Predicting Sports Activity on Endomondo Sensor Data

Shrikar Thodla, Dillon Quan,
Mikio Tada

Outline

- Dataset
- Related Works
- Analytic Goals
- Data Preprocessing
- Model Training / Comparison
- Lessons Learn

What is EndoMondo?

- A mobile app that collects data when the user is exercising to track their health and fitness data.

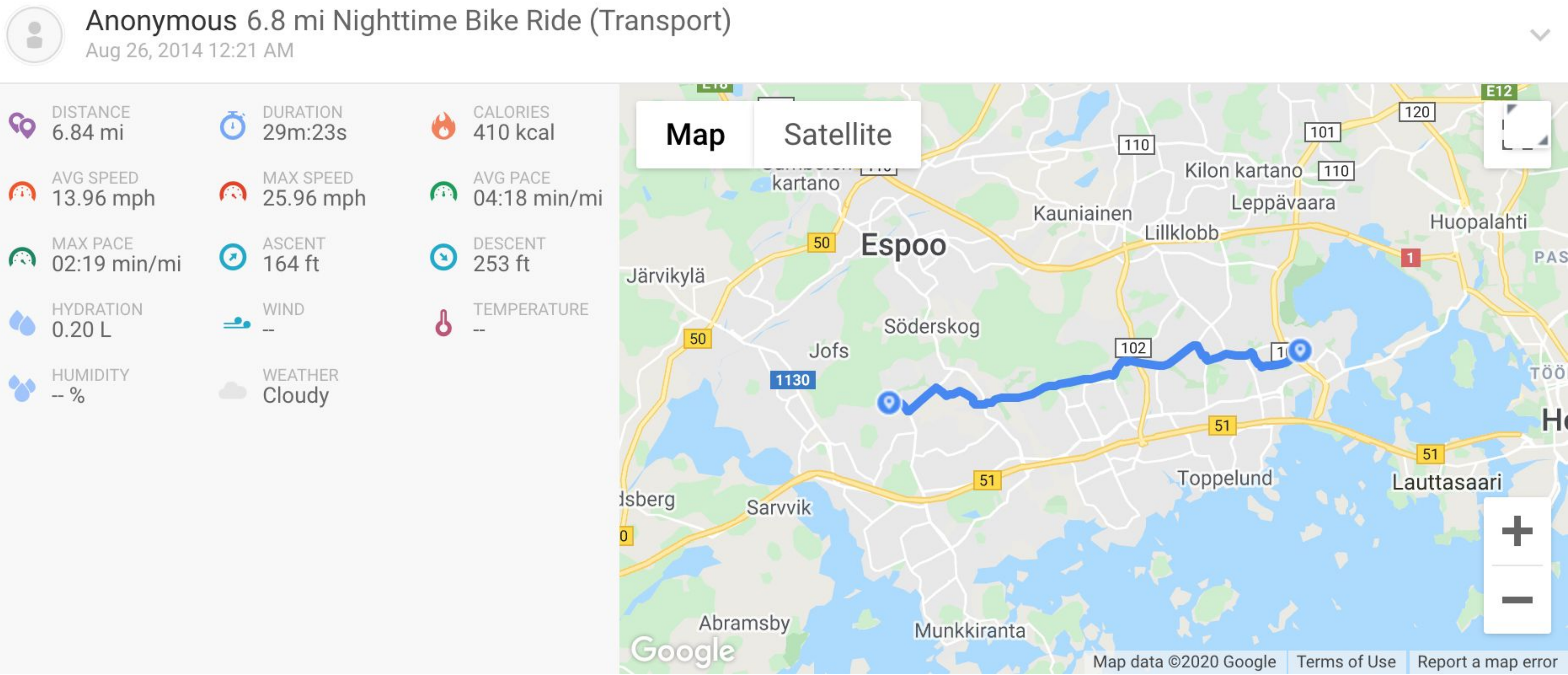
- Compatible with various mobile devices
  - Phone
  - Smart Watch
  - Smart Bracelet

Dataset

- Contained exercise data for various sports
  - Biking
  - Running
  - Biking (Transport)
  - Mountain Biking
- Size - 4.6Gb

endomondo

Data Source: https://sites.google.com/eng.ucsd.edu/fitrec-project/home?authuser=0

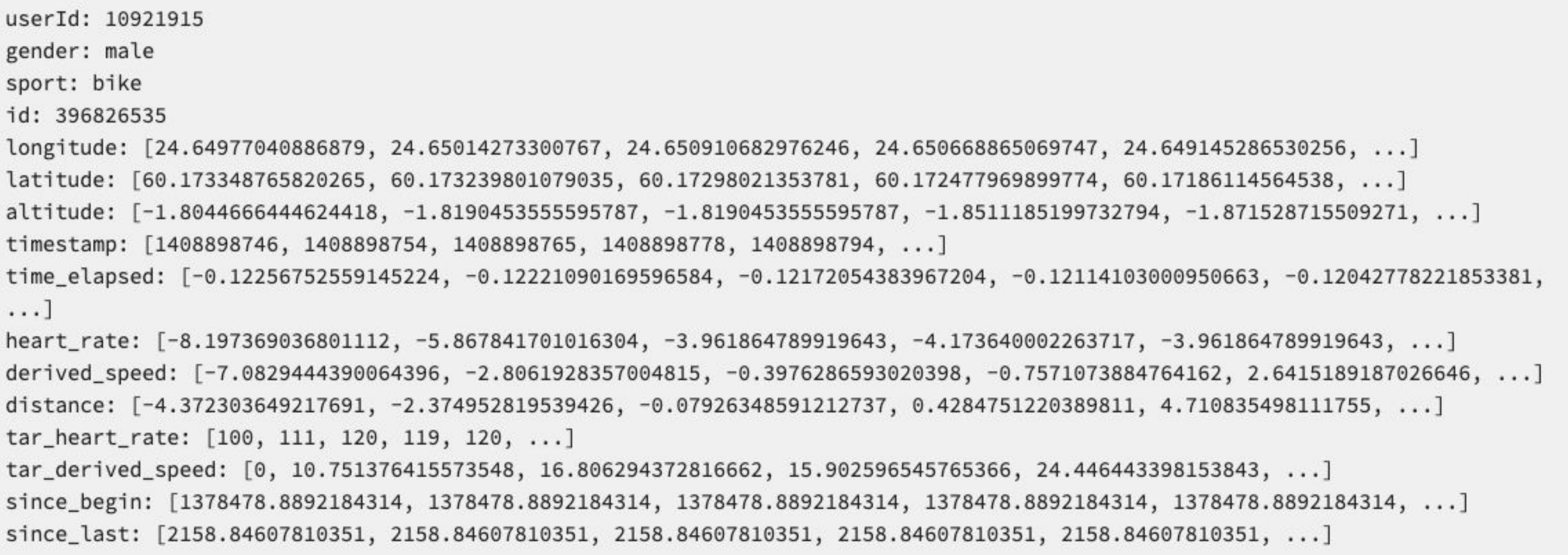


Anonymous 6.8 mi Nighttime Bike Ride (Transport)

Aug 26, 2014 12:21 AM

DISTANCE 6.84 mi	DURATION 29m:23s	CALORIES 410 kcal
AVG SPEED 13.96 mph	MAX SPEED 25.96 mph	AVG PACE 04:18 min/mi
MAX PACE 02:19 min/mi	ASCENT 164 ft	DESCENT 253 ft
HYDRATION 0.20 L	WIND --	TEMPERATURE --
HUMIDITY -- %	WEATHER Cloudy	

Dataset



```
userId: 10921915
gender: male
sport: bike
id: 396826535
longitude: [24.64977040886879, 24.65014273300767, 24.650910682976246, 24.650668865069747, 24.649145286530256, ...]
latitude: [60.173348765820265, 60.173239801079035, 60.17298021353781, 60.172477969899774, 60.17186114564538, ...]
altitude: [-1.8044666444624418, -1.8190453555595787, -1.8190453555595787, -1.8511185199732794, -1.871528715509271, ...]
timestamp: [1408898746, 1408898754, 1408898765, 1408898778, 1408898794, ...]
time_elapsed: [-0.12256752559145224, -0.12221090169596584, -0.12172054383967204, -0.12114103000950663, -0.12042778221853381,
...]
heart_rate: [-8.197369036801112, -5.867841701016304, -3.961864789919643, -4.173640002263717, -3.961864789919643, ...]
derived_speed: [-7.0829444390064396, -2.8061928357004815, -0.3976286593020398, -0.7571073884764162, 2.6415189187026646, ...]
distance: [-4.372303649217691, -2.374952819539426, -0.07926348591212737, 0.4284751220389811, 4.710835498111755, ...]
tar_heart_rate: [100, 111, 120, 119, 120, ...]
tar_derived_speed: [0, 10.751376415573548, 16.806294372816662, 15.902596545765366, 24.446443398153843, ...]
since_begin: [1378478.8892184314, 1378478.8892184314, 1378478.8892184314, 1378478.8892184314, 1378478.8892184314, ...]
since_last: [2158.84607810351, 2158.84607810351, 2158.84607810351, 2158.84607810351, 2158.84607810351, ...]
```

Related Works

- Original work on the dataset had 2 main goals
  - Workout forecasting Speed and Heart Rate
  - Recommending other sport activity based on their heart rate

- Reference: Jianmo Ni, Larry Muhlstein, Julian McAuley, "Modeling heart rate and activity data for personalized fitness recommendation", in Proc. of the 2019 World Wide Web Conference (WWW'19), San Francisco, US, May. 2019.

Related Works

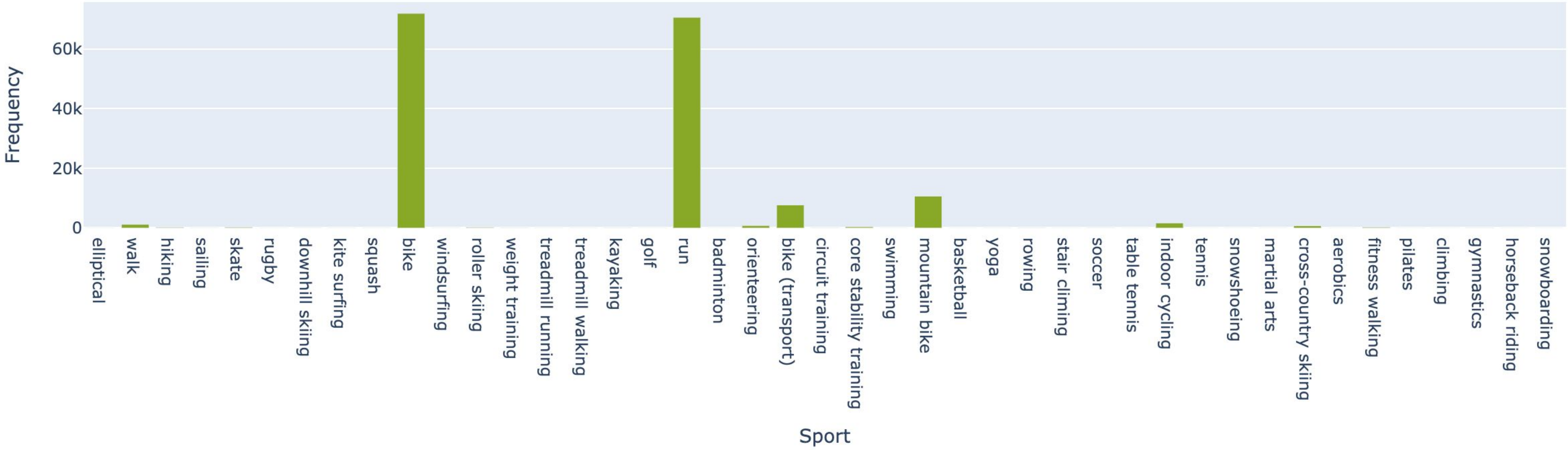
- Related work focused on classifying sport activity type using Endomondo data

- Reference: Alok Kumar Chowdhury, Aleksandr Farseev, Prithwi Raj Chakraborty, Dian Tjondronegoro, and Vinod Chandran. 2017. Automatic classification of physical exercises from wearable sensors using small dataset from non-laboratory settings. 2017 IEEE Life Sciences Conference (LSC) (2017), 111–114.

Analytic Goals

Goal: To explore other various machine learning algorithms for classifying sport activities using endomondo sensor data.

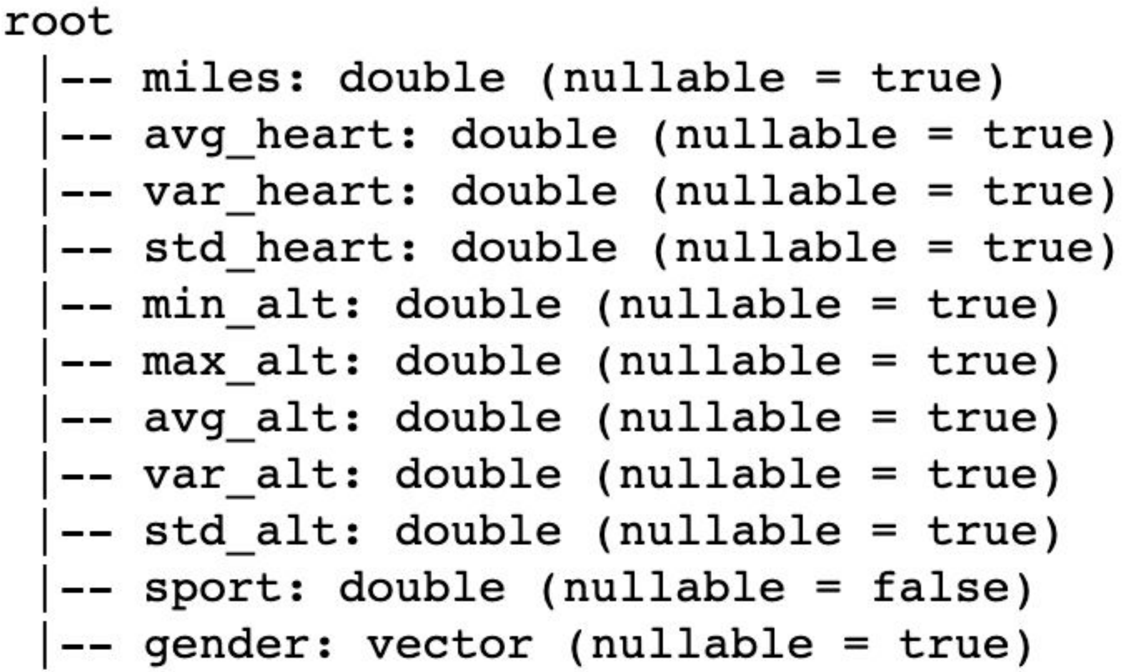
Label Distribution



Feature Engineering

- Miles
  - Using Haversine distance
- Standard Deviation, Variance, and Mean of Heart Rate and Altitude

Processed Dataset

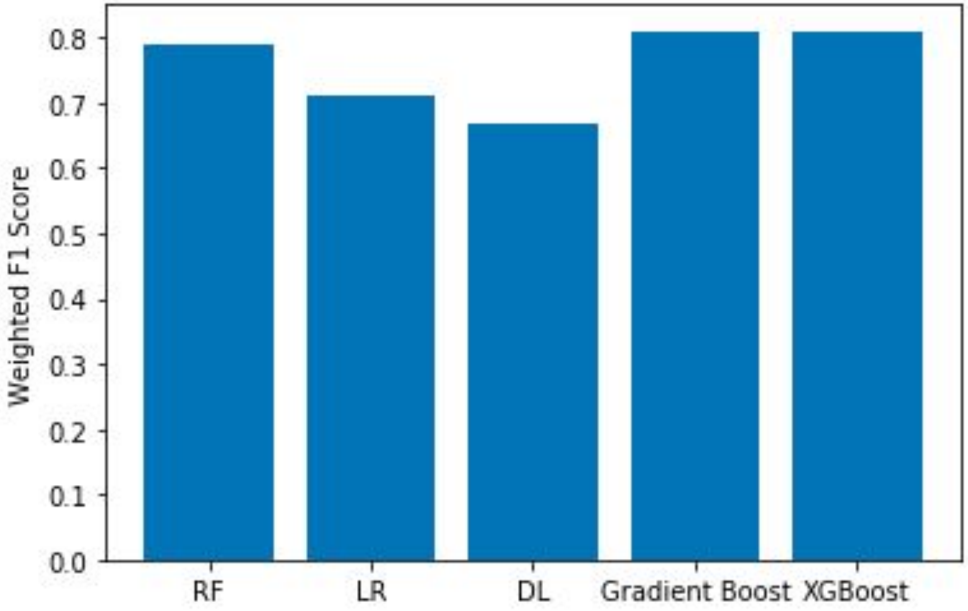


```
root
 |-- miles: double (nullable = true)
 |-- avg_heart: double (nullable = true)
 |-- var_heart: double (nullable = true)
 |-- std_heart: double (nullable = true)
 |-- min_alt: double (nullable = true)
 |-- max_alt: double (nullable = true)
 |-- avg_alt: double (nullable = true)
 |-- var_alt: double (nullable = true)
 |-- std_alt: double (nullable = true)
 |-- sport: double (nullable = false)
 |-- gender: vector (nullable = true)
```

Model Training

- Spark ML
  - Logistic Regression
  - Random Forest
- H2O
  - Neural Net
  - Gradient Boosting
  - XGBoost

Model Comparison



Best Model

- Gradient Boosting
  - Number of trees = 20
  - Max Depth = 20

- Weighted F1 = .814 (Unbalanced)
- Weighted F1 = .786 (Balanced)

AWS EMR Instances Comparison

EMR CONFIGURATION	EXECUTION TIME (SEC)
5 INSTANCES M5.XLARGE	170.31
4 INSTANCES M5.XLARGE	162.82
3 INSTANCES M5.XLARGE	96.62
3 INSTANCES M5.2XLARGE	161.28
2 INSTANCES M5.XLARGE	SparkContext would not start

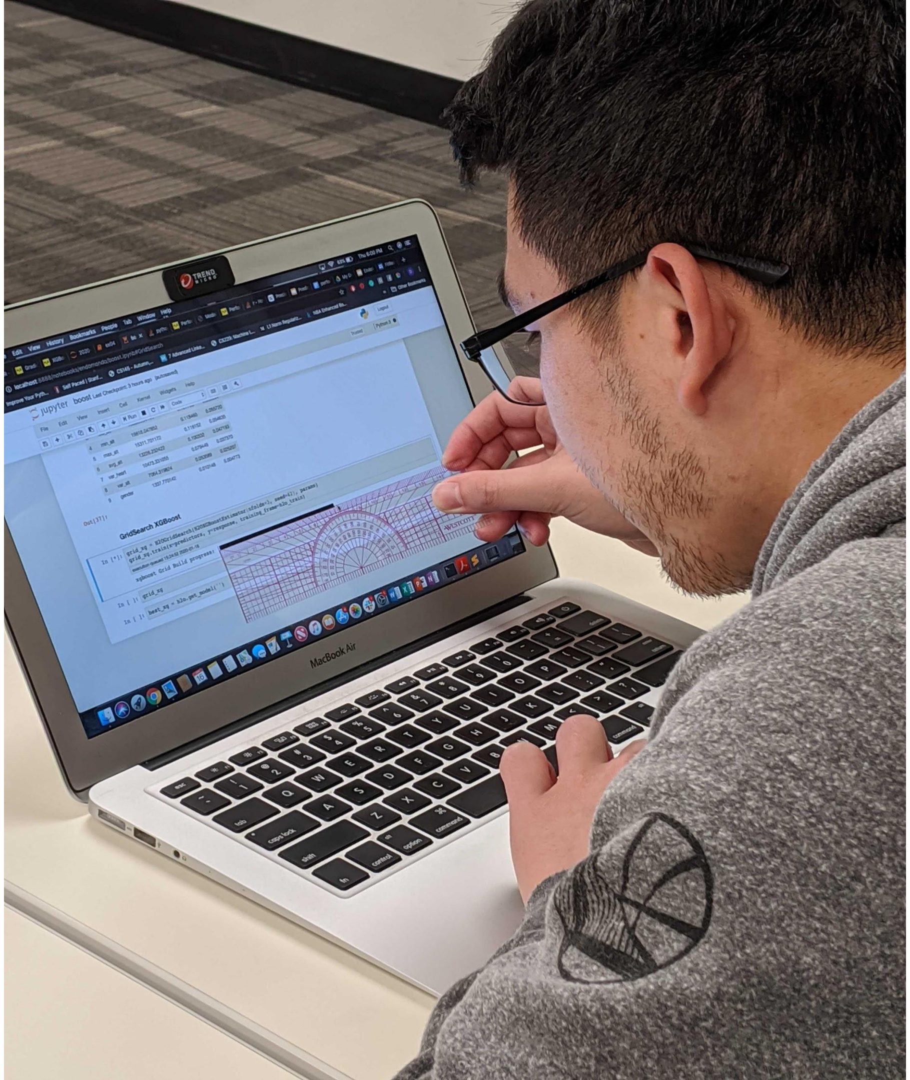
Pre-Processing Algorithms & Model Training Time Efficiency

- With 3 instances of m5.xlarge (4 vCore, 16 GiB memory, 64 GiB storage),
- Pre-Processing: 194.32 seconds.

Lesson Learned

- Reducing data traffic between nodes drastically decreases computational time.
- Figuring out what is sufficient for the task at hand and always start simple.
- Auto ML with max models is too slow given the time constraint. Try setting the time limit next time.

UNIVERSITY OF SAN FRANCISCO
Master of Science in Data Science
2:39



TREND MICRO
jupyter boost Last Checkpoint: 3 hours ago (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Run
Code
MacBook Air

```
4  min_alt    15816.047852  0.115460  0.055770
5  max_alt    15311.701172  0.116152  0.054635
6  avg_alt    13226.232422  0.103202  0.047193
7  var_heart  10473.331055  0.079449  0.027370
8  var_alt     7064.319824  0.053089  0.025297
9  gender      1337.770142  0.010148  0.004773
```

Out[37]:

GridSearch XGBoost

```
In [*]: grid_xg = H2OGridSearch(H2OXGBoostEstimator(nfolds=5, seed=42), params)
        grid_xg.train(x=predictors, y=response, training_frame=h2o_train)
        execution queued 15:24:52 2020-01-16
        xgboost Grid Build progress:

In [ ]: grid_xg

In [ ]: best_xg = h2o.get_model('')
```