Camille Anderson
Parallel and Distributed Computing
Assignment 2
Spring 2020
6/15/20

Dataset

https://www.kaggle.com/paultimothymooney/minneapolis-police-stops-and-police-violence
http://opendata.minneapolismn.gov/datasets/police-stop-data?geometry=-173.227%2C-5.468%2C79.898%2C48.789&selectedAttribute=personSearch

For this assignment, I used data from the Minneapolis, Minnesota Police Department. I originally found this data on kaggle, but used the minneapolismn.gov website for more information on the variables.

The data is from police stops from 2016 to present. The variables I chose to use were: problem (reason for the stop), personSearch ('yes' or 'no' for whether a search of the person was conducted), race, gender, and policePrecinct. All of the variables are categorical.

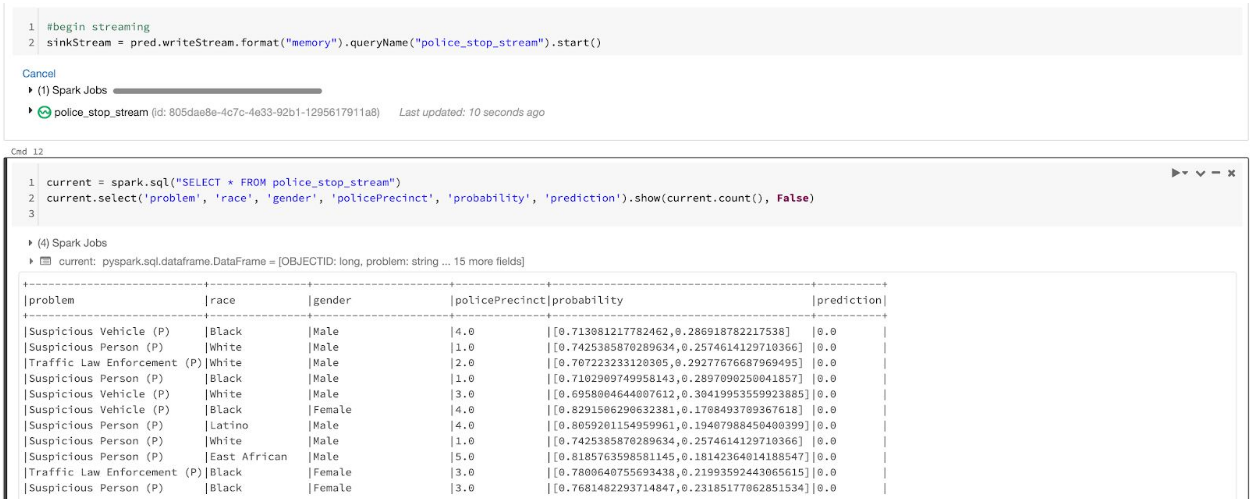
The original dataset included over 160,000 entries. After removing entries with missing values in my selected columns, I was left with only 36,000 entries. Using sklearn's 'train_test_split' I did a 70/30 split on the data into training and testing datasets. The testing dataset was repartitioned and written out into multiple files so that it could be read in one file at a time later for streaming.

Methods

Because of the binary 'yes' or 'no' classification of whether a person search was conducted at a police stop, I chose to use logistic regression to see if my other variables could act as predictors of this probability. I created indexers to index my independent variables for the logistic regression pipeline. Then, the logistic regression model was fit to my training data. I then streamed in the testing data and transformed it with the model.

Results

The model did not predict that any of the stops in the testing data would result in a person search. I think some contributing factors to this might be the large amount of missing values in the data that I had to exclude, and perhaps also that I used too many independent variables, however I believed that they were all important factors to include, and leaving any of them out might misrepresent the data. Additionally, the number of person searches conducted is relatively small. There were about 3 times as many 'no' values in the personSearch column as 'yes'.



```
#begin streaming
sinkStream = pred.writeStream.format("memory").queryName("police_stop_stream").start()
```

Cancel
(1) Spark Jobs
police_stop_stream (id: 805dae8e-4c7c-4e33-92b1-1295617911a8)
Last updated: 10 seconds ago
Cmd 12
(4) Spark Jobs

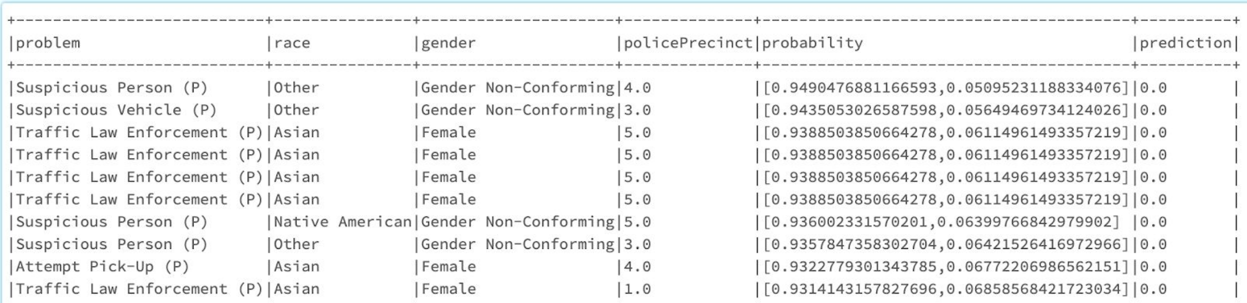
```
current = spark.sql("SELECT * FROM police_stop_stream")
current.select('problem', 'race', 'gender', 'policePrecinct', 'probability', 'prediction').show(current.count(), False)
```

```
current:  pyspark.sql.dataframe.DataFrame = [OBJECTID: long, problem: string ... 15 more fields]
```

```
+---------------------------+---------------+---------------------+--------------+----------------------------------------+----------+
|problem                    |race           |gender               |policePrecinct|probability                             |prediction|
+---------------------------+---------------+---------------------+--------------+----------------------------------------+----------+
|Suspicious Vehicle (P)     |Black          |Male                 |4.0           |[0.713081217782462,0.286918782217538]   |0.0       |
|Suspicious Person (P)      |White          |Male                 |1.0           |[0.7425385870289634,0.2574614129710366] |0.0       |
|Traffic Law Enforcement (P)|White          |Male                 |2.0           |[0.707223233120305,0.29277676687969495] |0.0       |
|Suspicious Person (P)      |Black          |Male                 |1.0           |[0.7102909749958143,0.2897090250041857] |0.0       |
|Suspicious Vehicle (P)     |White          |Male                 |3.0           |[0.6958004644007612,0.30419953559923885]|0.0       |
|Suspicious Vehicle (P)     |Black          |Female               |4.0           |[0.8291506290632381,0.1708493709367618] |0.0       |
|Suspicious Person (P)      |Latino         |Male                 |4.0           |[0.8059201154959961,0.19407988450400399]|0.0       |
|Suspicious Person (P)      |White          |Male                 |1.0           |[0.7425385870289634,0.2574614129710366] |0.0       |
|Suspicious Person (P)      |East African   |Male                 |5.0           |[0.8185763598581145,0.18142364014188547]|0.0       |
|Traffic Law Enforcement (P)|Black          |Female               |3.0           |[0.7800640755693438,0.21993592443065615]|0.0       |
|Suspicious Person (P)      |Black          |Female               |3.0           |[0.7681482293714847,0.23185177062851534]|0.0       |
```

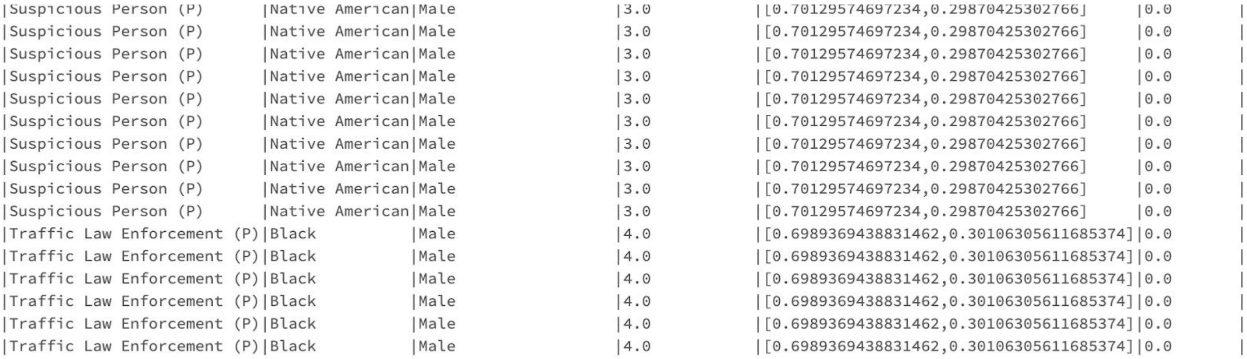
However, some variables seemed to make the probability of NOT having a search conducted higher than others. Here are the results sorted by probability:

Highest probability of not having a search conducted:



```
+---------------------------+---------------+---------------------+--------------+----------------------------------------+----------+
|problem                    |race           |gender               |policePrecinct|probability                             |prediction|
+---------------------------+---------------+---------------------+--------------+----------------------------------------+----------+
|Suspicious Person (P)      |Other          |Gender Non-Conforming|4.0           |[0.9490476881166593,0.05095231188334076]|0.0       |
|Suspicious Vehicle (P)     |Other          |Gender Non-Conforming|3.0           |[0.9435053026587598,0.05649469734124026]|0.0       |
|Traffic Law Enforcement (P)|Asian          |Female               |5.0           |[0.9388503850664278,0.06114961493357219]|0.0       |
|Traffic Law Enforcement (P)|Asian          |Female               |5.0           |[0.9388503850664278,0.06114961493357219]|0.0       |
|Traffic Law Enforcement (P)|Asian          |Female               |5.0           |[0.9388503850664278,0.06114961493357219]|0.0       |
|Traffic Law Enforcement (P)|Asian          |Female               |5.0           |[0.9388503850664278,0.06114961493357219]|0.0       |
|Suspicious Person (P)      |Native American|Gender Non-Conforming|5.0           |[0.936002331570201,0.06399766842979902] |0.0       |
|Suspicious Person (P)      |Other          |Gender Non-Conforming|3.0           |[0.9357847358302704,0.06421526416972966]|0.0       |
|Attempt Pick-Up (P)        |Asian          |Female               |4.0           |[0.9322779301343785,0.06772206986562151]|0.0       |
|Traffic Law Enforcement (P)|Asian          |Female               |1.0           |[0.9314143157827696,0.06858568421723034]|0.0       |
```

Lowest probability of not having a search conducted:



```
|Suspicious Person (P)      |Native American|Male                 |3.0           |[0.70129574697234,0.29870425302766]     |0.0       |
|Suspicious Person (P)      |Native American|Male                 |3.0           |[0.70129574697234,0.29870425302766]     |0.0       |
|Suspicious Person (P)      |Native American|Male                 |3.0           |[0.70129574697234,0.29870425302766]     |0.0       |
|Suspicious Person (P)      |Native American|Male                 |3.0           |[0.70129574697234,0.29870425302766]     |0.0       |
|Suspicious Person (P)      |Native American|Male                 |3.0           |[0.70129574697234,0.29870425302766]     |0.0       |
|Suspicious Person (P)      |Native American|Male                 |3.0           |[0.70129574697234,0.29870425302766]     |0.0       |
|Suspicious Person (P)      |Native American|Male                 |3.0           |[0.70129574697234,0.29870425302766]     |0.0       |
|Suspicious Person (P)      |Native American|Male                 |3.0           |[0.70129574697234,0.29870425302766]     |0.0       |
|Suspicious Person (P)      |Native American|Male                 |3.0           |[0.70129574697234,0.29870425302766]     |0.0       |
|Suspicious Person (P)      |Native American|Male                 |3.0           |[0.70129574697234,0.29870425302766]     |0.0       |
|Traffic Law Enforcement (P)|Black          |Male                 |4.0           |[0.6989369438831462,0.30106305611685374]|0.0       |
|Traffic Law Enforcement (P)|Black          |Male                 |4.0           |[0.6989369438831462,0.30106305611685374]|0.0       |
|Traffic Law Enforcement (P)|Black          |Male                 |4.0           |[0.6989369438831462,0.30106305611685374]|0.0       |
|Traffic Law Enforcement (P)|Black          |Male                 |4.0           |[0.6989369438831462,0.30106305611685374]|0.0       |
|Traffic Law Enforcement (P)|Black          |Male                 |4.0           |[0.6989369438831462,0.30106305611685374]|0.0       |
|Traffic Law Enforcement (P)|Black          |Male                 |4.0           |[0.6989369438831462,0.30106305611685374]|0.0       |
```

The results of the logistic regression show Asian females stopped for traffic enforcement are among the highest probability of not having a person search conducted, and black males stopped for traffic law enforcement have the lowest probability of not having a person search conducted, followed by Native American males stopped for a suspicious person report.
It is important to note that the race and gender values are as recorded by the officer. Additionally, the large number of missing values in these columns that I excluded make it difficult to have an entirely accurate depiction. It is also important to note that of course there are other circumstances that are not captured in this dataset that make having a person search

conducted more or less likely. The purpose of my project was to explore whether some factors may make having a person search conducted more probable, which might suggest implicit biases in policing. However, without more complete and thorough data, drawing any definite conclusions would be irresponsible.