Requirements: For each homework assignment you must create and submit a homework lab notebook. The notebook must include your name, the course (CPSC 324), and the assignment number (HW-5). Your notebook must be organized by step (see below), with each step number clearly marked (e.g., Step 1, Step 2, and so on). You may create your lab notebook in any editor/tool you like (e.g., google docs, microsoft word, latex, etc.), however, you must submit your notebook as a single PDF file named hw5-lab-notebook.pdf. Specific information to add to your notebook will be asked for in each step of the instructions. Additional details are provided below. Your notebook and any additional work (e.g., program files) must be submitted on or before the due date using the homework GitHub classroom repository provided via Piazza.

Instructions: Do the following steps and add the information requested to your lab notebook. Your lab notebook should be organized and well formatted. Note: You can (and should) look through the Google Cloud Skills Labs (part of some steps below) prior to starting them, which can help you get a sense for what the tasks are and each lab’s scope. You can also go back over the lab as well after you complete them and/or do them multiple times as needed (but watch out for those that require credits).

**STEP 0:** A number of the labs below use an ML classification approach based on logistic regression. If you are not familiar with logistic regression, take a few minutes to familiarize yourself with the idea and terminology. Here are some resources that can help.

- AWS provides a nice introduction to the basic ideas of logistic regression here: [What is logistic regression](#).

- Google has a similar resource (as a crash-course) for developers at: [Logistic regression (part of ML crash course)](#).

- Google provides a ML glossary, which defines a number of ML and related terms. The entry for logistic regression is [Logistic Regression], for sigmoid functions is [Sigmoid Function], and for log loss is [Log Loss]. Make sure you always “click to see the math”.

**STEP 1:** Within Google Cloud Skills Boost, complete the [Getting Started with BigQuery ML](#) lab. This is a free lab. Note the following concerning the lab before you begin.

- Be sure to do the “optional” part in the lab.

As you do the lab, add the following to your lab notebook for this step.
• In your own words, describe what the model information and training statistics provides (this is the optional step).

• In this lab you use three BQ ML functions: create a model, evaluate a model, and use the model for prediction. Write down, in your lab notebook, the basic syntax for doing each of these three functions.

• In your own words, state what feature is being predicted in the lab example and what features are used to build the model (i.e., to make the predictions).

• In your own words, state what each of the three steps (create model, evaluate model, prediction) above return.

• Modify the prediction query to SELECT all attributes returned (instead of the sum by country). You only need to select the first 10 rows. Write down the schema of what is returned by the query.

• Give a screenshot of the result of each step that was produced via the BQ ML commands in the lab.

STEP 2: Within Google Cloud Skills Boost, complete the Predict Visitor Purchases with a Classification Model in BigQuery ML lab. This is a free lab that goes over a few more examples than the lab in Step 1. Note the following before starting.

• This lab is a bit older, where at the time BigQuery ML only supported two different models. More are supported today.

• While the lab says it will take 5–10 minutes to create the models, it only took a couple of minutes each.

As you do the lab, add the following to your lab notebook for this step.

• Briefly describe each task in the lab in your own words. In particular, state what the goal of each step is, and for each step, what each query/command is doing.

• For Task 5 and Task 6, write down the date ranges used to train and evaluate the model. How many rows (instances) were used for the training set and testing set?

• Give a screen shot of the first few rows of the final result table produced.

STEP 3: Within Google Cloud Skills Boost, complete the Predict Taxi Fare with a BigQuery ML Forecasting Model lab. This lab requires 1 credit. Note the following before starting.
• As in the previous lab, this lab is a bit older (but says there are three types of models!).

• Similarly, it took only a couple of minutes to train the model.

As you do the lab, add the following to your lab notebook for this step.

• Give the number of rows that are returned in the training set query of Task 3.

• After you train the model, select the model, then select the schema. Take a screen shot of the schema and paste it into your notebook.

• Write down the number of rows produced by the evaluation query in step 6.

• Try different subsets of the data and check the resulting RMSE for Task 6. For instance, try a specific number of passengers, specific days, specific fare ranges, and so on. Try 4 different subsets, and for each write down what you tried and give a screenshot of the corresponding RMSE.

• In your own words, explain the difference between the first model trained and the second as well as their performance (evaluation) differences.

STEP 4: Within Google Cloud Skills Boost, complete the Bracketology with Google Machine Learning lab. This lab requires 5 credits. As you do the lab, add the following to your lab notebook for this step.

• Write down the number of seasons worth of data that are returned by the query in Task 4 as well as how far back in time the table goes.

• Describe the training data used in Task 7 to build the model, including the number of rows (instances) and the features used.

• Paste the training table generated for you in Task 7.

• In the last step of Task 7, you generate a series of weights for the seed attribute. Modify the query to generate the weights for school_ncaa (as hinted at in the query itself). Give the results and explain what they are saying. What is Gonzaga’s weight?

• Modify the query in Task 10 to only return Gonzaga’s games in the tournament. Paste the result in your notebook.

• Write down the number of rows resulting from the query in Task 12.

• Write down the number of rows used to train the model in Task 15.

• After Task 19, rewrite the query to show all the predictions made regarding Gonzaga. Give the result in your notebook.
• After Task 21, take a moment to reflect on using SQL for querying the predictions as a way to understand the model. What are your general impressions? What are possible advantages and disadvantages to this approach?

• Rerun the last query in Task 22 for Gonzaga. How did the predictions turn out for 2019? You can get the results from here: https://www.ncaa.com/brackets/print/basketball-men/d1/2019

**STEP 5:** Within Google Cloud Skills Boost, complete the **Fraud Detection on Financial Transactions with Machine Learning on Google Cloud** lab. This lab requires 5 credits. As you do the lab, add the following to your lab notebook for this step.

• Paste the result of the first query (Task 4) into your notebook.

• Write down your answers to the “PAUSE and REFLECT” questions at the end of Task 5 in your notebook.

• Write down the number of rows in each table created in Task 5.

• Give the resulting metrics for your model in Task 6.

• Write down your answer to the “PAUSE and REFLECT” question for Task 6 in your notebook.

• Write down your answer to the “PAUSE and REFLECT” question for Task 7 in your notebook.

• Note that the decision tree model in Step 9 took much longer to train than the previous models (around 6 minutes).

• Write down your answer to the “PAUSE and REFLECT” question for Task 9 in your notebook. Note there is a significant difference between the two models.

**STEP 6:** Within Google Cloud Skills Boost, complete the **Building Demand Forecasting with BigQuery ML** lab. This lab requires 1 credit. Before you start the lab, note the following.

• This lab walks you through a time-series forecasting example using an autoregressive integrated moving average (ARIMA) model.

• An overview of ARIMA is provided in the **The CREATE MODEL statement for ARIMA PLUS models** documentation page.

As you do the lab, add the following to your lab notebook for this step.
• Note that while saving the query in Task 2, after entering the dataset name, you need to select “Create Dataset”.

• At the end of Task 3, select the model you created, click on the Evaluation tab, and paste a screen shot of the evaluation table into your notebook.

• Rerun the last query in Task 5 with a smaller horizon (e.g., 2 days) and a higher confidence (e.g., 95%). Paste a screenshot of the query and the resulting predictions.

STEP 7: Create your own model using BigQuery ML, evaluate it, and use it to make predictions for one of the datasets we’ve previously used in the homework assignments. Explain each step of the process:

1. The data set you selected;
2. The label and features you are using;
3. How you are diving the dataset into train and test sets and the corresponding queries you used;
4. The result of the evaluation (via the ML.EVALUATE function);
5. The data you used for making predictions and the prediction results (note you need to save the prediction results as a separate table; and
6. The queries you ran over the prediction results (to understand what the quality of the predictions).

Give the queries you used for each step and the results in your notebook.

STEP 8: Within Google Cloud Skills Boost, complete the Vertex AI: Qwik Start lab. This lab requires 1 credit. Before you start the lab, note the following.

• Take a couple of minutes and watch the following video, which provides a brief overview of VertexAI: https://www.youtube.com/playlist?list=PLIivdWyy5sqJ1YuMdGjRwJ3fFYZ_vWQ62

As you do the lab, add the following to your lab notebook for this step.

• Note that it takes a few minutes for the instance to spin up in Task 3.
• You need to install graphviz for cell 30 to work. You can do this as you are in the terminal.
  – To install graphviz type: sudo apt install graphviz
– To install python graphviz module: pip install graphviz

• Note that cell 8 (downloading data) in the python notebook is finished once it prints “... [INFO] Loaded 3330 rows into ...”.
• Paste the recency histogram generated from cell 11 into your notebook.
• Paste the frequency histogram generated from cell 14 into your notebook.
• Paste the target monetary plot generated from cell 17 into your notebook.
• Paste the baseline plot generated from cell 21 into your notebook.
• Paste the resulting plots generated from cell 23 into your notebook.

**STEP 9:** Within Google Cloud Skills Boost, complete the **Deploy a BigQuery ML Customer Churn Classifier to Vertex AI for Online Predictions** lab. This lab requires 5 credits. As you do the lab, add the following to your lab notebook for this step.

• Describe the basic idea of each of the queries and their purpose in your lab notebook (up to and including Cell 18).

• Note that the decision tree model took around 10 minutes or so to finish.

• Paste into your lab notebook the result of the ML.EVALUATE call in Cell 22 (after creating the model).

• Paste into your lab notebook the result of the ML.CONFUSION_MATRIX call in Cell 23.

• Paste into your lab notebook the result of the ML.GLOBAL_EXPLAIN call in Cell 26.

• Briefly explain what the table in Cell 26 is saying.

• The rest of the steps from “Upload BQML model to Vertex AI from GCS” on are optional. These steps go through the process of uploading the model saved to Google Cloud Storage to Vector AI via XGBoost and then running it via the XGBoost python libraries. However, this part seems very finicky in the Jupyter notebook.

**STEP 10:** Within Google Cloud Skills Boost, complete the **Summarize Text using SQL and LLMs in BigQuery ML** lab. This lab requires 1 credit. As you do the lab, add the following to your lab notebook for this step.

• Note to locate the github_repos dataset, you first need to “star” the bigquery-public-data project.
• Note that I had to also add bigquery.connectionAdmin to the account (as well as the Vertex AI User).

• Look over the results in the generated text column, and provide your observations in your notebook on what the LLM returned (i.e., in the generated_text column).

• Try changing the prompt string and checking the results again. (As an example, I changed the prompt to “Can you read the code in the following text and generate a joke based on what the code is doing and what language it is written in.”.) Write down in your notebook what prompts you came up with and give some examples of what the results were.