Training and Racing: A story told in Data

Hi there.

My name is Keith Kruelskie.
Today I will talk about training, racing, and how we analyze it.
You can find me at https://github.com/keithkruelskie
(or out running or biking)

My Journey (So Far)

Graduated College

(2009)

Friends and 5ks

TOUGH MUDDER

(2012)

First Triathlon

Chevron Houston Marathon

(2014)

(2015)

First Ironman top 10 AG finish

(2016)

70.3 IM Worlds Qualifier

(2019)

No Data

With Data

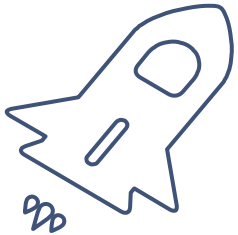
“In many ways, a race is analogous to life itself. Once it is over, it can not be re-created. All that is left are impressions in the heart, and in the mind.”
-Chris Lear

1

Training Data Fundamentals

Background and Problem Statement

Data in Training



- **Why is it important?**
  - Monitoring/Accountability
  - Improve Future training
  - **Predict** Athlete outcomes

Acquiring Data

- Fitness devices
- Fitness Data ecosystems
- User consent

GARMIN™

Important Training Metrics

Distance

- Known before race day
- Generally, defines if aerobic/ anaerobic effort
- Standardized

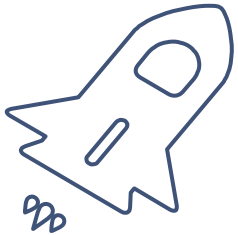
Pace

- Varies with distance, time, elevation
- Hard to predict accurately
- Determines race finish

Heart Rate

- Measure of intensity
- Recorded continuously
- Can be inaccurate

Can we predict an athlete’s race pace based on their previous training?



2

Training Data Exploration

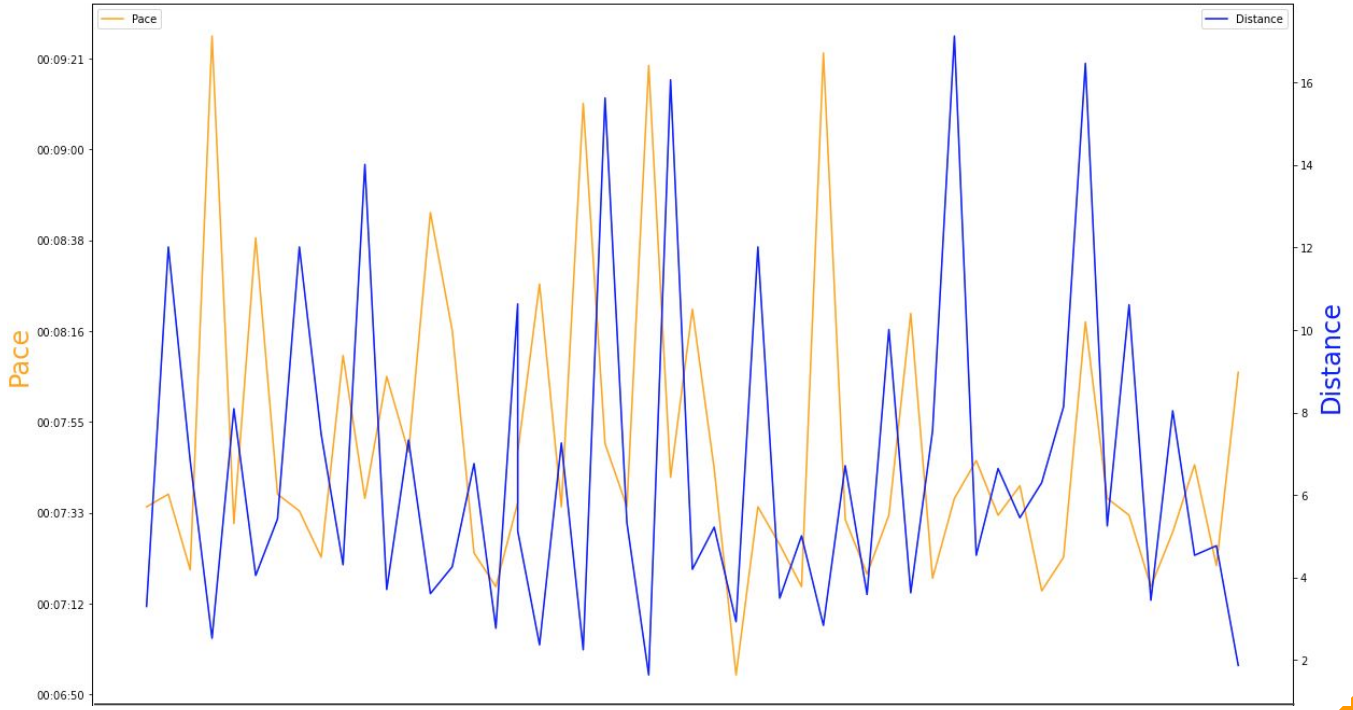
Data Processing and EDA

Importing and EDA

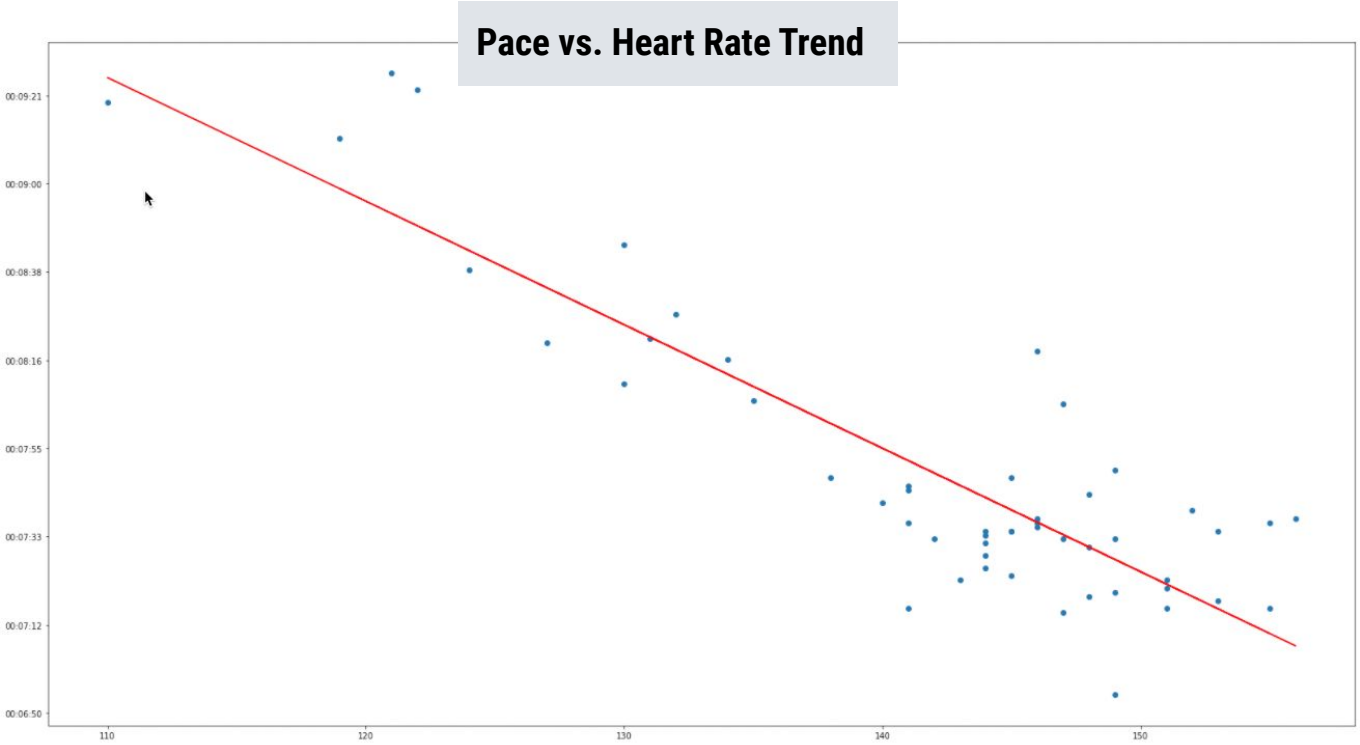
Wrangling Garmin Data

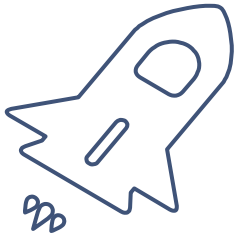
- Variable Features file to file
- Mixed Data Types
- Few data points to many per athlete
- Hobbyists and Racers

Initial Findings



Initial Findings





Key Takeaways:

- Pace over Intensity = “Threshold Pace”
- Athletes train at a variety of different intensities through their training
- Heart rate indicator of effort level

3

Modeling

autoARIMA and PyMC3 Bayesian Forecasting

autoARIMA

Bright Spots

- Quickly finds correlations(if any) based on time
- Accurate predictions, train to race data

Limitations

- Essentially defaults to linear regression in this dataset
- Does not account for seasonality of training

PyMC3

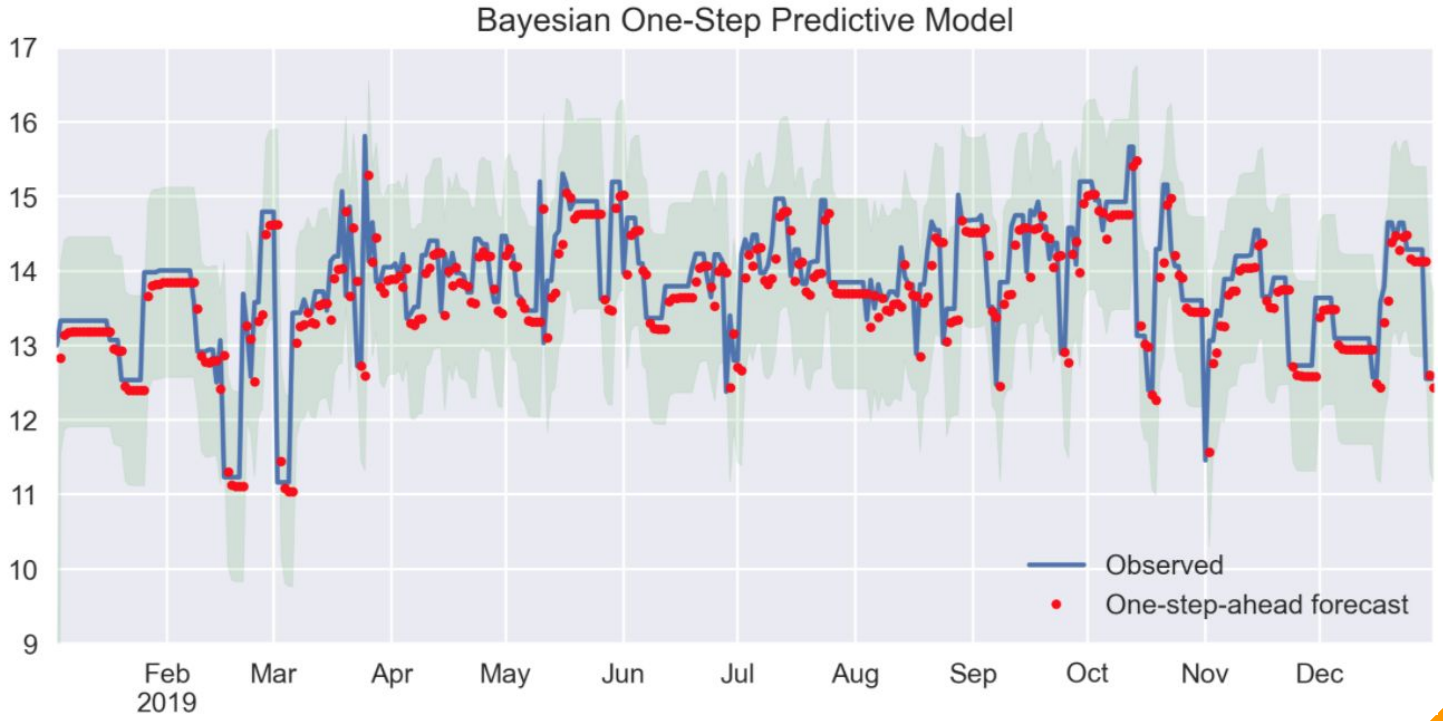
Bright Spots

- Accurate Short term forecasting
- Confidence Intervals

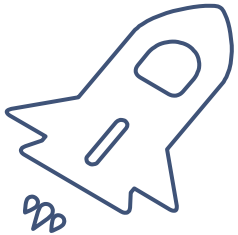
Limitations

- Can be construed as overfit/sensitive to data errors
- Not useful for long term projections

PyMC3 Predictions



Statsmodels SARIMAX (1,0,1) model wrapped by PyMC3



Results:

- Pace Predicted on most athletes within :10-15s/mile (linear autoARIMA)
- One step ahead predictions and confidence intervals accurate over time

4

Forecasting

PyMC3 Seasonality Prediction

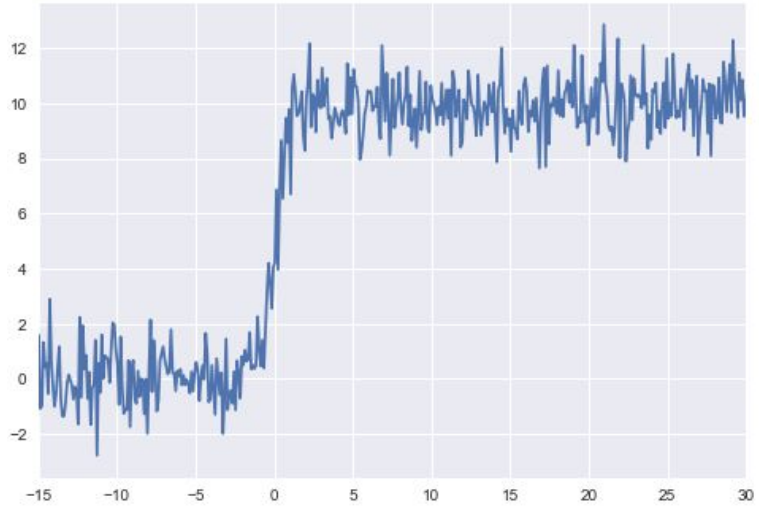
So, What's the Problem?

- We can predict athlete racing with a training history.
- But, what if we want to look into the future?

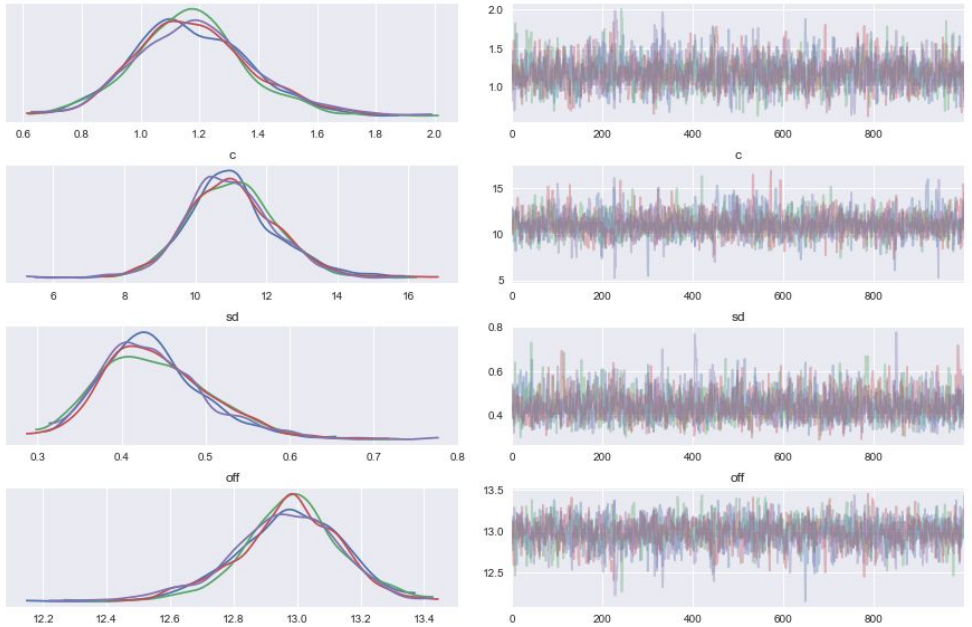
Athlete Training Pace Curve

- Valuable to Athlete and Coach
- Builds on the current pace
- Built on PyMC3
- Weekly smoothing

Athlete Training Pace Curve (Behind the scenes)



Synthetic Data Simulation for Model Design



Trace Plots on actual athlete data

Athlete Training Pace Curve - Results

- Predictions within 10-15s/mile on half Marathon and Marathon distances
- Pace curves match with predicted increases (for coached athletes)

THANKS!

Any questions?

You can find me at

@keithkruelskie & k.kruelskie@gmail.com