Employee Turnover

XYZ Corporation is a sportswear company. They are having difficulty controlling their employee turnover rates. It is important that XYZ understand first, who is likely to be leave the company and why, and second, activities they can engage in to start decreasing the volume of employees that leave. It is estimated that each employee that quits represents a cost of 300% of their annual salary.

They have collected the data on their employees that they have found so far and would like to use this dataset to predict which employees are at risk of leaving and why.

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 – you will notice that the survey data is especially piecemeal as they try to pull things together. However, despite the less-than-ideal state of this dataset, we still have a great opportunity to get Data Science front-and-center at the highest levels of XYZ leadership as the recommendations we make will be presented to the board of directors – so please do try to drive to as significant of conclusions as possible.

In addition, the business leaders of XYZ company would like to understand causality 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.

The variable that XYZ would like you to try to build a model around is labeled "EmployeeLeft".

<u>Please consider the following requirements for this challenge:</u>

--- Explainability: Please prepare to explain your model performance in terms of the variables given.

--- Bias: It is important that we know of any potential bias present in this model, please prepare to present on the impact of this model on any protected classes.

--- Deployment: Please deploy your models and be ready to demonstrate a live(online or local) API call to those models

XYZ company has a holdout dataset that they would like to put through your model(s) in 24 hours, 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 powerpoint or Jupyter presentation 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 in the course of your presentation)