K-Means, LOF, and DBSCAN) and in comparing these three model umbrellas. There is a visible upward trend in model performance within variations of the same type of model as the hyperparameter tuning done to the base model becomes more complex. This hyperparameter tuning essentially refines the model to make more precise predictions on the dataset. Despite this upward trend in performance as intra-model complexity increases, there is a downward trend in performance as inter-model complexity increases. The main explanation for this circumstance is that naturally simpler modeling approaches are more susceptible to high variance, where adjustments to hyperparameter values can significantly impact model performance. With this high variance, it is possible for these models to elicit low bias in some variations, as evidenced by our winning model, K-Means Variation 3. Our DBSCAN Models proved to be overly complex for this anomaly transaction classification task. Its complexity led to more stable predictions (low variance), yet it was unable to capture the trends held by the dataset as effectively as simpler modeling approaches (high bias). Our LOF Models represent the "middleground" of the three, eliciting moderate levels of both bias and variance.

Winning Model Performance on Test Dataset

```
{'Variation': 'km_v3', 'Init': 'k-means++', 'n_init': 1, 'k': 10, 'max_iter': 200, 'tol': 0.001, 'Silhouette Score': 0.4865
```

The winning model performed similarly on each of the train, validation, and test datasets. This is exemplary of a finely hyperparameter-tuned K-Means Clustering Model's potential to perform adequately on unseen data that has been formatted efficiently, preemptively. Based on the Silhouette Score Performance Metric used to evaluate the model's performance on test data, it appears the model has some predictive capabilities for this classification task, though it is not perfect by any means and does have limitations. A practical implementation of this model in industry may allow us to further analyze the effectiveness of this model at successfully identifying anomaly financial transactions, though the performance metrics evaluated in this project significantly aided the process of ensuring the selected winning model has considerable potential to perform this prediction task well.

Winning Model Performance Metrics by Stage

	Model Stage	Silhouette Score
0	Train	0.4638
1	Validation	0.4909
2	Test	0.4865

Fluctuations Across Different Stages

I attribute the fluctuations of this model across train, validation, and test stages to variances held within each of the three DataFrames (train, validation, and test), and not to inadequacies in the model training process). This model will of course not achieve the exact same silhouette score values across all possible transaction datasets, though should be expected to perform within a relatively narrow range when presented with new and unseen financial transaction data.

Next Steps

We have now selected a "winning model" and will work to improve the performance of this model through further data and model adjustments. By performing this next step we will better be able to determine if this model is appropriate to deploy on real-time financial transaction data to detect and flag potentially fraudulent transactions as they occur.