Best Predictive Model: Pedestrian Usage of Downtown Mall using Wifi Data

Alex P. Miller
Love Thy K-Nearest Neighbors

The Team

The Team

The Team

The Team

Wharton
University of Pennsylvania

Alex P. Miller

Ph.D. Student, Information Systems

I study:

- Recommendation systems
- A/B testing
- Algorithmic decision making

Big thanks to @BecomingDataSci (Data Science Renee) for tweeting about the competition!

The Data

The Data

Calendar features:

Day of week, quarter, federal holidays, etc.

Quite important given test period!

The Data

Weather data:

Temperature, wind, humidity

(from Weather Underground API)

The Data

Local events data:

http://downtowncharlottesville.net

Actually did have predictive power!

The Data

UVA Men's Basketball Schedule:

The Data


Downtown Business Association of Charlottesville
BASKETBALL


Retrievers
35

The Data

UVA Men's Basketball Schedule:

Home/away, ranked opponent, etc.

Not highly predictive, but still fun

The Data

Calendar

Weather

Local events

UVA Basketball

60+ possible features

Hand-picked < 30 that I thought would be most meaningful

The Model

Simple linear model with < 30 features

$$y_i = X_i \beta$$

The Model

Simple linear model with < 30 features

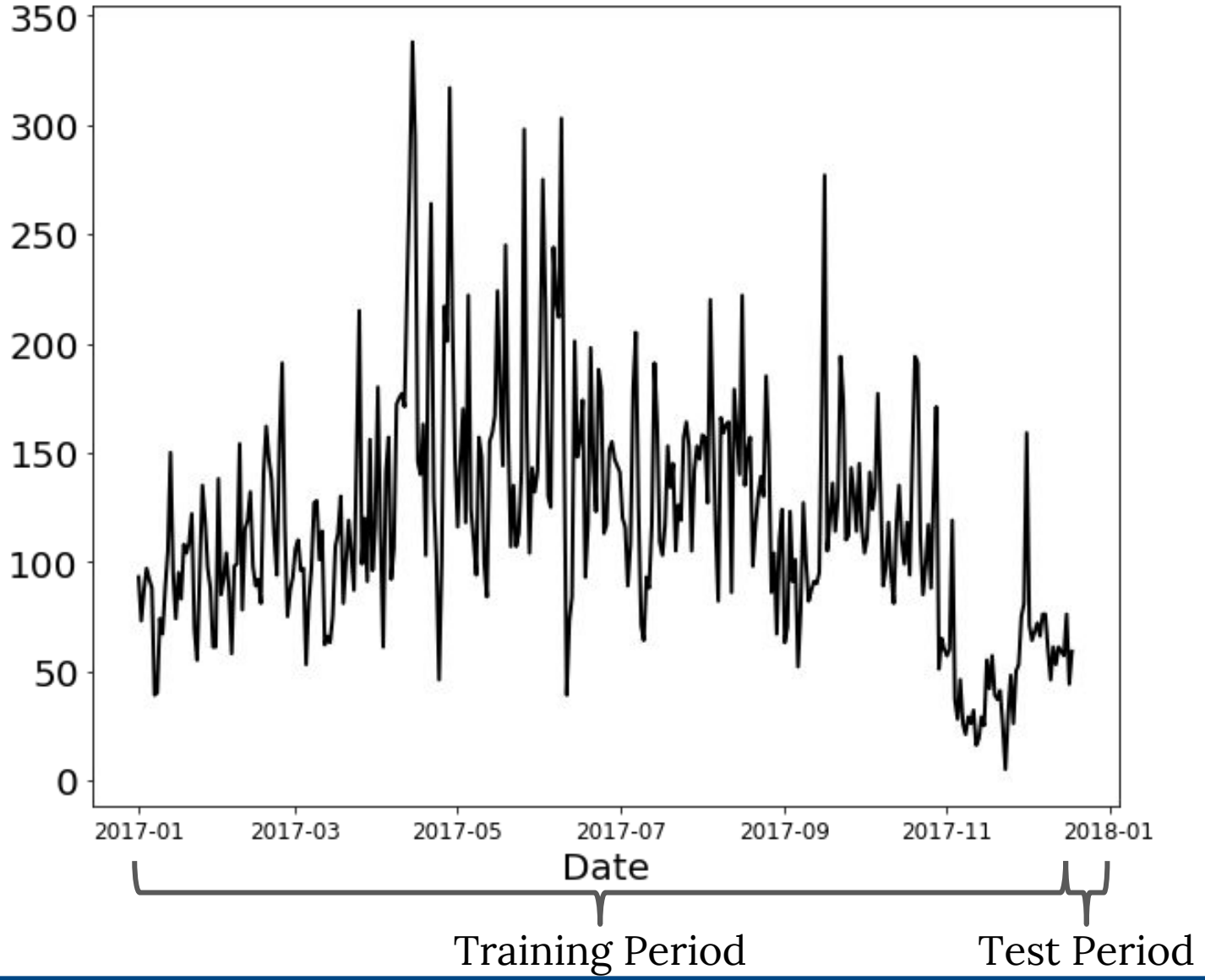
$$y_i = X_i \beta$$

Custom loss function:

- MAPE objective with L2 regularization (tuned with 5-fold CV)

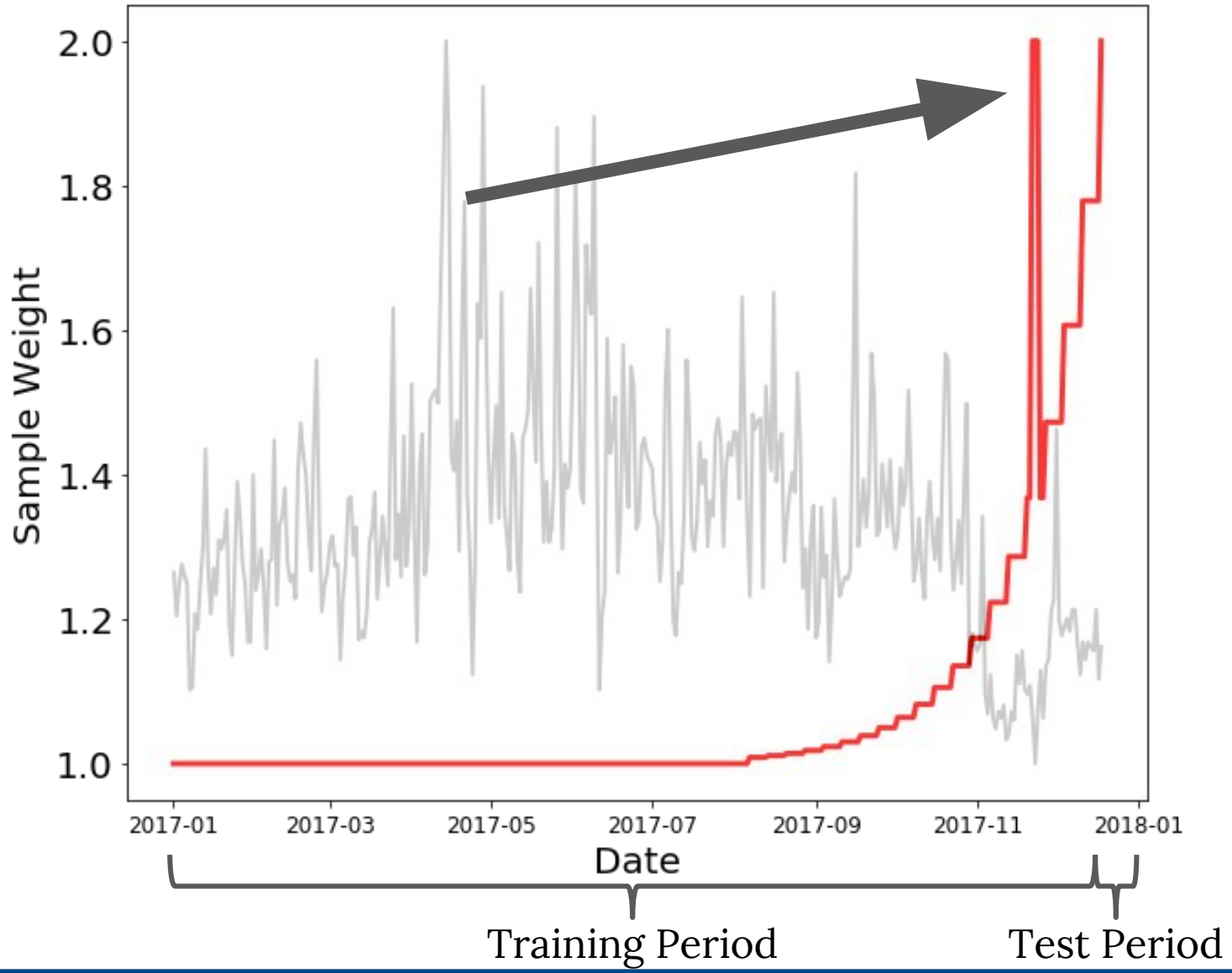
$$L(\beta) = 100 \left(\sum_{i=1}^{N} w_i \right)^{-1} \sum_{i=1}^{N} w_i \left| \frac{y_i - X_i \beta}{y_i} \right| + \lambda \sum_{k=1}^{K} \beta_k^2$$

The Tricks



- Very hard to de-trend time-series with only one year of data!
- Test period included Christmas

The Tricks



- Solution: exponentially smoothed weighting of time series + Thanksgiving

The Code

```
In [3]: # These didn't work for me!
        from sklearn.model_selection import cross_validate
        from sklearn.metrics.scorer import make_scorer
```

- Built my own linear model object, fit using numerical optimization (MAPE has no analytical minimum, like OLS)
- Built my own cross validator that could accommodate a MAPE model objective with L2 regularization

The Lessons

- Use better data, not better models!
- Optimize directly for your evaluation criterion
- Avoid overfitting during exploratory phase
  - Have a holdout dataset!
- Above all, use common sense
  - When N is small, weight your data intelligently
  - Make sure your model passes sanity checks (e.g., low traffic on Christmas)

Thanks!

Will write blog post about methodology… stay in touch!

@alexpmil

alexmill@upenn.edu