TEACHING ABCS TO BUSINESS STUDENTS

Undergraduate core and Masters in Business Analytics

ALOK CHATURVEDI, ALOK@PURDUE.EDU
Teaching ABCs

Artificial Intelligence

Undergraduate Core

Big Data

Cloud Computing

Graduate BAIM
Objectives

- Inspire students to be **fearless**
- Overcome **Technophobia**
- Develop **learning to learn** skills
- Learn tangible skills that can improve job prospects
  - Google Cloud platform

Tools

- Cloud Storage
- BigQuery
- Auto ML
- Data Studio
#standardsql
CREATE OR REPLACE MODEL `bts_data.ontime`  
OPTIONS
(model_type='logistic_reg', input_label_cols=['on_time']) AS 

SELECT  
   IF(arr_delay < 15, 1, 0) AS on_time, 
   carrier, origin, dest, dep_delay, taxi_out, distance
FROM
`datapipe-1.bts_data.flights_unpart`
WHERE 
arr_delay IS NOT NULL

#standardsql
SELECT * FROM ML.EVALUATE(MODEL `bts_data.ontime`,  
( 
SELECT  
   IF(arr_delay < 15, 1, 0) AS on_time, 
   carrier, origin, dest, dep_delay, taxi_out, distance
FROM
`datapipe-1.bts_data.flights_unpart`
WHERE 
arr_delay IS NOT NULL
))

SELECT * FROM ml.PREDICT(MODEL `bts_data.ontime`,  
( 
SELECT  
   'AA' as carrier,  
   'DFW' as origin,  
   'LAX' as dest,  
   dep_delay,  
   18 as taxi_out,  
   1235 as distance
FROM
UNNEST(GENERATE_ARRAY(-3, 10)) as dep_delay
))
Tech Show Video

Sophia AndreOTTi, Rajvi Desai, Brandon Diltz, Haley Johnston, & Sarah Panikkacherry
Objectives

- Inspire students to be **fearless**
- Develop **learning to learn** skills
- Learn to build **end-to-end complex data pipelines**
Building Data Pipeline

Batch Data Pipeline

Batch Data

(CLOUD FUNCTION > STORAGE)
Building Data Pipeline

Batch Data Pipeline
Building Data Pipeline

Streaming Data Pipeline

Cloud Function -> Cloud Storage -> Cloud Dataflow -> Cloud Pub/Sub -> Cloud Dataflow -> BigQuery -> DataStudio

Data Stream
Building Data Pipeline

Streaming Data Pipeline
ML ANALYSIS ON BATCH AND STREAMING DATA
Question 1: Most major airlines operate across the globe and have agreements with other regional airlines that further extend their reach. Still, each airline has a geographic area that they operate within the most. But which airlines operate in similar spaces?

Question 2: How predictable are itineraries? Can we tell the origin of a flight from the airline, time and day of week?

Before modeling, let’s add two fields, one for the **Airline code** and one for the **day of the week**, using the following SQL expression:

```sql
SELECT *
FROM flightsapilr.flightsdata
```

Let’s store this as query “fightsdata2”
We can build a k means clustering ML model to answer Question 1:

```sql
CREATE OR REPLACE MODEL `flightsapilr.kmeans`
OPTIONS (model_type='kmeans', num_clusters = 2) AS
SELECT latitude, longitude, airline, origin
FROM `flightsapilr.flightsdata2`
WHERE spi != false
```

The number of clusters was selected through iterative trials to minimize the Davies-Bouldin index and MSD and maximize cluster interpretability. The values tried were integers 2 through 7.
• Big Query: ML Clustering
Big Query: ML Clustering

The results from building this model are:

**Metrics**

- **Davies-Bouldin index**: 1.7234
- **Mean squared distance**: 1.2103

<table>
<thead>
<tr>
<th>Centroid Id</th>
<th>Count</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>296</td>
<td>35.1396</td>
<td>-92.3316</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>-36.9986</td>
<td>143.0028</td>
</tr>
<tr>
<td>3</td>
<td>57</td>
<td>41.4898</td>
<td>35.1548</td>
</tr>
<tr>
<td>4</td>
<td>214</td>
<td>37.8800</td>
<td>-92.9642</td>
</tr>
<tr>
<td>5</td>
<td>154</td>
<td>44.2159</td>
<td>-98.0174</td>
</tr>
</tbody>
</table>
Summary

- Challenge the students
- If you are fearless, students will be fearless
- Focus on learning to learn
- Zoom actually helps in peer-to-peer as you can ask the students to share their screens
- Working with large, real world data helps UG students grow cognitively
- Masters students appreciate what it takes to prepare data before you can use sexy ML models
References

- The Building Blocks of a Modern Data Platform

- Features in BigQuery’s New UI for 2021
  - https://towardsdatascience.com/5-great-features-in-bigquerys-new-ui-for-2021-yes-it-has-tabs-c4bac66d66b

- BigQuery SQL Cheat Sheet

- BigQuery — Almost All You Need to Know
  - https://medium.com/swlh/bigquery-almost-all-you-need-to-know-f239e6b52279

- Working with Joins, Nested & Repeated Data

- Bigquery ML
  - https://towardsdatascience.com/search?q=Bigquery%20ML