Name: Anim Ohene, Henry Nguyen

Progress 1

The following tasks have been completed: Selecting the dataset, Data Exploratory/summary statistics, Data Preprocessing, and Feature Engineering.

The first task completed was choosing a dataset–a prerequisite to the data exploration phase. Selecting an appropriate dataset is essential in gaining the proper insights into fraudulent transactions. It's important to note that, privacy concerns associated with financial transactions disallow the publication of big datasets containing fraudulent transaction metadata. As such, synthetic datasets derived from transaction simulators that mirror real-world financial transactions are made. One such dataset called “Financial Fraud Detection Dataset” was selected by the students. To quote, “Derived from a simulator named PaySim, which utilizes aggregated data from actual financial logs of a mobile money service in an African country, this dataset aims to fill the gap in publicly available financial datasets for fraud detection studies. ”

Subsequently, the data exploration task was completed. The dataset contains 6362620 rows and 11 columns. The column's features are type, amount, nameOrig, oldbalanceOrg newbalanceOrig, nameDest, oldbalanceDest, newbalanceDest, isFraud, isFlaggedFraud. There are 3 categorical features: type, nameOrig, nameDest, and the rest are numerical features. The type feature represents the type of transaction that took place, the amount feature is the transaction amount in the local currency, nameOrig is the customer who initiated the transaction, and nameDest is the recipient of the transaction. The oldbalanceOrg feature is the initial balance before the transaction, and newbalanceOrig is the balance after the transaction. The oldbalanceDest feature is the recipient's balance before the transaction, and newbalanceDest is the recipient's balance after the transaction. Now, isFraud is the target variable represented by binary values 0 and 1, for whether the transaction is fraudulent or not fraudulent. Lastly, the isFlaggedFraud feature, to quote, “Flags large-scale, unauthorized transfers between accounts, with any single transaction exceeding 200,000 being considered illegal.”

After, exploring the data, the data was preprocessed. To preprocess the data the students checked for missing values, and checked whether or not the data set was balanced. The dataset had no missing values, however, there was a great class imbalance among the approximate 6 million samples of the dataset, with only about 8500 samples being fraudulent cases. This gives a ratio of approximately 705 non-fraudulent cases to 1 fraudulent case.

After preprocessing the data, the next step of the pipeline was to perform feature engineering. First, visibly noisy columns were removed. These were the name columns–nameOrig, nameDest. The specified columns were the ID values of the sender and the receiver of the transactions. Afterward, the remaining(s) categorical variables were encoded using label encoding. Next, the mutual information gain score of the features and target variable were calculated. At a high level, mutual information shows the dependency of one variable to another. In the context of machine learning, It can reveal the importance of each feature. A higher score

shows higher dependencies between the feature and the target variable. As such, noisy features can be removed. After obtaining the mutual information scores of the features the students refrained from removing such features and left them for further investigation throughout the pipeline. That is, they reserve the given task for the optimization stage later in the pipeline if necessary. The next task was to address the imbalanced dataset. Despite the large disparity, this is an accurate representation of fraudulent transactions to regular transactions, as most financial transactions that occur are not fraudulent. However, despite this accurate depiction, leaving the dataset in this manner can result in adverse effects when constructing our clustering models. Take K-means for example which starts by randomly initializing centroids. If a dataset contains a large class imbalance, centroids would by default have a greater chance of initializing in the vicinity of samples in the majority class. In addition, another drawback is that clusters may be dominated by the majority class. To resolve these potential problems, the students proposed to oversample the minority class to create a more balanced dataset and optimize the clustering task. Remark, oversampling has some drawbacks as well due to the addition of synthetic data points. Another notable obstacle was the sheer size of the data. With over 6 million samples, manipulation of the data, as well as visualizing it becomes a hard task. As such, the next stage of the pipeline was performing stratified random sampling. In this task, stratified random sampling was more appropriate than simple random sampling in order to preserve the class distribution. If left to a mere simple random sampling, there's no guarantee that the new sample dataset would contain the same distribution as the preceding dataset it drew samples from, since samples are taken at random in simple random sampling. However in stratified random sampling, the data is divided into mutually exclusive subgroups–called strata– and samples are taken from each stratum. The result is a sample dataset more representative of the parent set. With this stage complete, the next step of the pipeline was to perform PCA on the set of features. After obtaining the stratified samples, Principle Component Analysis was applied. In short, PCA is a dimensionality reduction technique for projecting features of a high dimensional space into a lower dimensional space whiles maintaining the variability of the original dataset. First and foremost, the input features are scaled to ensure that all variables contribute equally to the principal components. Then, using n = 2 as the number of dimensions after performing the dimensionality reduction. Here, 2 is preferable because it is more manageable than a higher number of n while still retaining a significant percentage of the data's variability. As a result, PC1 and PC2 are collected and PC1 covers the most variance in terms of features from the data set and PC2 would be second. The scaled PC1 and PC2 are then concatenated with the target variable. Finally, represent their relationship on the scatter plot.

The following tasks have not yet been done: Elbow method for optimal number of clusters, other model optimization(Hyperparameter tunning), Partitioning model with K-means, k-means++, agglomerative clustering, evaluation of model with silhouette score(bar graph to compare each model score), Interpretation of clusters, and seeing ranges in susceptibility to fraudulent transaction. Model fitting, optimization, tuning, evaluation, and interpretation of the result have not yet been worked on. These tasks required the dataset to be explored, preprocessed, and engineered–the tasks that were completed this week.

The following will be completed next week: Elbow method for the optimal number of clusters, other model optimization(Hyperparameter tunning), and partitioning model with K-means, k-means++, and agglomerative clustering. Now that the data has been explored, preprocessed, and engineered, next week's tasks are to construct the clustering models to partition the data, and to perform optimization techniques to the said models.