**MMF Assignment for AML/ATF Models and Analytics Lecture**

This assignment is for you to build an Anti-money laundering (AML) client risk rating (CRR) model. A CRR model is a model that assigns a rating (high, moderate, low) to a customer at the time of customer onboarding and for periodic updates; the rating is from AML perspective, and considers customer KYC (know your customer) information and transaction patterns.

**Model**

You are required to form a group of 4 people and build a model that assigns a rating to a client according to the features provided (refer to the separate data file). If your model output is a score, you need to decide how to set a threshold to classify a client as a model high or a model non-high. Although the assignment is designed for building a statistical or machine learning model, if you’re not familiar with this area, feel free to build a judgemental one based on your research; please clearly state your rationale and use analysis to support whenever necessary.

**Data**

Data is in the file data\_std.xlsx. A data dictionary is enclosed in the same file for you to understand the meaning of the data elements. All features have already been standardized. Below is a summary of the type of data elements.

* ID (cust\_id\_masked): Unique Customer ID
* Target (rating): This is the target of the model, or the true rating assigned by adjudicators. The 3 categories (high, moderate, and low risk) have been simplified to high and non-high, therefore, a binary classification model would suffice.
* Features: There’re 149 features provided in the data. You don’t need to include them all in the final model. You can (i) perform feature selection by machine learning methods, or (ii) make judgemental decisions on inclusion or exclusion according to your understanding of each feature’s impact on a client’s AML risk.

**Report**

Please submit a report of your model techniques and outcomes. Please also answer the questions below when you write the report. Answers to some of the questions are not unique and could depend on the type of model you choose to build. You can use any programming language you prefer.

**Questions**

(If you’re building a judgemental model, ignore question 3)

1. Is it necessary to transform the data, and why? If yes, how do you transform?
2. Is it necessary to do feature selection, and why? If your answer is yes, what techniques do you use and why? (what problems do these techniques address?)
3. How do you split the data into training and testing dataset? Do you consider sample weight, and why?
4. What model do you choose? What’re the hyperparameters in the model (if any), and how do you decide the values of them? How do you cut the threshold for high and non-high.
5. How do you evaluate the performance of your model? Hint: you can evaluate on the model score (e.g., ROC AUC) or on the model-predicted label (e.g., precision, recall, F1 score)
6. How do you interpret the model? Which are the important features, and how they drive the model outcome. Is there any feature with counter-intuitive behavior? For example, higher risk occupation should indicate higher client risk rating, but the data or trained model might indicate the opposite. Can you think of a reason for this kind of problem, and how do you address it?
7. Are there any other issues you observe or actions you take to address potential issues while train the model? If yes, please describe them, and state the reason and solution.