EDSP Mentoring
Preparing for the Challenge

Microsoft

Agenda

- Announcements
- Data Science Challenge
- Tips for Success
- Sample Challenge
- Questions & Discussion

The Data Science Challenge

Announcements

- Congratulations to James Xu and Vishnu Pamula the winners of the Mentoring Program Leaderboard! They have Amazon Gift Cards coming their way!
- Thank you to all the mentors, you also have Amazon Gift Cards coming your way!
- Please complete the Post-Program Survey (this one is short we promise!)

Microsoft Data Science Challenge

What is it?

- A Challenge that must be solved, using Python or PySpark in 24h by training a Machine Learning Model.
- It is done in a role play scenario, as if it was a real customer problem – based on a real situation.
- The challenge must not be solved using AutoML or ML Designer

Challenge Phases:

Phase 1

- A candidate receives the issue dataset and a business prompt. They then have 24 hours exactly to build a model using python to achieve the objectives of that prompt and to build a PowerPoint presentation detailing their work on that model.

24 hrs.

Phase 2a

- The day following the initial 24 hours, the candidate presents their findings to their challenge proctor in a simulated role-play scenario (30 minutes), including a quick walkthrough of the code (5 minutes).

30 + 5 mins

Phase 2b

- The candidate has 15 minutes to score a holdout dataset against the model that they have already built. The candidate must be ready for the holdout dataset to have differences compared with the initially provided dataset.

15 mins

Phase 2c

- The candidate will then have 5 minutes to receive feedback on their performance and ask any questions they may have.

5 mins

Your proctor will remind you about the agenda at the day of the presentation and keep track of the time!

**Randomly generated dataset*

Microsoft Data Science Challenge

Challenge Composition

- You will receive a dataset alongside with a set of mandatory requirements. They may include (but are not limited to):
  - Bias mitigation and FairLearn
  - Local and/or Global Explanations
  - Deployment
- Both the dataset and the challenge requirements are randomly generated
- Independent of your final score, you will need to complete the mandatory challenges to be certified

Evaluation

- A Certified proctor will run the process and score the challenge using a well-defined set of criteria
- You will be evaluated in 4 main areas:
  - Presentation
  - Business Understanding
  - Data Engineering
  - Data Science
- The goal is not to evaluate how good your model is, but your entire approach to solve the problem and your presentation

What to Expect

1. The challenge is offered Monday - Thursday. You have exactly 24 hours to complete the challenge

2. Register for the challenge: Challenge Registration Form at least 2-3 weeks prior to your target date to give time to schedule a proctor.   The earlier you register the better, so don't delay once you have a date in mind

3. At the top of the challenge registration form note the link to the Sample Challenge – be sure to complete this before pursuing the challenge

4. Once a proctor is assigned, they will contact you to schedule a kickoff and a presentation (24 hours apart). Besides the scheduling, your proctor will schedule/ask you to schedule a short briefing on the challenge logistics and expectations

5. During your kickoff meeting your proctor will provide the dataset and instructions on the submission. You will then have 24 hours to complete the challenge and prepare to present your results

6. For your final presentation, you will have **35 minutes to present your results**, and **15 minutes to score the holdout dataset**. During the presentation the proctor will ask questions from both a business and technical perspective. Note that the presentation will be recorded

7. You will receive feedback and your outcome within a few days of the presentation. If you do not pass, you can retake the challenge at least 30 days after the attempt. Your line manager will only be notified about a successful attempt. Failed attempts are not logged and only you and your proctor will know

Tips for Success

Tips for Success - Be Prepared

- Complete the sample challenge!

- Keep your coding efforts as minimal as possible during the challenge - rather “adjust” existing notebooks & code (from the sample challenge) to the new dataset

- Prepare your environment ahead of time and double check the day before your challenge

- Carefully read through the challenge instructions – each challenge has custom requirements

Tips for Success – Technically

- You must use either Python or Pyspark; Azure Machine Learning / Databricks is not required, but may be beneficial

- AutoML is not allowed in the challenge results/presentation

- Be prepared to explain the choices you make for algorithms, hyperparameters, metrics, data preparation.

- You will need to score a holdout set that may not look exactly like your training/test data, have your process repeatable enough that this is possible

- Don’t forget Bias mitigation, FairLearn, and Local/Global Explanations

Tips for Success - Presentation

- Do not forget the presentation part of the challenge! **50% of the points** are awarded for a compelling presentation that caters to both a business and technical audience.

- Be sure you can connect the Data Science work and modeling to business recommendations – think of it as Story Telling

- Explain any assumptions you make throughout your process and show your exploratory analysis of the data as well as the results

- Responsible AI and Model Interpretability are key – demonstrate and show how this is incorporated in your results

- Keep track of the time & follow a “good meeting routine”. While your proctor will remind you about how much time is left, it will help you pace your presentation.

Sample Challenge

Sample Challenge

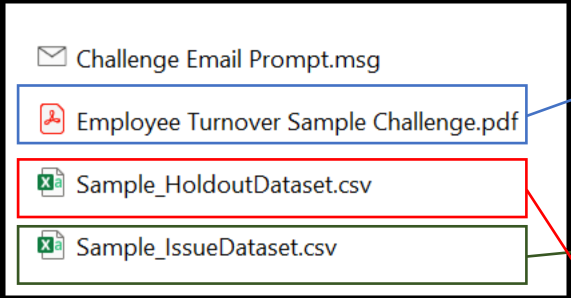
- Download the dataset and instructions for the Sample Challenge and review the sample email prompt and instructions carefully

- Complete the challenge in a repeatable fashion – think of this as a way to prepare your process, pipelines, and methods ahead of time

- Make sure you test evaluating your model with the holdout set provided

- If you would like to simulate a live presentation/feedback session, please contact your mentor to arrange for this. We may have proctors or one of the mentors do this with you

Sample Challenge

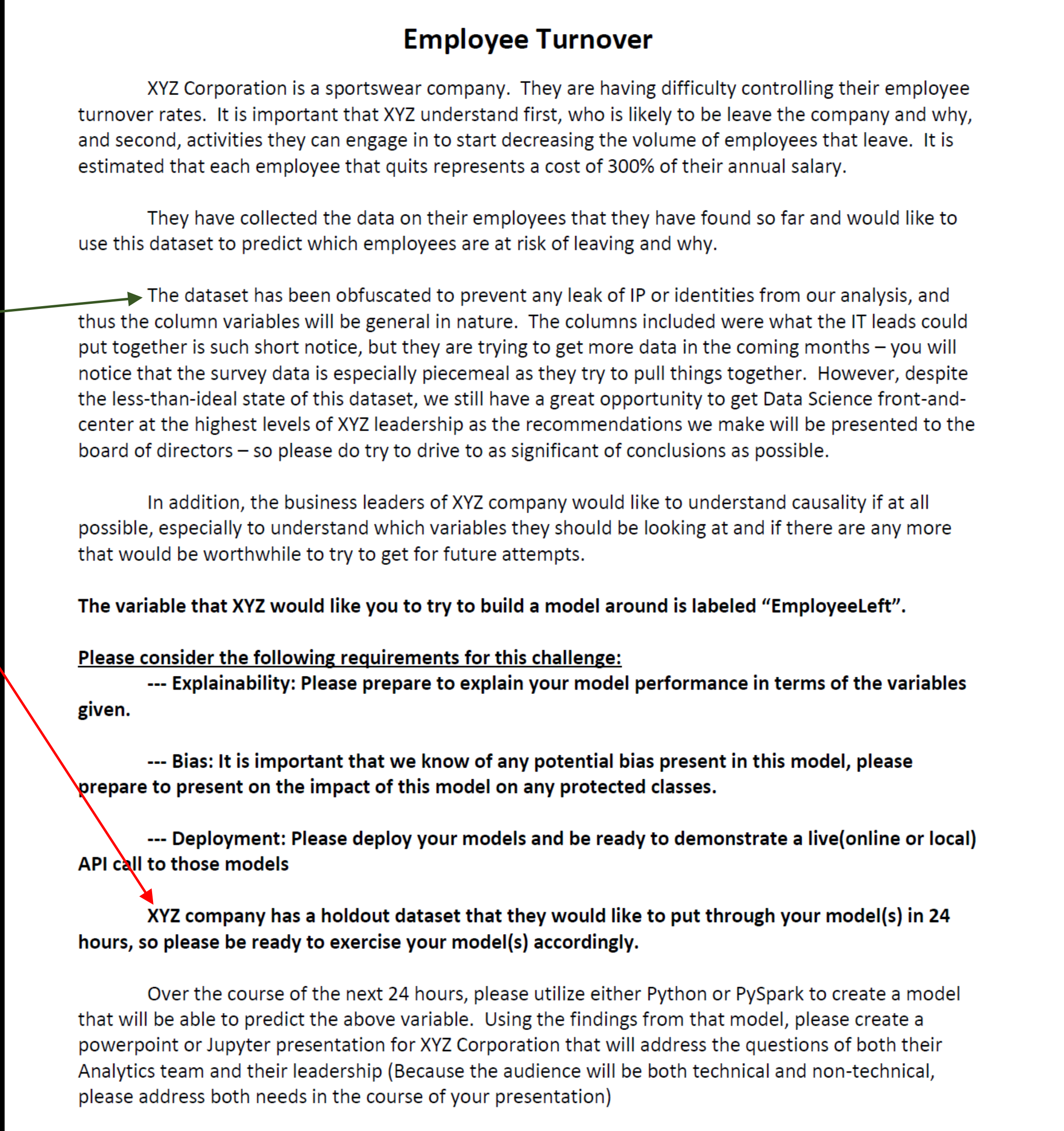


Challenge Email Prompt.msg

Employee Turnover Sample Challenge.pdf

Sample_HoldoutDataset.csv

Sample_IssueDataset.csv



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)

Questions & Discussion