***Momentum Insurance Company*** has been ruling the personal vehicle insurance market for many years in the US. Till this time, they have relied on the traditional judgment of their business experts like underwriters and claims managers in selling personal vehicle insurance policies, evaluating which claims maybe fraudulent and what may be the expected final claim payout once it is reported.

However, with the massive growth in business over the last decade and cut-throat competition with more customized products as well as insurance provision in the gig economy, there is a need to overhaul the products portfolio and drive innovation. They need their business experts to take up more innovation responsibilities, but first they need to relieve the experts from their current responsibilities of manually evaluating the entire policy and claim lifecycle. So, the owner of the company, Mr. Shylock William, being the shrewd businessman he is, has turned his eyes to **machine learning**.

Shylock has reached out to you for help and has given you a lot of data on active policies and claims paid. He has a problem in mind, which is described as follows:

1. Predicting whether a claim would be referred for fraud investigation or not? [SIU\_Referral\_Flag = “Yes” in the claim\_details\_data.xlsx file]

**Key understanding of the data:**

All the datasets have been described in the Data\_Dictionary provided. However, one key point is each policy and policyholder who is purchasing the vehicle insurance is identified by “Poilcy\_ID”. However, some policyholders drive well and do not have any accidents or claims while others may have multiple. As a result, many Policy\_IDs would have no claims reported while others may have multiple claims. Unique ID for claims is “Claim\_ID”.

**Guidelines and Evaluation criteria:**

1. Please clearly lay out the mathematical problem you are trying to solve/optimize
2. Clearly lay out all your assumptions
3. Even if you have a brilliant method of connecting external data sources to this database, please resist yourself
4. Whichever problem you choose to solve, feel free to use any number of datasets at your disposal (in the zipped folder) if it helps your analysis
5. Please submit all your codes used to do the analysis
6. What you will deliver:
   1. A few slides capturing the problem to solution
   2. Present before the panel and convince them about your way of using the solution
   3. There is no test data separately provided. Please use all the data at your disposal to develop the model and prove its accuracy/interpretability/robustness
7. ***Evaluation Criteria:***
   1. There can be multiple ways to reach a solution, so, some weight will be given to your solution design aesthetics
   2. Code functionality
   3. Presentation
   4. Business usability of the solution
8. **Dataset Query Resolution**: Please send al the consolidated queries to Novartis SPOC no later than 2PM IST Thursday (9th Sept); the team will get back to you by EOD with answers to all queries.