**Phishing Attacks**

**XYZ Corporation** is a credit card company. They are having difficulty controlling their fraud detection and recovery costs, especially through phishing attacks. It is important that XYZ understand first, who is likely to be attacked and why, and second, activities they can engage in to start decreasing the volume of attacks they receive.

They have collected data on their on their email phishing attacks that they have found so far and would like to use this dataset to predict which employees are being targeted and why. The dataset contains records for email users stretching back for one year.

The dataset has been obfuscated to prevent any leak of IP or identities from our analysis, and thus the column variables will be general in nature. The columns included were what the IT leads could put together is such short notice, but they are trying to get more data in the coming months. The business leaders of XYZ company would like to understand the correlations if at all possible, especially to understand which variables they should be looking at and if there are any more that would be worthwhile to try to get for future attempts, and more importantly, understand what actions to take from a company perspective.

**The variable that XYZ would like you to try to build a model around is labeled “EmployeeTargetedOverPastYear”.**

**Please consider the following requirements for this challenge (If you are unable to meet these requirements, it will result in an automatic non-certification):**

* 1. **Please do NOT use any form of AutoML for your models**
  2. **(others)**

**Additionally, XYZ company has a holdout dataset that they would like to put through your model(s) in 24 hours, during the presentation to them, so please be ready to exercise your model(s) accordingly.**

Over the course of the next 24 hours, please utilize either Python or PySpark to create a model that will be able to predict the above variable. Using the findings from that model, please create a slide deck for XYZ Corporation that will address the questions of both their Analytics team and their leadership. Because the audience will be both technical and non-technical, please address both needs during your presentation.

Finally, please remember this challenge is confidential and the certification and all associated work is individual.