Credit Card Fraud

By Yeon Hwa Choi

Questions

The security team in a bank wants to identify fraudulent transactional activities.

Which transactions are fraud?

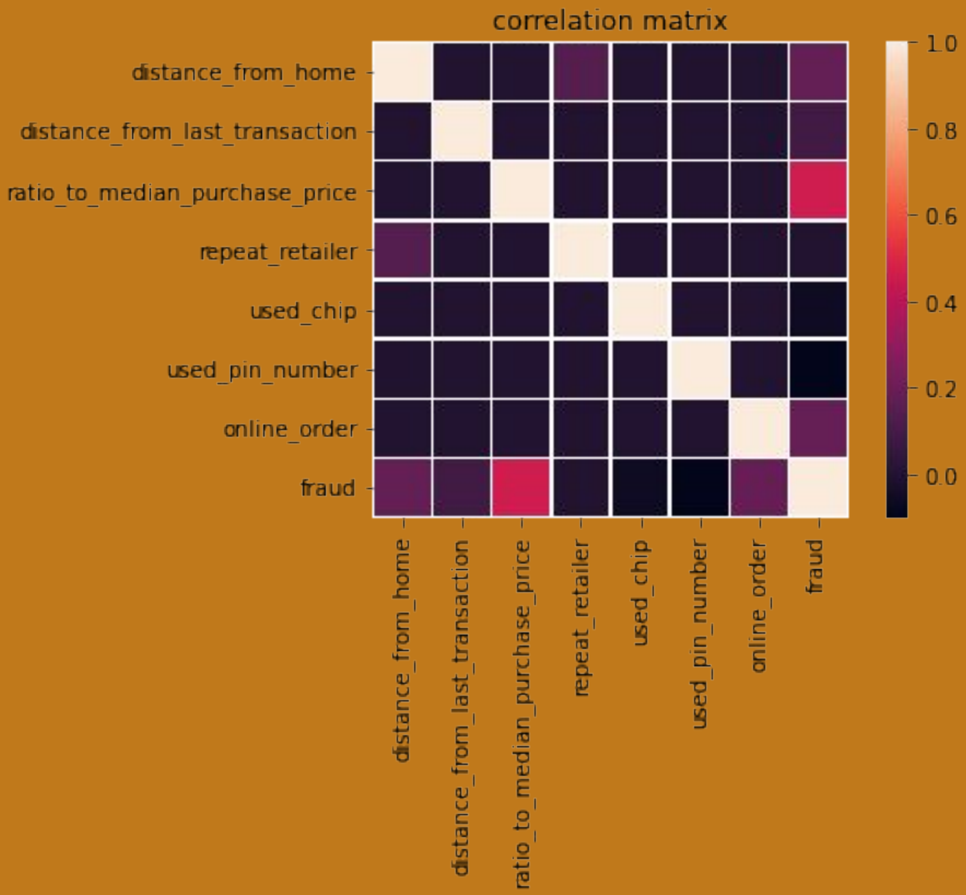
Dataset

Total number of data: 1,000,000

Number of Columns: 8

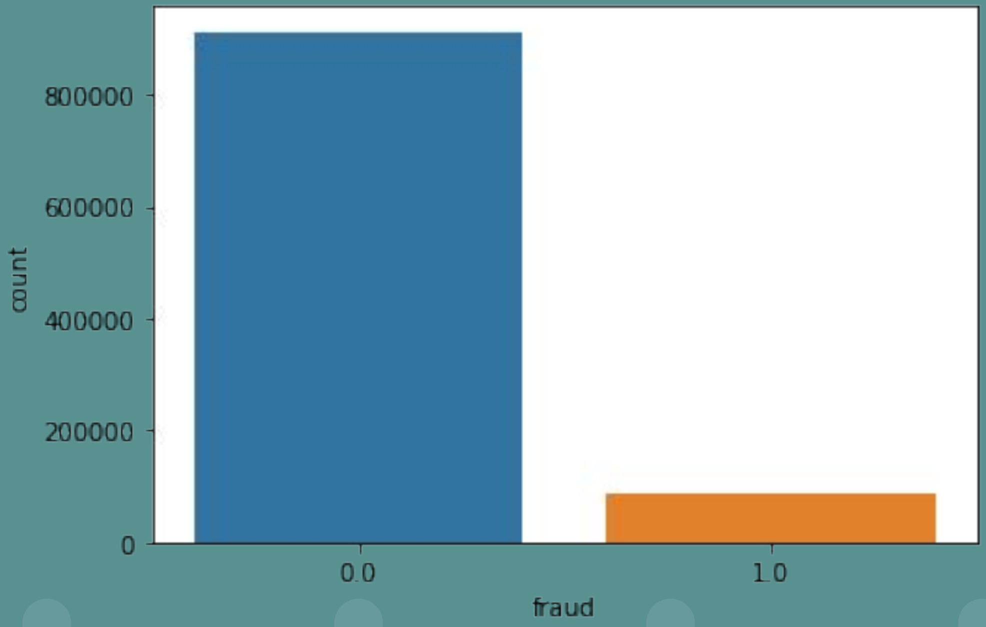
Columns: sex, age, hypertension, heart disease, ever married, work type, Residence type, avg glucose level, bmi, smoking status, stroke

Correlation Table Using Heatmap



- Ratio to median purchase price and fraud have the highest correlation.
- Fraud shows slight correlation with distance from home and online order

Fraud Transactions Count



- 87,403 number of transactions out of total 1 million data are fraud activities.
- In other words, 8.7% of the transactions are fraud transactions.

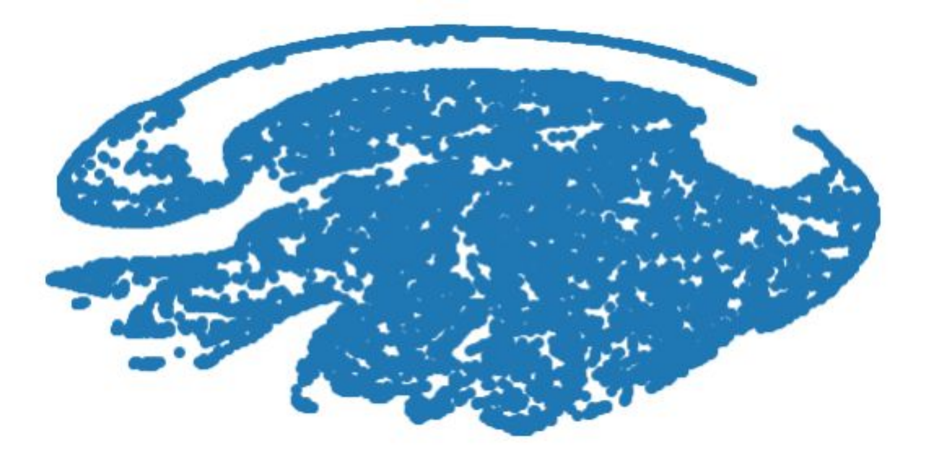
X = all columns except Fraud

Y = Fraud (Dependent Variable)

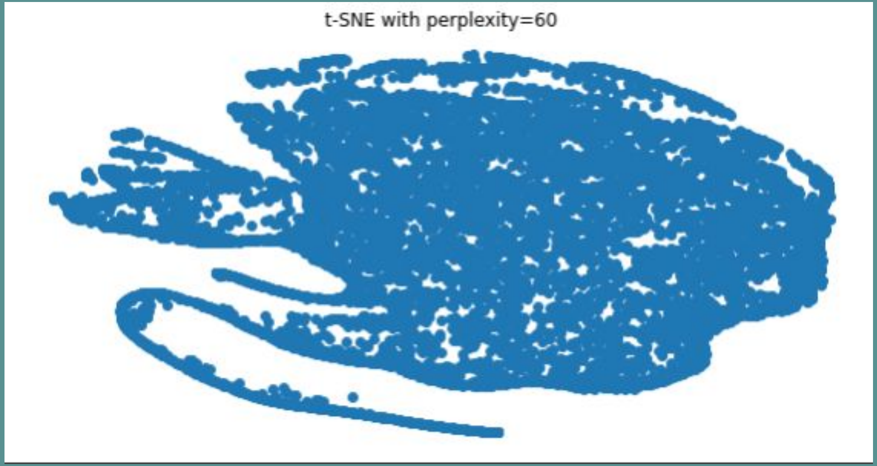
PCA Dimensionality Reduction

- Visualizing the dataset in two-dimensional space using the PCA's first two components.
- The graph do not visually gives a clear distinction between two values.

t-SNE Dimensionality Reduction



n_components=2, verbose=1, perplexity=40, n_iter=300

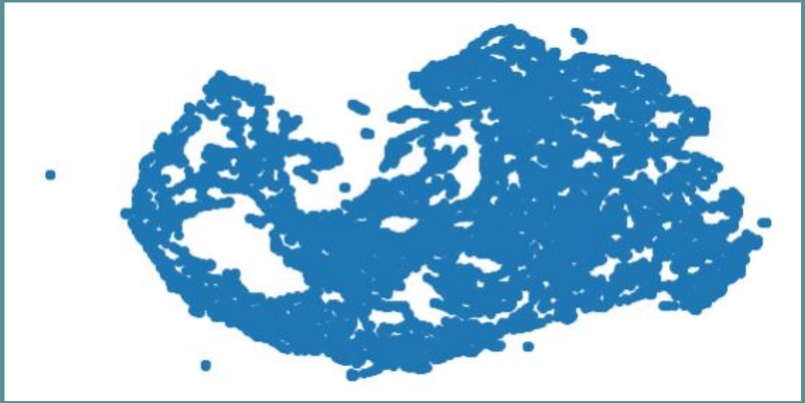


t-SNE with perplexity=60

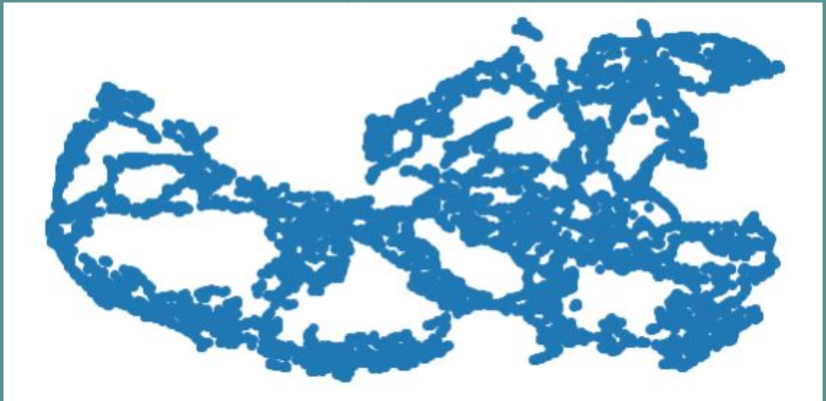
n_components=2, verbose=1, perplexity=60, n_iter=300

- t-SNE do not provide a good local similarities.
- Increasing perplexity do not give any better result.

UMAP Dimensionality Reduction



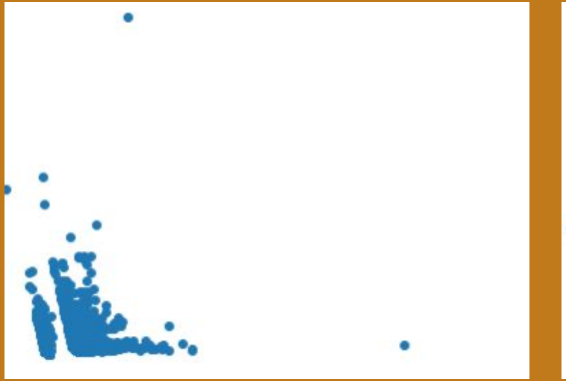
Number of neighbors=5, min distance =0.3, metric='correlation'



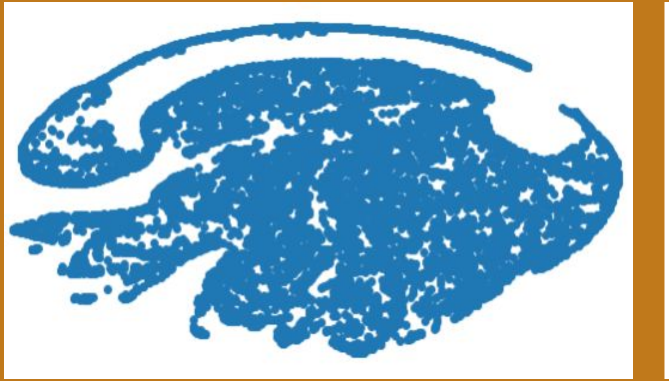
Number of neighbors=7, min distance =0.1, metric='correlation'

- Comparing two UMAP models with different parameters.
- UMAP model with number of neighbors of 7 and minimum distance of 0.1 visually more divided into groups.

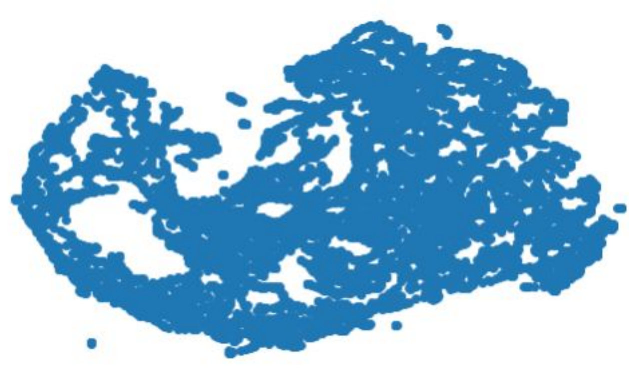
Comparison Between Dimensionality Reduction Models



PCA Model



t-SNE Model



UMAP

- Local similarity graphs for three different Dimensionality reduction models are compared.
- All three models give different images and shapes.
- It is difficult to compare and choose better model by looking at the graphs.

K-Means Clustering

ARI for two cluster k-means: 0.20296998519306872
ARI for three cluster k-means: 0.10120324153042251
ARI for four cluster k-means: 0.10059730975901982

Silhouette score for two cluster k-means: 0.56866939947411
Silhouette score for three cluster k-means: 0.6760462998301284
Silhouette score for four cluster k-means: 0.6788334300627757

- K-Means clustering models resulted in relatively high scores in both ARI and Silhouette.
- Two-Cluster K-Means model gives the best ARI score while Four-Cluster K-Means model gives the best Silhouette score.

DBSCAN Clustering

Adjusted Rand Index of the DBSCAN solution: -0.000807629043919622

The silhouette score of the DBSCAN solution: 0.618181023643096

- DBSCAN Clustering model gives a very small negative number for ARI score.
- DBSCAN’s Silhouette score is favorable.

Gaussian Mixture Model Clustering

ARI score of Gaussian Mixture: 0.22344945949428616

Silhouette score for Gaussian Model : 0.4749669884265485

- For Gaussian Mixture Model, both ARI score and Silhouette score are relatively favorable.
- In terms of ARI score, Gaussian Mixture Model gives the highest score.

Agglomerative Clustering

ARI score of linkage method average: 0.17508812359776627

Silhouette score of linkage method average: 0.4057609808208884

- Agglomerative Clustering Model gives moderately good scores in both ARI and Silhouette.
- Among all four models, Agglomerative Clustering method resulted in lowested Silhouette score.

Conclusion

Three different Dimensionality Reduction models has been used for the analysis: **PCA**, **t-SNE**, and **UMAP**. Based on my research, it was difficult to make a conclusion based on visualization of research using graphs.

Also, four clustering models were utilized: **K-Means**, **DBSCAN**, **Gaussian Mixture**, and **Agglomerative Clustering**. Four models were evaluated based on ARI score and Silhouette score. **K-Means** model exhibited the highest Silhouette score, and **Gaussian Mixture** model exhibited the highest ARI score.

Recommendation: K-Means model with four-cluster using PCA and Gaussian Mixture model are recommended to train the unsupervised credit card dataset and identify fraudulent credit card activity.