Assessment Task: Propensity Model for Anti-Money Laundering
Adam Fagan

In this task, the key focus is on identifying suspicious clients using transactional data. The process is structured into four main stages: **Load Data**, **Data Preprocessing and Exploratory Data Analysis**, **Feature Engineering and Modelling**, and **Evaluation**.

Load Data

First, I merged the four parts of the dataset, which were divided across two sheets in **part1** and **part2** files, to create a coherent dataset for analysis. I performed an initial exploration of each dataset to ensure compatibility for merging and identified any potential issues early on. Since there were no duplicate or missing `client_id` values, the join proceeded without any problems, such as mismatched or lost data. Both datasets shared the `client_id` as the key column, and I performed an inner join to merge them, ensuring the data combined correctly based on the shared `client_id`. This approach could have used a left, right, or outer join, but the result would not differ in this specific scenario.

Data Preprocessing and Exploratory Data Analysis

The data showed extreme outliers in transaction-related columns, especially in volumes for incoming and outgoing transactions, which suggested that accounts with high transaction values could be involved in suspicious activity. Key findings included the presence of high-risk transaction behaviors in features like international transactions, where suspicious clients often exhibited unusual patterns. However, the presence of flagged clients within typical ranges complicated the identification of outliers, as not all unusual behavior correlated with flagged clients. Additionally, other features like client age and account age showed no significant outliers, indicating they were not useful for detecting suspicious clients. In terms of distributions, transaction volumes and client income were heavily skewed to the right, with most clients showing low transaction volumes and income, while a small number of high-value clients distorted the mean. Gender and nationality distributions suggested only slight differences in suspicious activity across groups, with a marginally higher rate of suspicious activity among females and Slovak nationals. Importantly, transaction behavior and client demographics showed no correlation, emphasizing the need for a multi-feature approach to detecting anomalies.

Feature Engineering and Modelling

Technique of Feature Engineering - Added New Features

For the first feature engineering technique, I created new features to capture more nuanced patterns in the data. These included the **Incoming-Outgoing Ratio**, which assessed the balance between incoming and outgoing transactions to identify unusual activity; the **Cash Transaction Ratio**, which highlighted the proportion of cash transactions as a potential risk indicator; the **High-Risk Ratio**, evaluating the frequency of high-risk transactions within international activities; the **International Volume Ratio**, which captured the proportion of international transaction volume to flag unusual international behavior; and the **Normalized Risk Score**, which standardized risk scores relative to client income for fairer comparisons across clients. Each of these new features was utilized with the three selected models: **Autoencoder**, **One-Class SVM**, and **Isolation Forest**. These models were trained separately within this feature engineering technique.

Technique of Feature Engineering – Encoding Categorical Columns

For the second feature engineering technique, I applied **One-Hot Encoding** to categorical features such as gender and nationality to avoid introducing ordinal or numerical bias. This method converted categorical variables into binary indicators, enriching the dataset with more granular features and eliminating implicit ordinal relationships. This encoded dataset offered a clearer representation of group-specific behaviors, which is critical for anomaly detection. The encoded features were paired with the same three models: **Autoencoder**, **One-Class SVM**, and **Isolation Forest**, with each model trained separately to detect anomalies.

Technique of Feature Engineering – Only Selected Features

For the third feature engineering technique, I focused on a reduced set of features that showed the greatest variation between flagged and non-flagged clients. This streamlined the dataset by retaining only the most informative variables while ensuring computational efficiency. These selected features were instrumental in emphasizing transaction-related patterns critical for detecting anomalies. As with the other techniques, I trained each of the three models—**Autoencoder**, **One-Class SVM**, and **Isolation Forest**—separately using the selected feature set.

By systematically applying each model within every feature engineering technique, I ensured that the modeling approaches were rigorously evaluated to optimize anomaly detection performance across different perspectives.

Evaluation

The models were evaluated using several techniques: PCA visualization, silhouette scores, and cluster purity. PCA was used to reduce the dataset to two dimensions, allowing for a visual assessment of how well flagged suspicious clients overlap with the top 100 most suspicious clients predicted by the models. Silhouette scores measured the closeness of each data point to its assigned cluster, while cluster purity assessed the alignment between the top 100 predicted clients and the flagged clients. Among the models, the Isolation Forest with categorical features performed the best. It had a silhouette score of 0.2374, indicating good overlap between flagged and predicted suspicious clients, and a high cluster purity score of 0.7745, ensuring that the top 100 clients closely resembled the flagged group.

In conclusion, while transaction-related features are the most critical for identifying suspicious behavior, combining multiple features and using advanced machine learning techniques, such as Isolation Forest, provided the most effective results in detecting high-risk clients. The best model demonstrated strong alignment with flagged suspicious clients, confirming its robustness for anomaly detection in this context.