**S&P Global Market Intelligence – Senior Data Scientist Take home assessment**

Please try to complete at least one of the exercises below and send your results (including answers to the questions listed and code) to [saraansh.arya@spglobal.com](mailto:saraansh.arya@spglobal.com) within 72 hours of receiving the assessment. In your interview we would like to talk about your solution/results and the process you used. If you don’t have time to finish the exercise, please send us whatever you’ve got and come to your interview prepared to talk about the task and chosen approach.

**EXERCISE 1**

The Bankruptcy dataset is made up of financial ratios pertaining to companies. One row belongs to one company and the definition of each column has been provided in the sheet ‘*column\_definitions*’. For each row there is a target class ‘*bankruptcy?’* which mentions whether this company went bankrupt within the next 5 years or not. In this exercise, we would like you to create a model on this data to predict whether a company would go bankrupt within the next 5 years. Additionally, please answer the below questions:

* Can you predict the class label of a sample based on its features?
* What is the accuracy of the classifier on the test set?
* How does the classifier perform for each class label?
* Can you identify the most important features for the classifier?
* Can you explain why the classifier misclassified a specific sample?
* How well does the classifier generalize to new, unseen data?
* Can you compare the performance of the classifier with other models?
* How does the performance of the classifier change with different hyperparameters?
* How does the classifier perform on imbalanced datasets?

**EXERCISE 2**

The Movie Review dataset is made up of movie reviews. For each review there is a target class mentioned which reflects whether a viewer gave a positive review to the movie or a negative one. In this exercise, we would like you to train a model on this data to predict whether a new review is positive or negative. Additionally, please answer the below questions:

* What is the accuracy of the classifier on the test set?
* How does the classifier perform for each class label?
* What model did you use and why?
* What major challenges did you face while working on the task?
* Can you explain why the classifier misclassified a specific sample?
* How does the performance of the classifier change with different hyperparameters?
* Are there any other observations that you had while working on this task?