**Project 5: Data Drift Experiment**

**CS4981 ML Production Systems**

**Overview**

Over the quarter, you implemented a spam classification system. In this last part, you are going to explore the impact of a phenomena called "data distribution shift" that occurs when the data being sent to the system radically changes. In this case, it can occur when spammers change their approach to work around spam classifiers. You will learn one way to deal with this problem.

**Instructions**

**Part I: Test It Using the Same Data**

1. In one terminal instance, run the prediction service

2. In a separate terminal instance, run the simulator to see how it performs:

(venv) $ python spam\_simulator.py evaluate-model --email-dir path/to/email\_data --classifier-url http://127.0.0.1:8888/ --number-emails 10000

**Part II: Test It Using Different Data**

1. Download the email\_json\_dataset2.zip file from the shard Box folder and unzip it:

<https://msoe.box.com/s/l9xl4udcwvqzrzwtxwrqw4bejehomk2b>

2. Evaluate the model using the second set of data:

$ python spam\_simulator.py evaluate-model --email-dir path/to/email\_data --classifier-url http://127.0.0.1:8888/ --number-emails 10000

**Part III: Ingest the New Data and Retrain the Model**

1. Start the ingestion service.

2. Run the simulator to add the second data set to the database:

$ python spam\_simulator.py simulate-user --email-dir path/to/email\_data --email-url http://127.0.0.1:8888/ --mailbox-url http://127.0.0.1:8889/ --number-emails 100000 --average-events-per-email 5

3. Retrain the model and restart the prediction service with the new model

5. Evaluate the updated model twice, once with each set of data:

$ python spam\_simulator.py evaluate-model --email-dir path/to/email\_data --classifier-url http://127.0.0.1:8888/ --number-emails 10000

**Part IV: Report**

Write a 3-4-page report that includes:

1. A summary of the overall system (with a diagram!)
2. Describe the details of your modeling approach (including your features, type of model, any model parameters used, etc.) and experimental setup (including your choice of metrics).
3. Describe the results of your experiment. How did the original model perform on the new data set?
4. What you observed is called "data distribution shift." This example is a bit extreme, but it is a real problem that operators of ML-powered services face. Come up with examples of other types of ML services that would have to deal with this problem and how it could impact users and the business.
5. How did the updated model perform on the two data sets? Did retraining the model help? Many production systems periodically retraining and deploy models on a regular basis (e.g., every 12 or 24 hours). Explain the utility of periodic retraining models to the continued efficacy of the system.
6. Reflect on the quarter. What did you learn? What went well? What could be improved?

**Submission Instructions**

Submit your final report to Canvas as a PDF.