XGBoost Algorithm

Jared Mindel, Isaac Lee, Jeannine Hall, and Jed Dryer

Process Overview

- Selected the dataset
- Prepared the data
- Built and tested the model/tuned hyperparameters
- Compared XGBoost to Logistical Regression
- Issues we encountered
  - Git control within group
  - Long runtimes and crashing computers

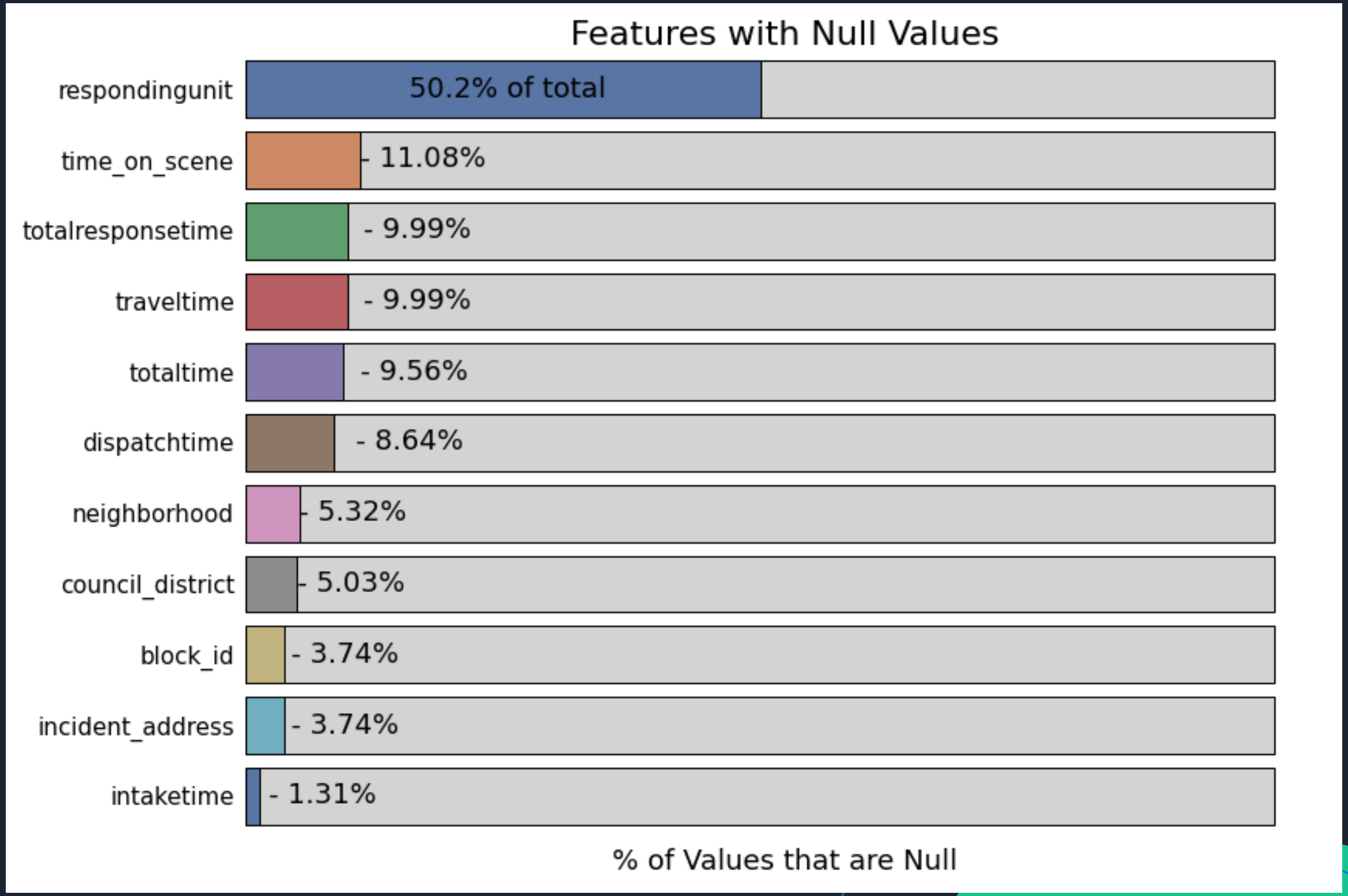
Selecting the Data

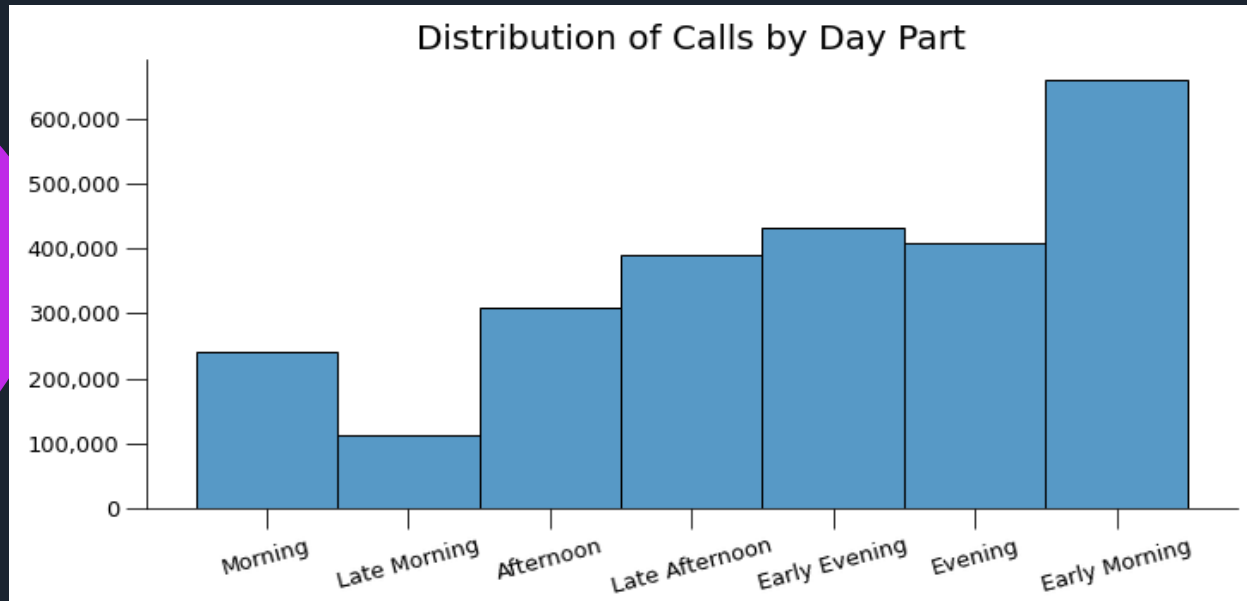
- Datasets we considered:
  - Books
  - Diabetes
  - Cars 93
  - Airports
  - Spotify Songs
  - Meteorite Landings

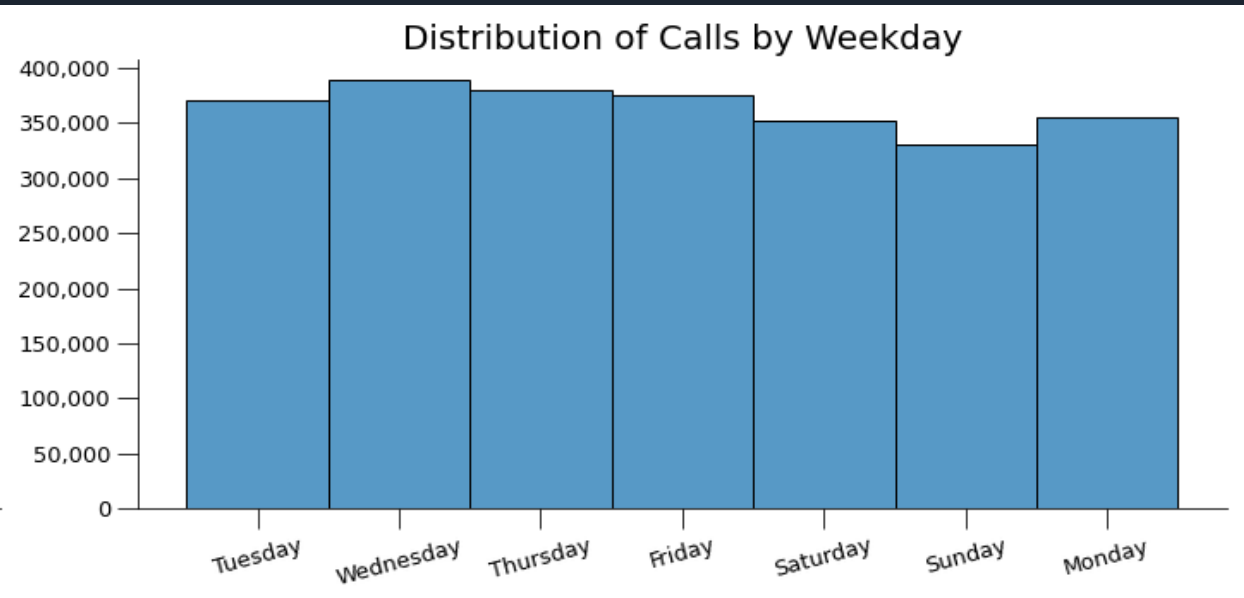
- Our final choice: 911 Service Calls in Detroit

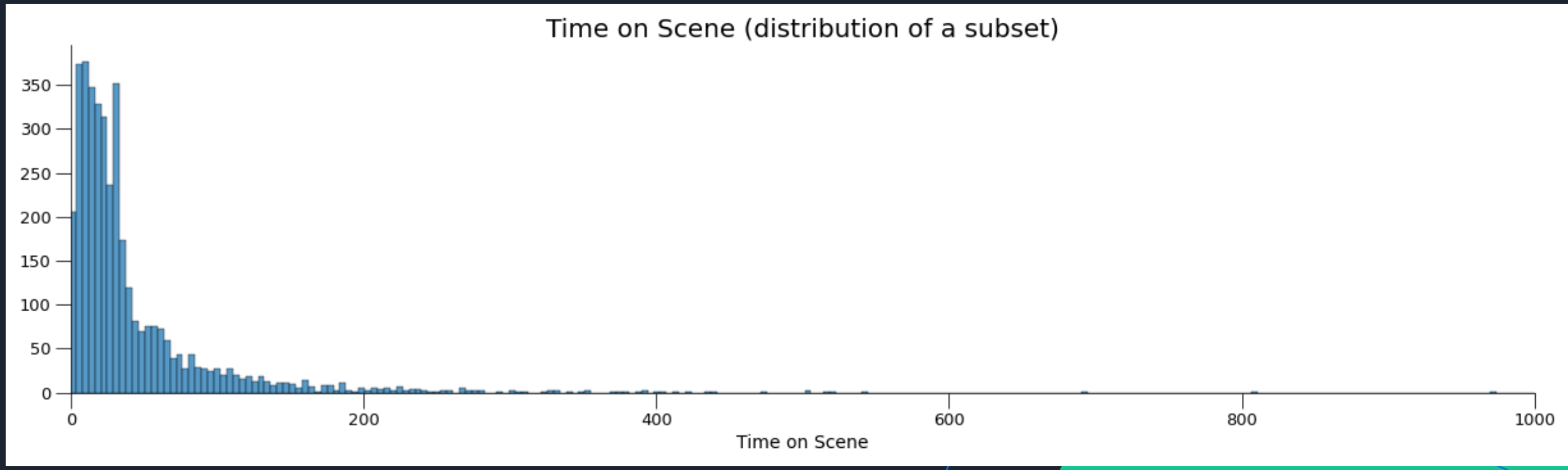
Preparing the Data

- Determined where the null values were
- Removed unnecessary columns
- Filled in missing values in Priorities column
- Created Weekday and Day_Part columns
- Dummied categorical data
- Did not normalize data because XGBoost uses decision trees


Features with Null Values
respondingunit
50.2% of total
time_on_scene
- 11.08%
totalresponsetime
- 9.99%
traveltime
- 9.99%
totaltime
- 9.56%
dispatchtime
- 8.64%
neighborhood
- 5.32%
council_district
- 5.03%
block_id
- 3.74%
incident_address
- 3.74%
intaketime
- 1.31%
% of Values that are Null


Distribution of Calls by Day Part
600,000
500,000
400,000
300,000
200,000
100,000
0
Morning
Late Morning
Afternoon
Late Afternoon
Early Evening
Evening
Early Morning


Distribution of Calls by Weekday
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Monday


Time on Scene (distribution of a subset)
350
300
250
200
150
100
50
0
0
200
400
600
800
1000
Time on Scene

Running the model

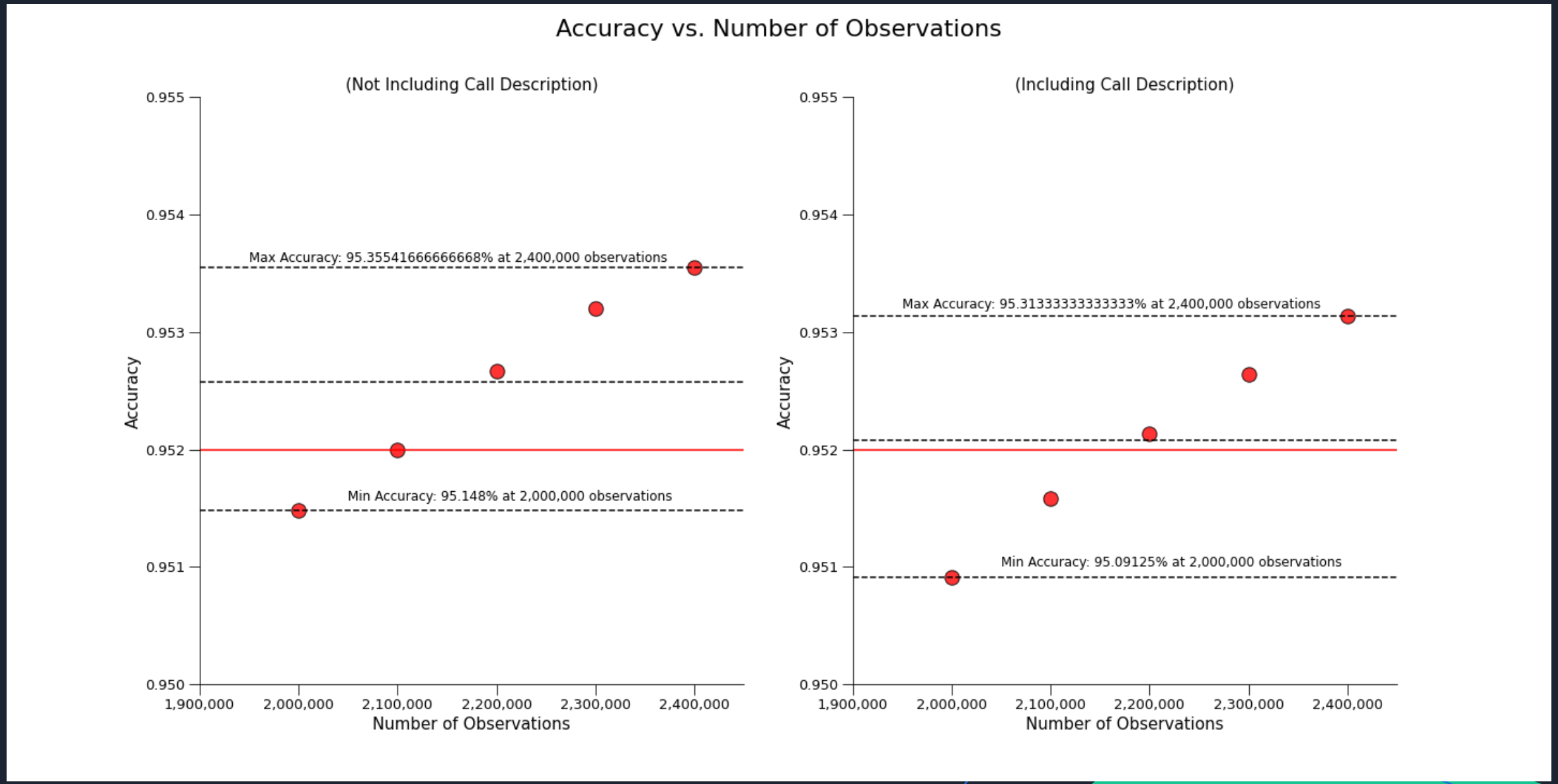
Tried several different varieties of XGBoost

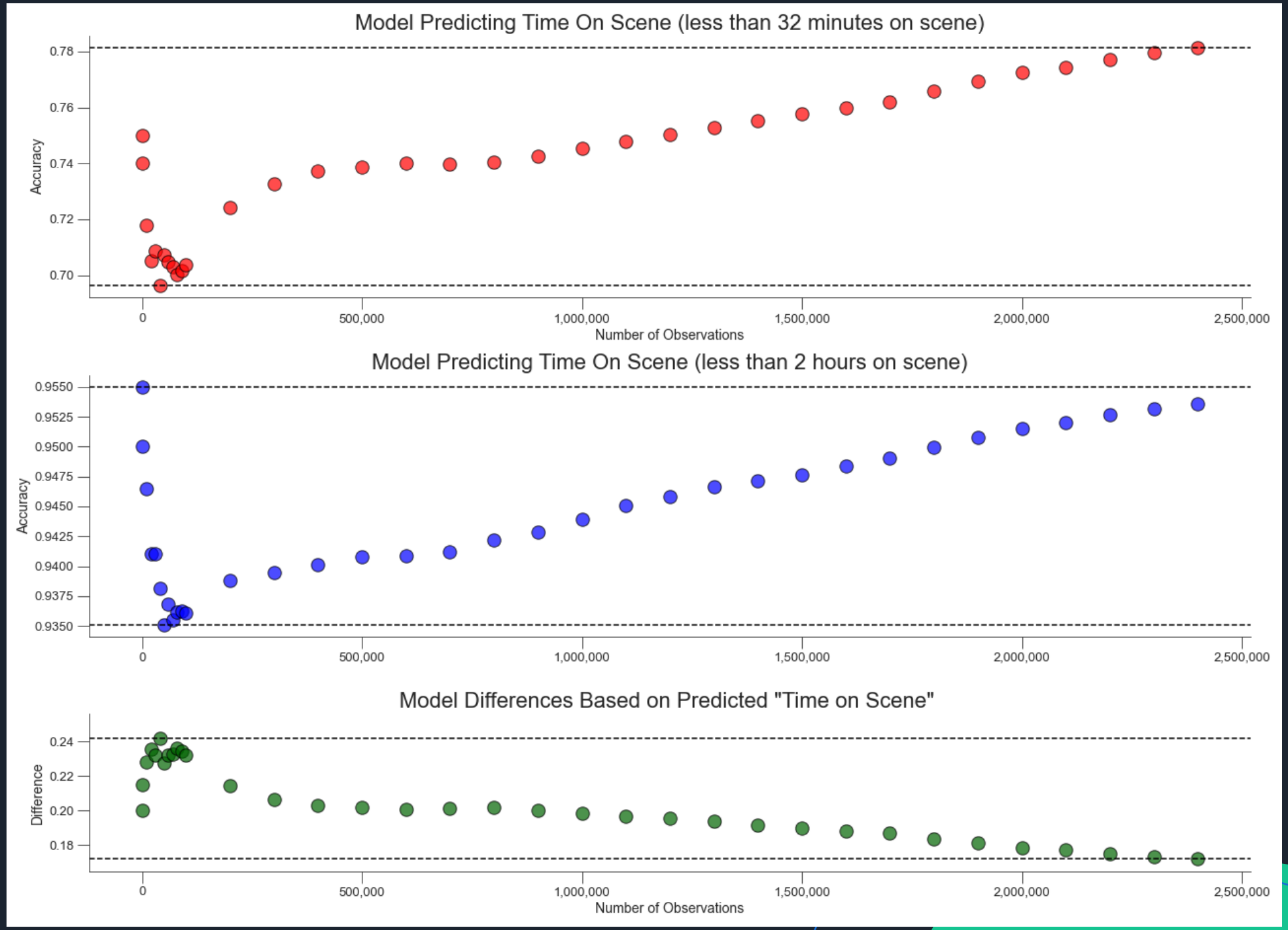
- DMatrix
- Classifier
- Regressor
- Booster

Settled on XGBoost Classifier

- Initially tested two-hour "time on scene"
- Changed our target to the average of the time on scene

Running the model

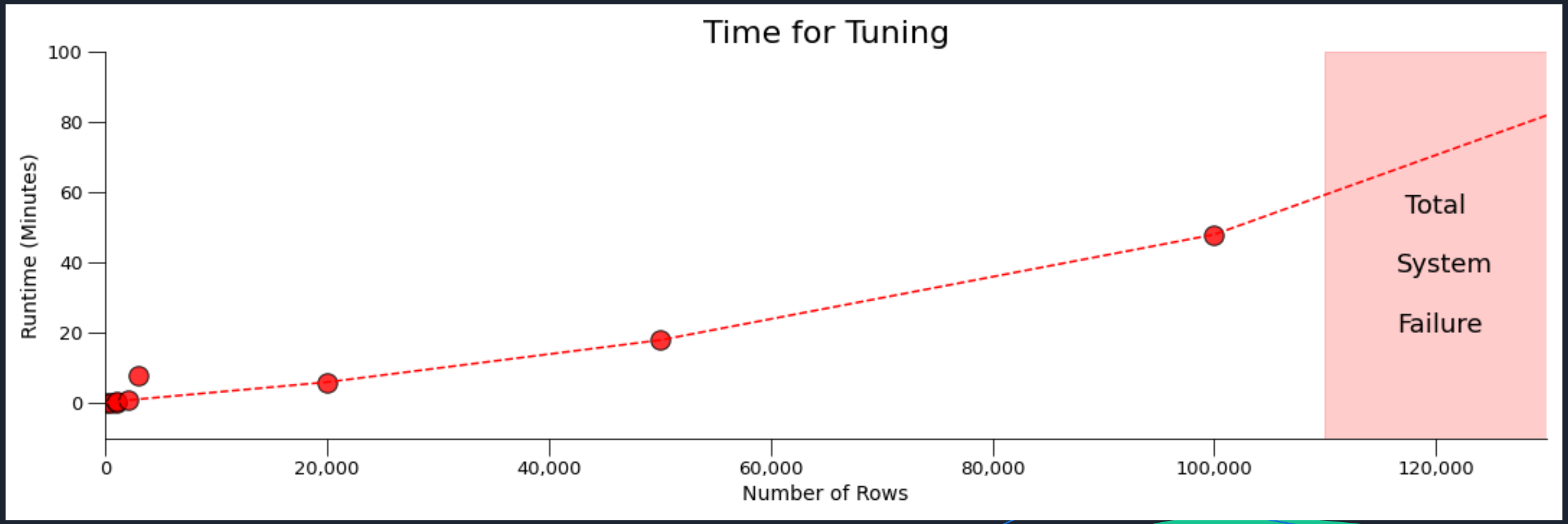


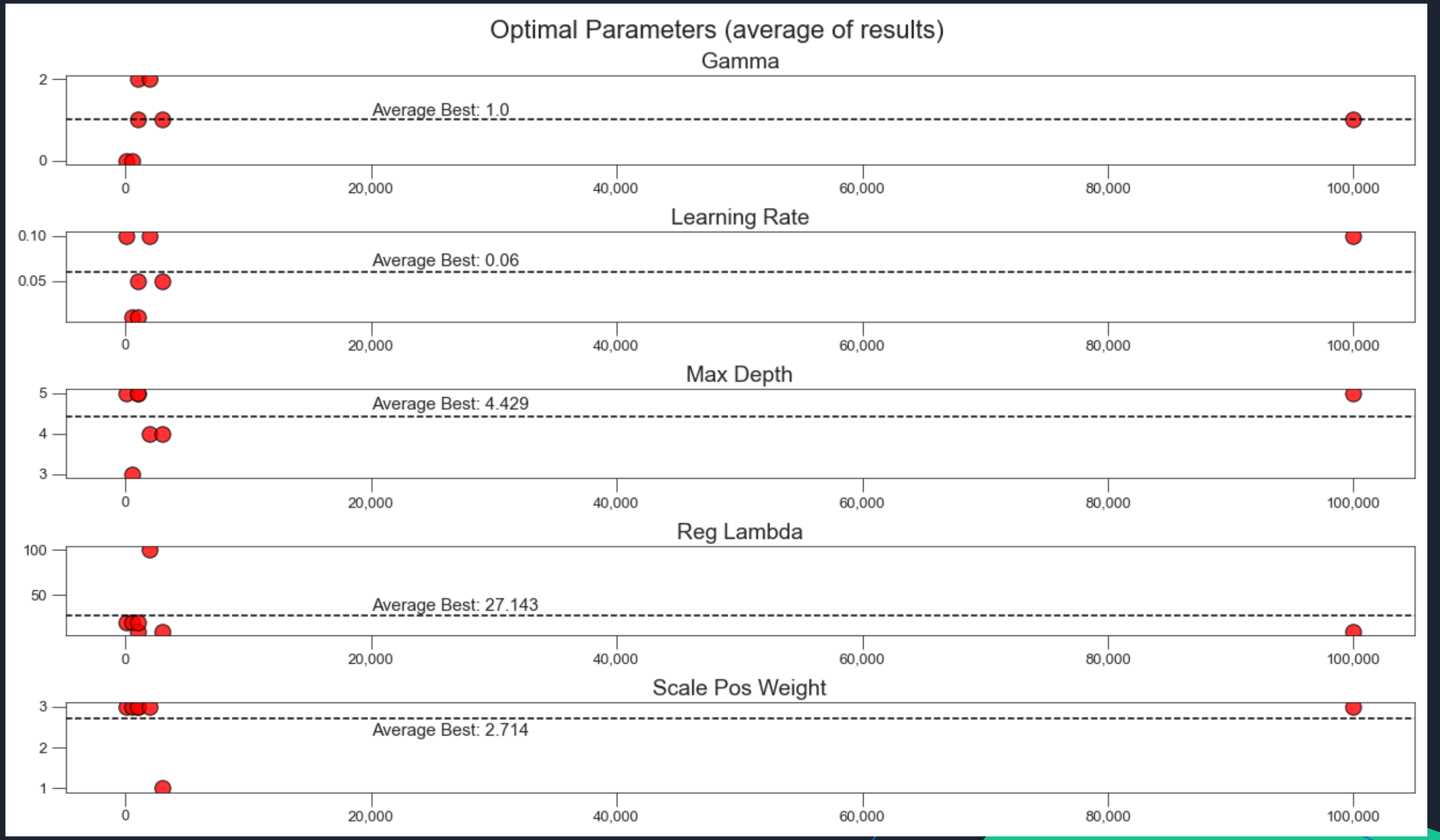

Model Predicting Time On Scene (less than 32 minutes on scene)
Accuracy
0.78
0.76
0.74
0.72
0.70
0
500,000
1,000,000
1,500,000
2,000,000
2,500,000
Number of Observations
Model Predicting Time On Scene (less than 2 hours on scene)
Accuracy
0.9550
0.9525
0.9500
0.9475
0.9450
0.9425
0.9400
0.9375
0.9350
0
500,000
1,000,000
1,500,000
2,000,000
2,500,000
Model Differences Based on Predicted "Time on Scene"
Difference
0.24
0.22
0.20
0.18
0
500,000
1,000,000
1,500,000
2,000,000
2,500,000
Number of Observations

Tuning the Model

- Cross Validation with GridSearchCV
  - Passed in Hyper-Parameters and evaluated the results
  - System Crashes above 100,000 observations
  - During tuning, the best parameters from each round are preserved as the sole value for that parameter.
    - Used the average of the returned "best fit" in our final model.

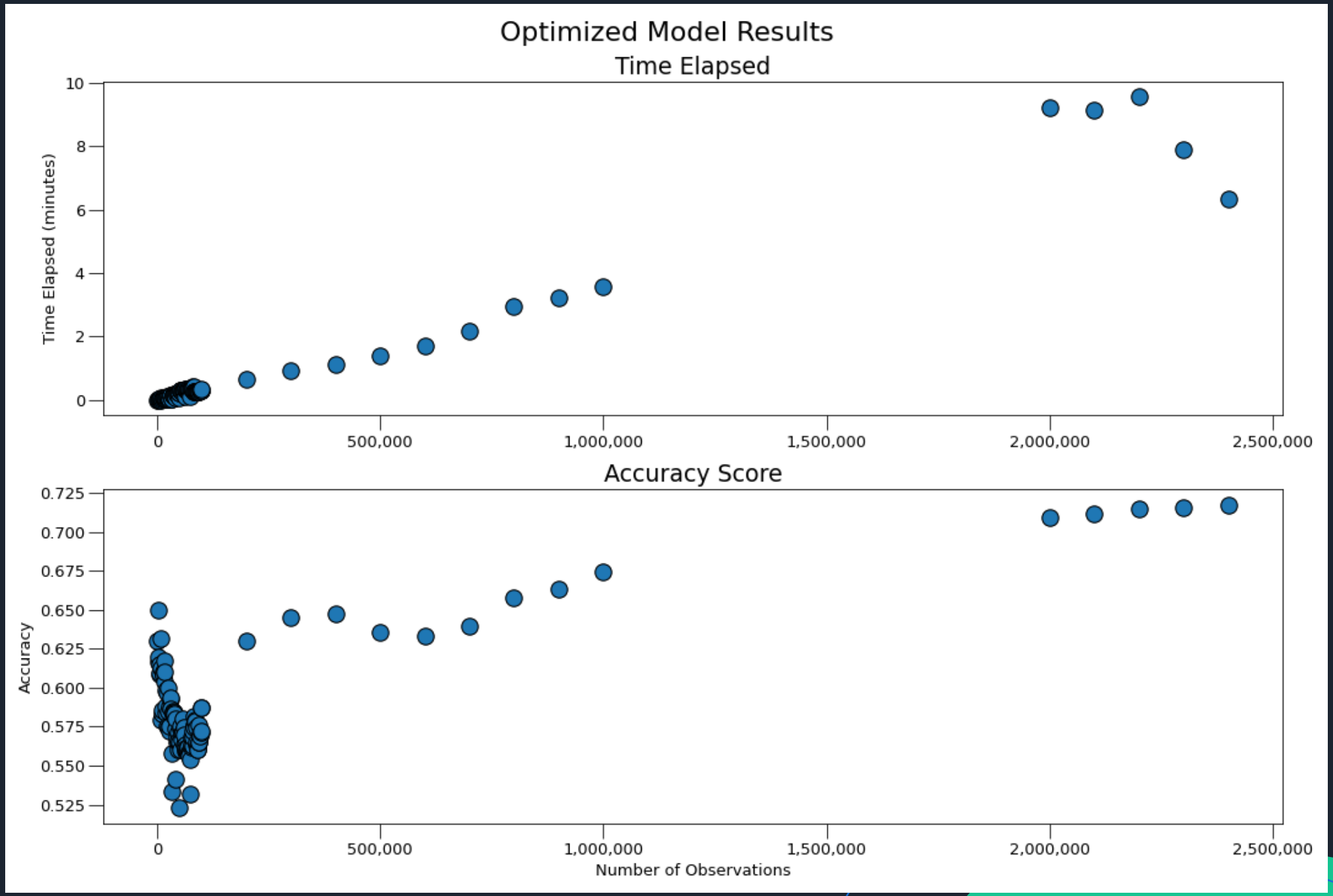
Tuning with GridSearchCV



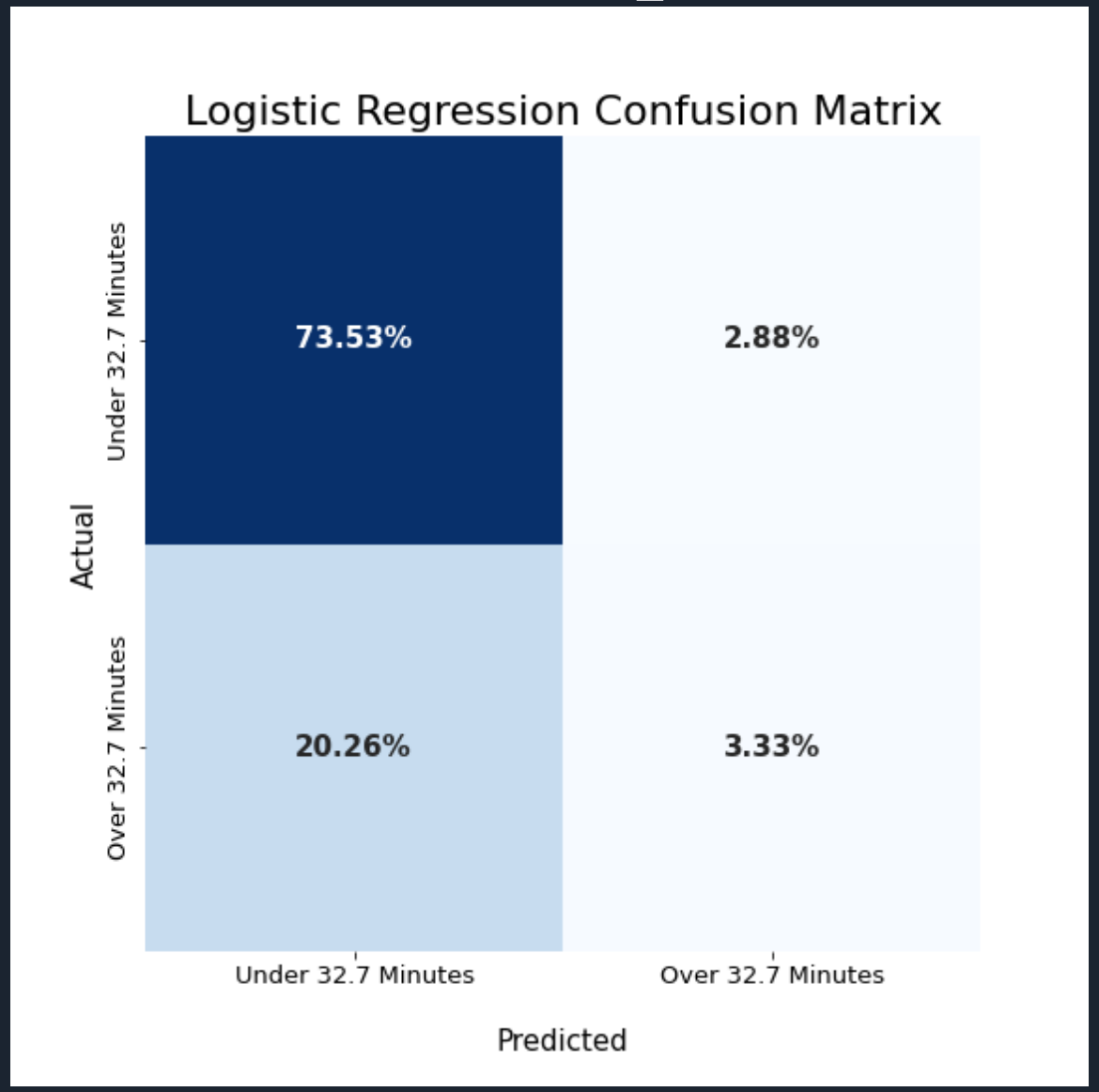

Optimal Parameters (average of results)
Gamma
Average Best: 1.0
Learning Rate
Average Best: 0.06
Max Depth
Average Best: 4.429
Reg Lambda
Average Best: 27.143
Scale Pos Weight
Average Best: 2.714
0
20,000
40,000
60,000
80,000
100,000

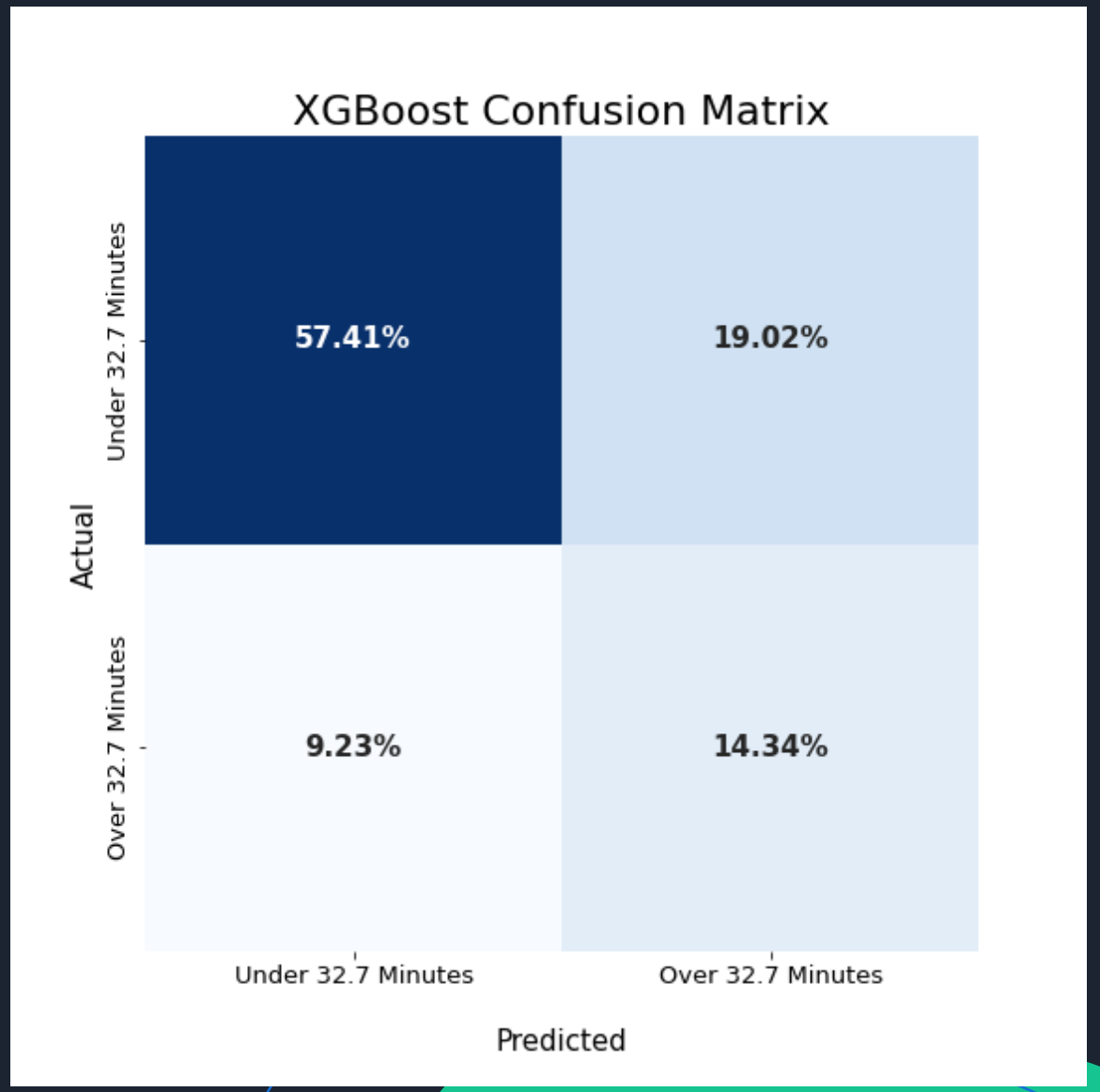
Model Performance Improvements

- Iterate fewer times – for sake of speed
- Boost memory performance
- Removed Call Description from the model
- Adjusting the random state may change the accuracy of the results
- Reverting to the pre-tuning hyper-parameters.

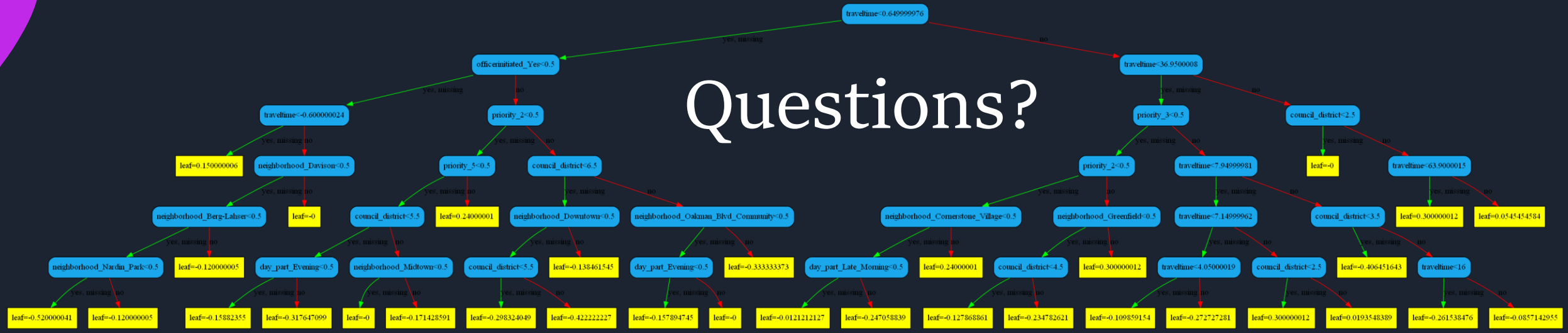

Optimized Model Results
Time Elapsed
Time Elapsed (minutes)
10
8
6
4
2
0
0
500,000
1,000,000
1,500,000
2,000,000
2,500,000
Accuracy Score
Accuracy
0.725
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0.525
0
500,000
1,000,000
1,500,000
2,000,000
2,500,000
Number of Observations

Model Comparison: Confusion Matrices





Questions?


traveltime<0.649999976
yes, missing
no
officerinitiated_Yes<0.5
traveltime<36.9500008
yes, missing
no
traveltime<-0.600000024
priority_2<0.5
priority_3<0.5
council_district<2.5
yes, missing
no
leaf=0.150000006
neighborhood_Davison<0.5
priority_5<0.5
council_district<6.5
priority_2<0.5
traveltime<7.94999981
leaf=-0
traveltime<63.9000015
neighborhood_Berg-Lahser<0.5
leaf=-0
council_district<5.5
leaf=0.24000001
neighborhood_Downtown<0.5
neighborhood_Oakman_Blvd_Community<0.5
neighborhood_Cornerstone_Village<0.5
neighborhood_Greenfield<0.5
traveltime<7.14999962
council_district<3.5
leaf=0.300000012
leaf=0.0545454584
neighborhood_Nardin_Park<0.5
leaf=-0.120000005
day_part_Evening<0.5
neighborhood_Midtown<0.5
council_district<5.5
leaf=-0.138461545
day_part_Evening<0.5
leaf=-0.333333373
day_part_Late_Morning<0.5
leaf=0.24000001
council_district<4.5
leaf=0.300000012
traveltime<4.05000019
council_district<2.5
leaf=-0.406451643
traveltime<16
leaf=-0.520000041
leaf=-0.120000005
leaf=-0.15882355
leaf=-0.317647099
leaf=-0
leaf=-0.171428591
leaf=-0.298324049
leaf=-0.422222227
leaf=-0.157894745
leaf=-0
leaf=-0.0121212127
leaf=-0.247058839
leaf=-0.127868861
leaf=-0.234782621
leaf=-0.109859154
leaf=-0.272727281
leaf=0.300000012
leaf=0.0193548389
leaf=-0.261538476
leaf=-0.0857142955