Applied Analytics Project - Week 11 Assignment
Explain the Model, Analyze Risk, Bias, and Ethical Considerations

Brady Levenson
ADAN 8888 - Applied Analytics Project
Professor Nurtekin Savas
November 17, 2024

Revisitation of Problem Statement for Motivation to Analyze Risk, Bias, and Ethical Considerations

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.

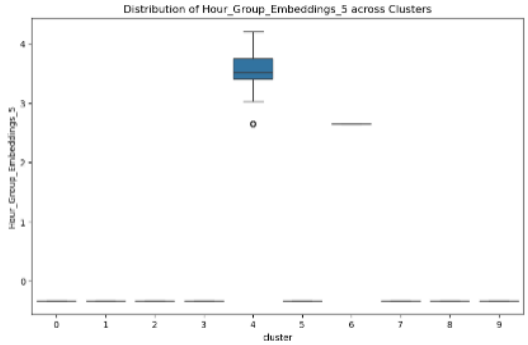
Important Features in the Model

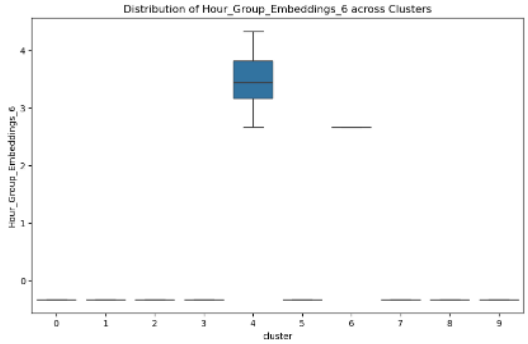
```
Top features per PC based on their contribution to variance:
PC1: ['Hour_Group_Embeddings_5', 'Hour_Group_Embeddings_6', 'Hour_Group_Embeddings_1']
PC2: ['Transaction_Embeddings_4', 'Transaction_Embeddings_6', 'Transaction_Embeddings_7']
PC3: ['Location_Embeddings_6', 'Location_Embeddings_4', 'Location_Embeddings_3']
PC4: ['Account_Embeddings_3', 'Account_Embeddings_5', 'Account_Embeddings_4']
```

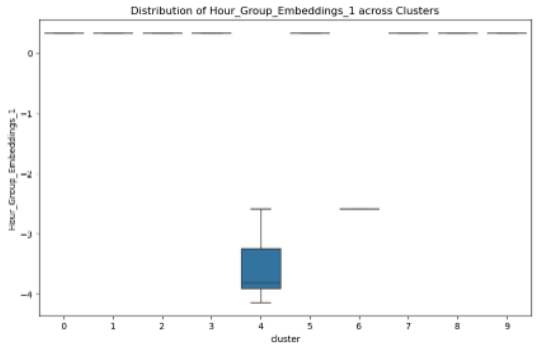
Following a Principal Components Analysis approach to identifying the most contributory features of our data to model results, we have determined that Time, Transaction Type, Location, and Account ID features are strong indicators for our model results. The results above exhibit 12 specific varieties of the aforementioned feature types, detailing the specific principal component that each listed feature has prevalence in.

Sample Spreads of Data in Model:

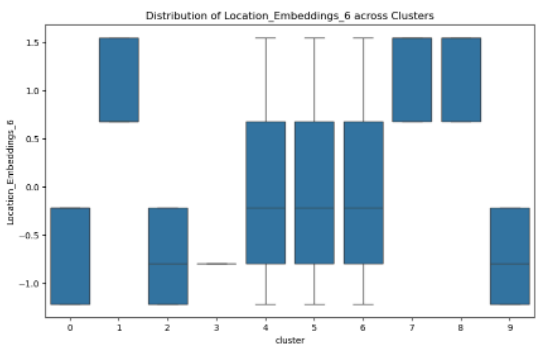
Hour Group Distributions

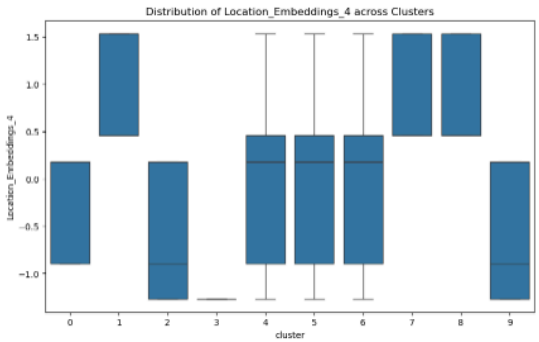


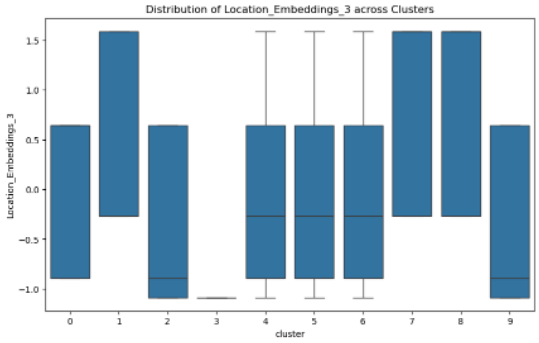




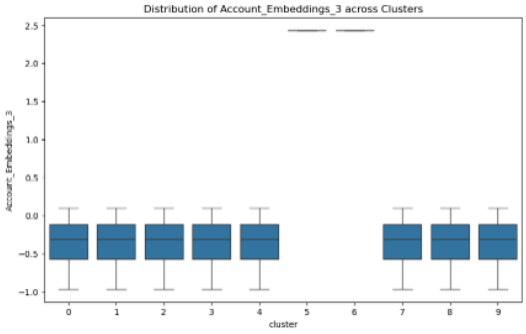
Location Distributions

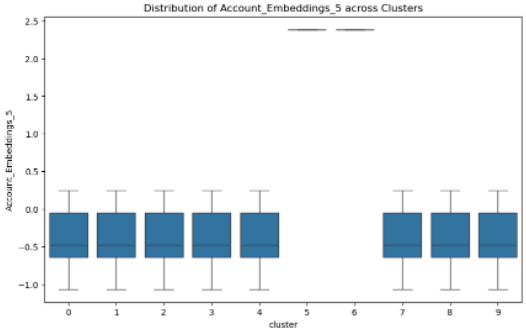


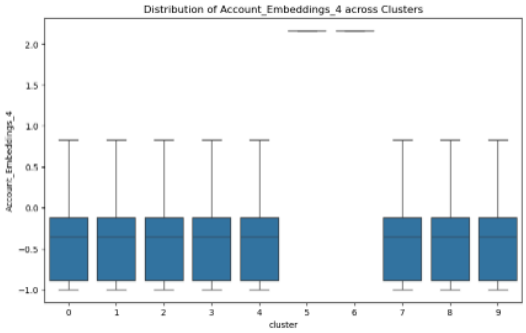




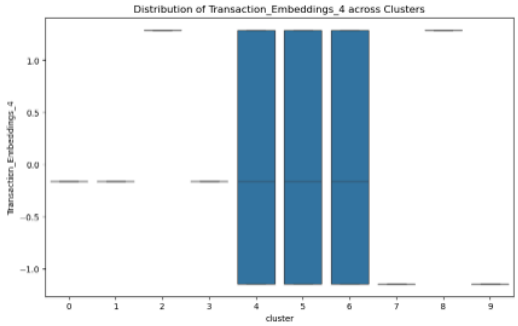
Account ID Distributions

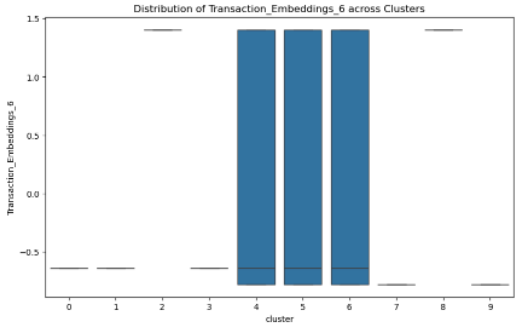


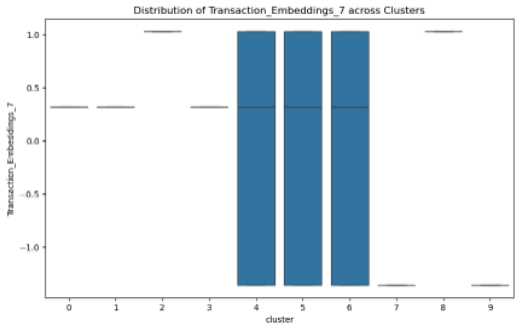




Transaction Type Distributions







Analysis of 5 Random Predictions

We isolated five random predictions to analyze how they were individually classified as anomalies within the given dataset. We first extracted some of the data points with the greatest distances from their cluster's centroid. We then randomly selected five of them to discuss the decision weight of each of that specific transaction's feature weights. Below are some sample feature distances from that specific transaction's cluster's centroid, both originally and after applying a transformation to "move" that transaction 10% closer to its cluster's centroid.We took this approach due to the unsupervised nature of this experiment. By making an anomaly transaction 10% closer to its cluster's centroid, this moves that transaction back within the bounds of what constitutes a normal transaction (which we defined in Week 9 when we selected our winning model and applied it on test data. From the attached results, we can see certain feature contributions to model outcomes. Larger values within the results represent greater feature influence on transaction classifications. The Time, Transaction Type, Location, and Account ID- based features, as alluded to in the previous section, have the most influence on these classification decisions. As evidenced below, each feature only needs to adjust slightly for the entire classification of a transaction to change directions, which is expected given the subtleties that the anomaly transactions within this dataset are being classified based on.

Distance from CLuster Centroids Before 10% Adjustment

Feature-wise contributions to distance from centroid for selected points:

	Day_0	Day_1	Day_2	Day_3	Day_4	Day_5	Day_6
0	0.172981	0.172447	0.173179	0.163272	0.162168	6.234886	0.167943
1	0.170637	0.164565	0.172248	6.156493	0.172364	0.162026	0.161660
2	0.174762	0.179575	5.821240	0.159325	0.170867	0.152382	0.154537
3	5.865028	0.174708	0.163556	0.161403	0.154641	0.173756	0.168029
4	0.174762	0.179575	0.176124	6.182390	0.170867	0.152382	0.154537

	Deviation_From_Mean	Account_Embeddings_0	Account_Embeddings_1	...
0	70.459024	2.339557	0.381855	...
1	68.323767	1.139947	0.060031	...
2	32.048874	0.000049	0.001941	...
3	68.550852	0.518583	0.383460	...
4	55.089934	0.721595	1.055179	...

	Location_Embeddings_6	Location_Embeddings_7	Hour_Group_Embeddings_0
0	0.252400	0.006409	7.499109e-29
1	0.002823	3.066500	3.462360e-29
2	0.653537	3.687054	4.727678e+00
3	0.223810	0.902671	3.865418e-29
4	2.346223	0.332543	3.899476e-01

	Hour_Group_Embeddings_1	Hour_Group_Embeddings_2	Hour_Group_Embeddings_3
0	2.899372e-29	9.222585e-29	1.701752e-28
1	7.595560e-29	1.159749e-28	3.038224e-28
2	1.053798e+00	5.283089e+00	3.366768e+00
3	7.595560e-29	1.159749e-28	3.057606e-28
4	4.913953e-03	4.730771e-01	2.682538e-03

	Hour_Group_Embeddings_4	Hour_Group_Embeddings_5	Hour_Group_Embeddings_6
0	4.738096e-29	7.888609e-31	4.930381e-32
1	1.020712e-28	7.888609e-31	2.829083e-28
2	4.111301e+00	7.395323e-01	6.868740e-01
3	1.100738e-28	6.933348e-31	2.961310e-28
4	7.773590e-01	1.356645e-01	1.636402e-01

	Hour_Group_Embeddings_7
0	5.127904e-29
1	1.100738e-28
2	3.042660e+00
3	1.112417e-28
4	8.748556e-02

Distance from Cluster Centroids After 10% Adjustment

Adjusted feature values for points moved closer to centroids (10% adjustment):

	Day_0	Day_1	Day_2	Day_3	Day_4	Day_5	Day_6
0	-0.371191	-0.372631	-0.371779	-0.362359	-0.361778	2.239072	-0.369701
1	-0.371474	-0.373591	-0.371891	2.234708	-0.360531	-0.361552	-0.370475
2	-0.370977	-0.371781	2.177730	-0.362850	-0.360712	-0.362769	-0.371371
3	2.180409	-0.372359	-0.372951	-0.362591	-0.362724	-0.360121	-0.369691
4	-0.370977	-0.371781	-0.371426	2.234187	-0.360712	-0.362769	-0.371371

	Deviation_From_Mean	Account_Embeddings_0	Account_Embeddings_1	...
0	-7.558157	1.633973	-0.906967	...
1	-7.439646	-0.710725	-0.129331	...
2	-5.069675	0.230465	-0.305022	...
3	-7.457654	0.890389	0.209219	...
4	-6.654653	1.001293	-1.269169	...

	Location_Embeddings_6	Location_Embeddings_7	Hour_Group_Embeddings_0
0	-1.164619	-0.771620	0.311277
1	-0.789408	1.739231	0.311277
2	-0.713879	1.722328	-2.235387
3	-1.167550	-0.684617	0.311277
4	1.392260	-0.524824	-4.754290

	Hour_Group_Embeddings_1	Hour_Group_Embeddings_2	Hour_Group_Embeddings_3
0	0.327891	-0.308537	0.317453
1	0.327891	-0.308537	0.317453
2	-2.685324	2.178219	-2.379826
3	0.327891	-0.308537	0.317453
4	-3.546126	4.865892	-4.077828

	Hour_Group_Embeddings_4	Hour_Group_Embeddings_5	Hour_Group_Embeddings_6
0	-0.314228	-0.329445	-0.329433
1	-0.314228	-0.329445	-0.329433
2	2.300350	2.742903	2.750950
3	-0.314228	-0.329445	-0.329433
4	4.918732	3.848362	3.860923

	Hour_Group_Embeddings_7
0	-0.319338
1	-0.319338
2	2.420499
3	-0.319338
4	4.256591

Protected Categories Discussion

My dataset does not directly contain protected categories such as age, race, or ethnicity, though it does contain categories that may potentially be correlated with protected categories, which is an important factor to consider when analyzing whether or not my dataset or model output is biased. As a financial anomaly detection dataset, this dataset includes a series of financial transactions with original features of Timestamp, AccountID, Merchant, TransactionType, Location, TransactionID, and Amount. My model incorporates time variables, AccountIDs, Merchants, Locations, and Amounts, which may all have close associations with protected classes.

Bias Based on Correlations to Protected Categories

As mentioned in the previous section, while not explicitly containing protected category data, our dataset does contain categories that may have close associations to protected categories. Since we do not have any given protected class data in the original dataset, we will use inference to estimate the relative level of protected category-based bias present in our data and model. We will focus mainly on the most relevant features to the model, as computed earlier.

Time: Using this feature to explain model variance may unintentionally add bias to our modeling process. It is quite possible that younger (and potentially less likely to have solidified daily work schedules than their older counterparts) people may elect to make their financial transactions during the day, whereas older, more seasoned professionals may primarily make their financial transactions in the evening. This factor could lead to the classification of anomaly financial transactions indirectly on an age variable, which is problematic given Age's status as a protected category.

Location: The location field of this dataset contains various different global locations. Using this variable as part of our analysis may accidentally incorporate racial demographics into our modeling process. This is potentially problematic as Race is another protected class that must be accounted for in bias analysis.

Account ID: Using specific individuals' Account ID's as part of the financial anomaly transaction classification process may add bias to our modeling process by numerous means including the

incorporation of Gender, Race/Ethnicity, Age, and other protected categories in our classification analysis. This notable concern arises from indirectly utilizing users' personal demographic information when inputting transaction data to our model for the model to classify.

Bias Removal Strategies

While it would be rather difficult to remove protected category-based bias from this dataset as it stands due to the nature of us merely prospecting the amount of bias contributed by each variable based on assumptions of how the data was collected, we can still discuss some theoretical strategies for removing said bias and their impact on our dataset and model. One approach, if reasonably possible to complete, would be to resample the original data with consideration of bias to begin this project with a more neutral and unbiased dataset. This approach would allow us to obtain more or less unbiased model results, which would lead to us feeling more confident about deploying this model in the real world. Another approach to bias mitigation is to identify features that may have high correlations with protected categories and undersample from these features to give them less weight in the model's final classification decisions. This strategy creates a more balanced dataset that accounts for reducing bias, which may lead to more ethically sound model outputs. The main difficulty with executing this approach on this dataset is the uncertainty of our prospective protected category correlations to features present in our dataset. Given a dataset that includes explicit protected categories, we could precisely calculate feature correlation to these categories and perform undersampling of those highly-correlated features accordingly.

Risks of Using Model on Stakeholders

While this model has proven to have moderate potential to successfully identify potentially fraudulent financial transactions in real time and flag them to undergo further review, it poses some considerable ethical risks that stakeholders need to take into account when deciding if this modeling approach is practical for their specific use. One primary concern is the model acting in a discriminatory manner towards people of certain protected categories including but not limited to elderly people, men or women, or people of certain ethnicities. This factor alone could decentivize stakeholders from deploying the model or make them want to work with the developers of the model to remove inherent biases. Using a model that could contain biases of this variety may also result in serious legal ramifications including lawsuits or fines. If a stakeholder is found to have used this model in a discriminatory manner, they may be subject to these legal consequences.

Beyond unfair practices and legal aspects, stakeholders may also avoid using this model out of a desire to preserve their reputations and images. By utilizing potentially unfair systems or services, stakeholders may gain bad rapport with their customers or other close industry associates.

Next Steps

Now that we have analyzed risk, bias, and ethical considerations, our next step in the end-to-end modeling process will be to save and package our model for deployment and to build a model monitoring plan.