**Data Science Exercise**

Imagine it is March 2023 and you are working for HubPay. The attached file (Customer\_Churn\_Data\_v2.xlsx) contains mocked-up data on retention versus churn (departure) of customers form HubPay, along with a set of data definitions. All retail service providers experience churn, and there is competitive advantage to be gained from predicting the customers that will depart and taking steps to retain them.

Your objectives (using either R or Python) are to:

1. Perform basic descriptive analytics to understand whether any of the features in the data are associated with churn;
2. Prepare and cleanse the data to make it suitable for modelling;
3. Build classification models to predict churn, using at least two alternative machine learning techniques, and perform appropriate validation upon these models;
4. Evaluate the performance of the models and any shortcomings that are evident, and opine as to whether the models are adequate for decision-making;
5. Build logistic regression model(s) explaining churn in terms of the explanatory variables and provide interpretations of coefficients and coefficient standard errors in these model(s); and
6. Include this information in a short, well-presented report or slide deck. **Send the report/deck and your code file back to me within seven days of receiving the email that carried these instructions.**

**Hint:** this is not (!) the type of data science challenge where the winner is the person with the highest model performance, like Kaggle. Rather, we aim to verify that you have the analytical and coding skills you talked about in your interview, and to see how you deal with data that is limited in various ways.