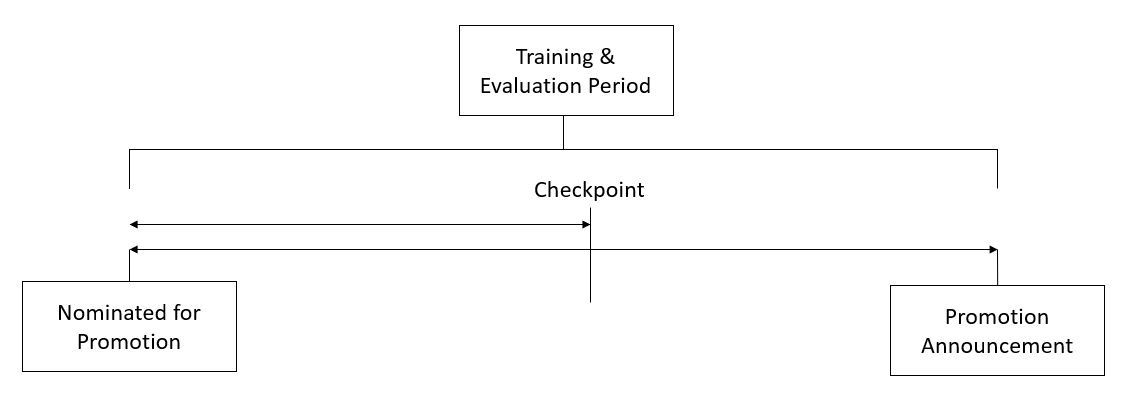
The problem statement below is the project from the website I worked on.

Problem Statement

Your client is a large MNC and they have 9 broad verticals across the organization. One of the problems your client is facing is around identifying the right people for promotion (only for manager position and below) and prepare them in time. Currently the process, they are following is:

1. They first identify a set of employees based on recommendations/ past performance
2. Selected employees go through the separate training and evaluation program for each vertical. These programs are based on the required skill of each vertical
3. At the end of the program, based on various factors such as training performance, KPI completion (only employees with KPIs completed greater than 60% are considered) etc., employee gets promotion

For above mentioned process, the final promotions are only announced after the evaluation and this leads to delay in transition to their new roles. Hence, company needs your help in identifying the eligible candidates at a particular checkpoint so that they can expedite the entire promotion cycle. 

They have provided multiple attributes around Employee's past and current performance along with demographics. Now, the task is to predict whether a potential promotee at checkpoint in the test set will be promoted or not after the evaluation process.

Summary of the project:

Why this project?

I am very interested in data science and am constantly trying to learn more about it. This hackathon project seemed like a good way to get my feet wet into the world of data science. I enjoyed working on it, and I might continue in my spare time to try and improve my final score, but I ended it so I could better pursue other certifications and pursuits.

What did I do for this project?

I had to learn about the R programming language in order to visualize the data and understand how the variables all relate to one another. I had never used R before, but It was recommended to me by a recent data scientist that I had been speaking too. While the whole project could have been utilized within R, ultimately, I am more familiar with Python and it’s more relevant to most people. After I did the data visualization with the graphs and charts, I started applying what I learned to weight the independent variables. First, I did multiple iterations of what weights each variable should have on Excel with the training data. Since I was able to get the promotion percentages by department, it was easy for me to determine how many people form each department should get promoted on the training data. Therefore, I was able to do some tests just using a little bit of VBA from Excel.

Once I felt I had a sufficient amount of knowledge from the initial tests, I started applying various machine learning classification models to the training data. I picked the best one, which was XGSBoost, and I decided to try and fine tune it a little bit more. With it being my first time doing this, it was hard to fully understand all the parameters, so I did a little bit of research and asked some people with more experience what would be best to tweak. It took some playing around and a lot of runs, but I was satisfied with the tweaks that were made.

What are the final results?

There was a lot of outside promotions that really affects the data. For example, approximately 20% of the promotions in technology didn’t seemingly meet any of the requirements, yet they got promoted. For reasons like that, the overall scores are pretty low on this hackathon for everyone participating. I am currently at 1818 out of the 8200 participants at this point.

What did I learn from this project?

I learned a lot about R. It is such a nice statistical language that is pretty straightforward to learn. I utilized Spark in Python in the beginning of this project. That was hard to set up, but I got that running within Pycharm. The experience with Spark made me realize it was a little overkill for this data set, so I ultimately took it out of the project and focused more on the pandas and sklearn packages. I learned a lot about data analysis and being able to filter out what is important. It was also a lot of fun.

The MNC’s efficiency for promotion would improve with this algorithm. However, if I was recommending it to them, I would caution them that they could potentially overlook many candidates who they would have selected in the past. The algorithm can’t consider all of the intangibles that candidates have, which would make it approximately an 80 percent solution.