Taxi Travel Time Prediction

Assignment 3 - Outcome Lecture

Sebastian Caldas and Nicholay Topin

This lecture has 2 objectives:

Summarize the students’ solutions to the assignment

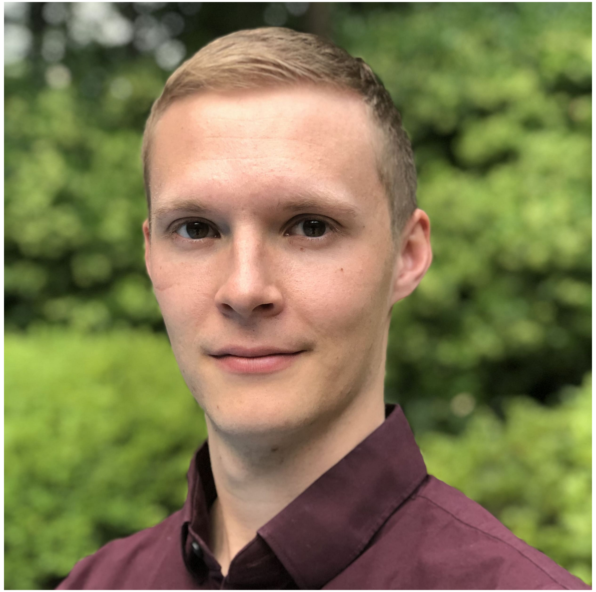
Understand how the assignments have related to the **course’s goals**

This lecture has 2 objectives:

Summarize the students’ solutions to the assignment

Understand how the assignments have related to the **course’s goals**

Helen Zhou



Jacob Tyo

Global summary

“By 5pm on April 15, 2019, make a submission to Kaggle that beats the baseline.”

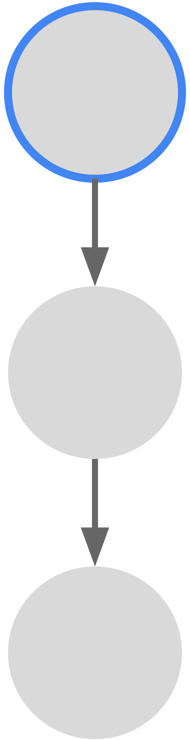
- We did some feature engineering
    - For a given pick up-drop off pair, we calculated the first, second and third quartiles for the travel time.
    - We added these as 3 new features to our samples

- Our model was a 2-layer neural network (with ReLU non-linearities)
    - We first made sure the network could overfit the training data
        - We increased the size of the layers to 2048 neurons
    - We then added some regularization in the form of dropout
    - We trained on 5% of the data using Adam

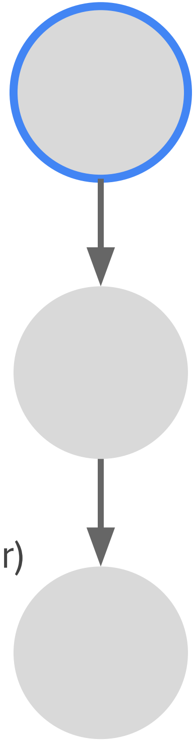
Any comments?

“Provide a clear, detailed description of your overall pipeline sufficient to reproduce your exact pipeline.”

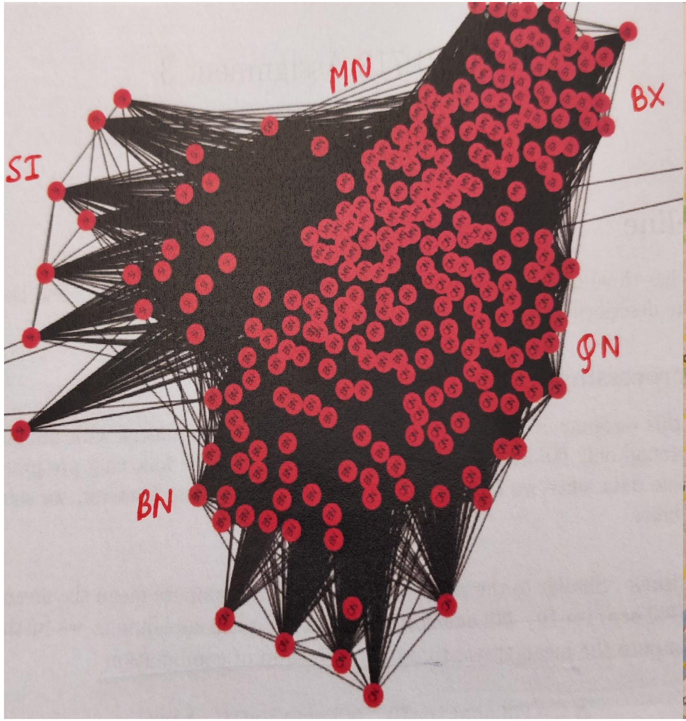
1. Preprocessing
   - Mostly done for you (Thanks again, Nicholay!)
   - Convert time *t* to *ln(t + 1)* to easily optimize RMSLE
   - Subsample the data (to account for limited resources)

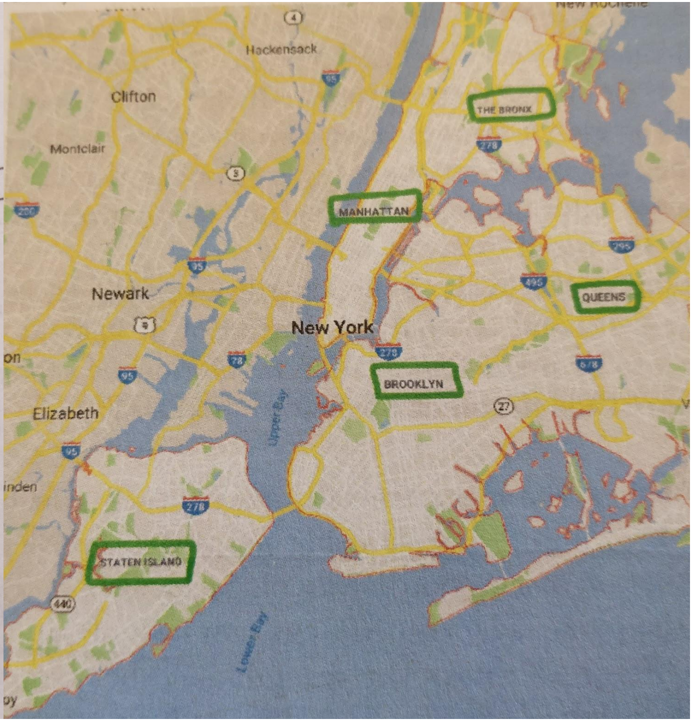


“Describe the pipeline used for your submission and present your results.”



1. Preprocessing
   - Mostly done for you (Thanks, Nicholay!)
   - Convert time *t* to *ln(t + 1)* to easily optimize RMSLE
   - Subsample the data (to account for limited resources)
2. Feature engineering
   - Remove “vendor id”, “payment type” and “passenger count” (?)
   - Month (?), day of week, hour of day (categorical)
   - Distance between locations
   - Average time for pick-up/drop-off pair
   - Traffic estimates (count for pick-up/drop-off pair, sometimes hour)
   - Additional external data (described later)
   - Embeddings of the pick-up/drop-off locations

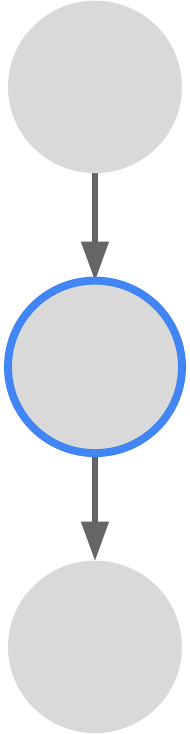




Figures by Biswajit Paria

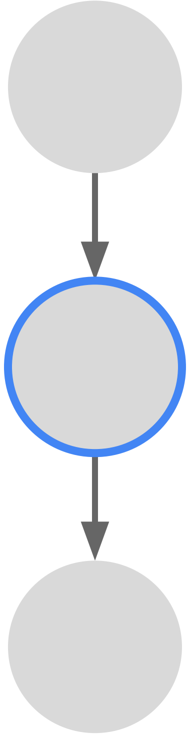
“Describe the pipeline used for your submission and present your results.”

3. Split into train/val sets
   - Test set was given
   - Best estimates if train happened before val

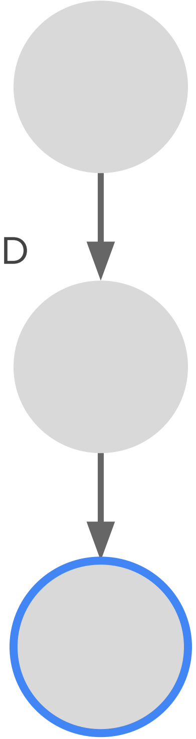


“Describe the pipeline used for your submission and present your results.”

3. Split into train/val sets
    - Test set was given
    - Best estimates if train happened before val
4. Method Selection
    - Dictionaries
    - Random forests (most popular)
    - Boosted trees
    - Nearest neighbors (not very flexible)
    - Shallow feed-forward neural network (quite unpopular?)
    - Classifier per pick-up/drop-off pair (sometimes band of day)
        - Requires handling sparsity



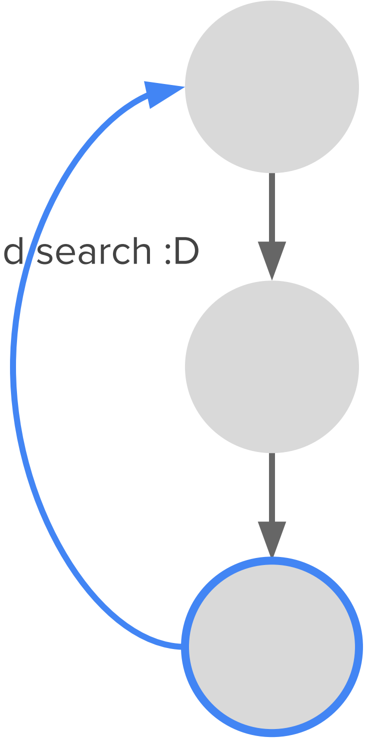
“Describe the pipeline used for your submission and present your results.”



5. Tuning
   - Tune on a developer set (different from train/val)
   - Cross-validation, grid-search, random-search
   - People learned not to pick an extreme value of the grid search :D
6. Evaluation
   - Convert back from log-space
   - Evaluate on val set (before submitting to Kaggle)

“Describe the pipeline used for your submission and present your results.”

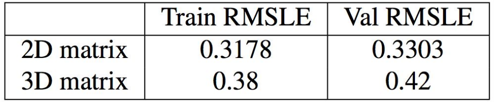
5. Tuning
   - Tune on a developer set (different from train/val)
   - Cross-validation, grid-search, random-search
   - People learned not to pick an extreme value of the grid search :D
6. Evaluation
   - Convert back from log-space
   - Evaluate on val set (before submitting to Kaggle)
7. Iterate
   - First method did not work for many



Any comments?

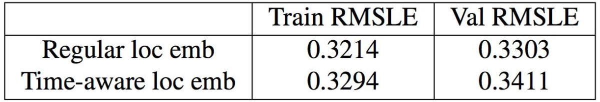
“Describe the process you used to select your pipeline and improve it.”

- Ablation studies



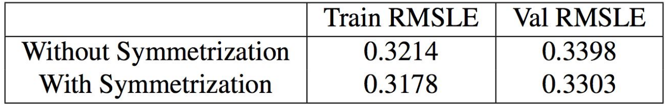
	Train RMSLE	Val RMSLE
2D matrix	0.3178	0.3303
3D matrix	0.38	0.42

Table 2: Effect of using (PU, DO, hour) mean time instead of (PU, DO)



	Train RMSLE	Val RMSLE
Regular loc emb	0.3214	0.3303
Time-aware loc emb	0.3294	0.3411

Table 6: T-SNE-based time-aware location embeddings



	Train RMSLE	Val RMSLE
Without Symmetrization	0.3214	0.3398
With Symmetrization	0.3178	0.3303

Table 1: Effect of symmetrization on the mean travel time matrix

Tables by Srinivas Ravishankar

“Describe the process you used to select your pipeline and improve it.”

- Hyperparameter tuning

Any comments?

“Describe the additional data you used.”

- Most popular types of external data:
    - Weather (different granularities)
        - https://www.timeanddate.com/
        - https://www.kaggle.com/selfishgene/historical-hourly-weather-data#weather_description.csv
        - https://darksky.net/dev
        - https://w2.weather.gov/climate/index.php?wfo=okx
    - Holidays
        - Wikipedia
    - Real-time traffic speed data
        - https://data.cityofnewyork.us/Transportation/Real-Time-Traffic-Speed-Data/qkm5-nuaq

“Describe the additional data you used.”

- Most popular types of external data:
  - Weather (different granularities)
    - https://www.timeanddate.com/
    - https://www.kaggle.com/selfishgene/historical-hourly-weather-data#weather_description.csv
    - https://darksky.net/dev
    - https://w2.weather.gov/climate/index.php?wfo=okx
  - Holidays
    - Wikipedia
  - Real-time traffic speed data
    - https://data.cityofnewyork.us/Transportation/Real-Time-Traffic-Speed-Data/qkm5-nuaq
- Most pipelines could easily handle the additional features

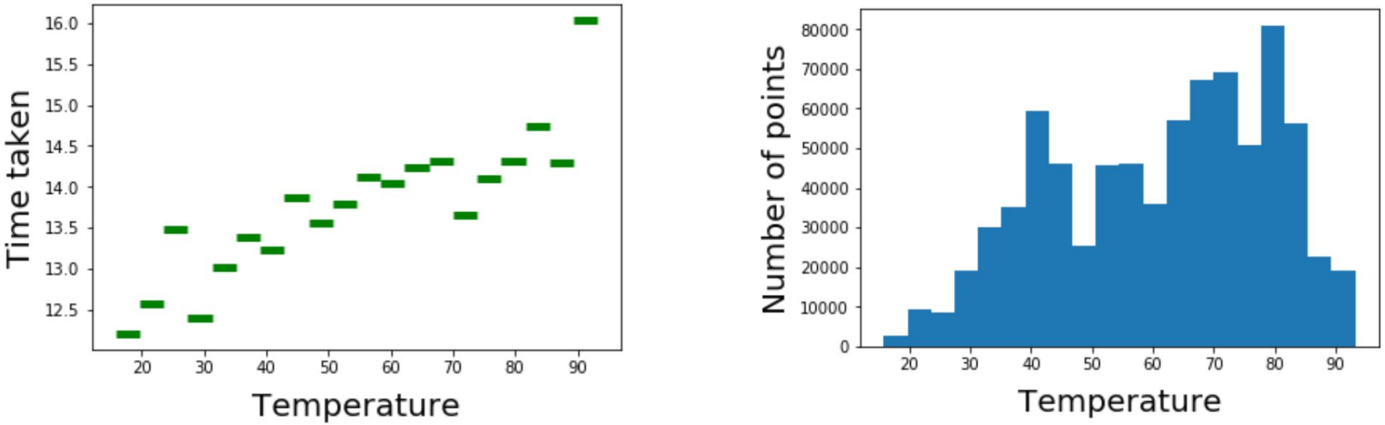


Figure 2: Average travel time for different temperatures, as well as the histogram of temperatures.

Figure by Ritesh Noothigattu

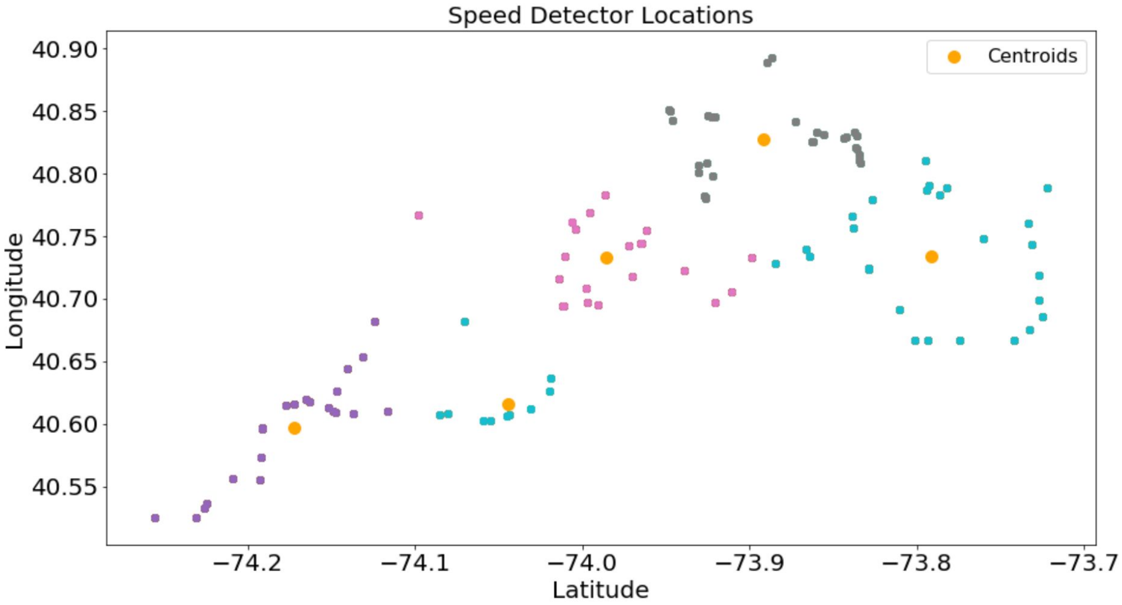
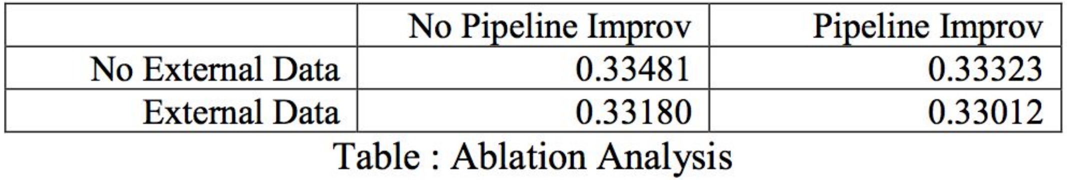


Figure 1: Distribution of speed detectors throughout New York City and their associated centroids.

Figure by Zachary Wojtowicz

“Perform a basic ablation analysis.”

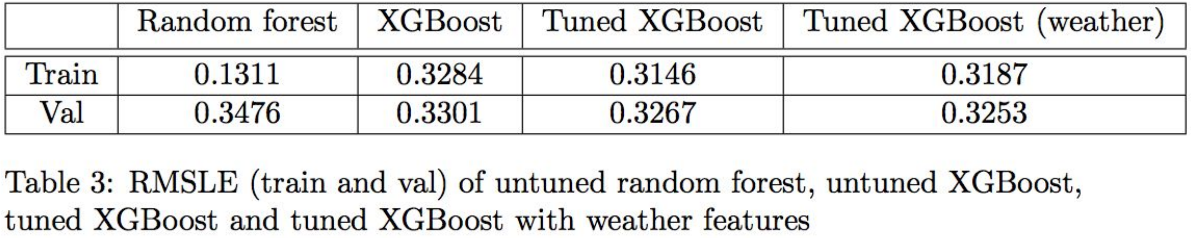
- Students had mixed results when adding external data



	No Pipeline Improv	Pipeline Improv
No External Data	0.33481	0.33323
External Data	0.33180	0.33012

Table : Ablation Analysis

Table by Aditya Galada



	Random forest	XGBoost	Tuned XGBoost	Tuned XGBoost (weather)
Train	0.1311	0.3284	0.3146	0.3187
Val	0.3476	0.3301	0.3267	0.3253

Table 3: RMSLE (train and val) of untuned random forest, untuned XGBoost, tuned XGBoost and tuned XGBoost with weather features

Table by Jie Xie

Any comments?

“Justify your choice of overall pipeline.”

- Most students did quite well in this regard
- The strongest arguments were usually:
    - Improved performance
    - Better computational cost

“Propose concrete and meaningful modifications or extensions to your solution.”

- Better models
- More data (e.g., from previous years)
- Error analysis

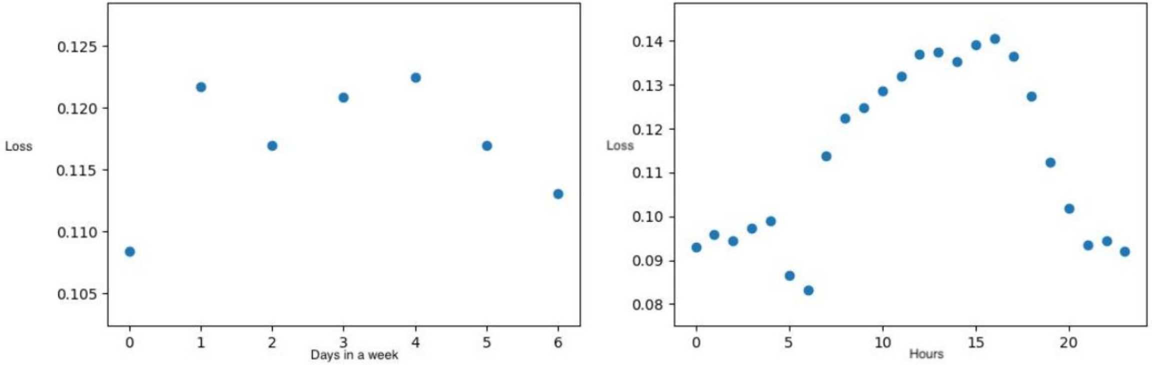


Figure by Fan Yang

“Propose concrete and meaningful modifications or extensions to your solution.”

- Better models
- More data (from previous years, for example)
- Error analysis
- More feature engineering

“Propose concrete and meaningful modifications or extensions to y

- Better models
- More data (fro
- Error analysis
- More feature e

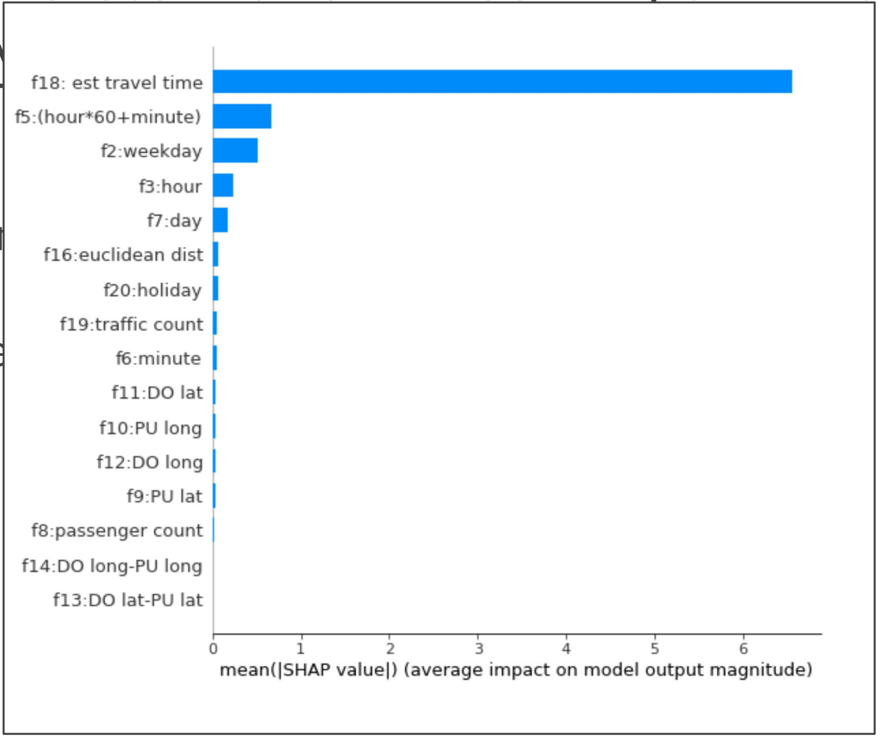


Figure by Jing Mao

Any comments?

This lecture has 2 objectives:

Summarize the students’ solutions to the assignment

Understand how the assignments have related to the **course’s goals**

Typical Steps of Applied Data Analysis

Steps

Step 1

Step 2

Step 3

Typical Steps of Applied Data Analysis

Steps

Overview of research
Some research questions the data might answer
Description of data
Data checks / transfer
Return to questions and translating them
Present to collaborators

Step 2

Step 3

Typical Steps of Applied Data Analysis

Steps

Overview of research
Some research questions the data might answer
Description of data
Data checks / transfer
Return to questions and translating them
Present to collaborators

Simple methods to give preliminary answers
Present to collaborators

Step 3

Typical Steps of Applied Data Analysis

Steps

Overview of research
Some research questions the data might answer
Description of data
Data checks / transfer
Return to questions and translating them
Present to collaborators

Simple methods to give preliminary answers
Present to collaborators

Do better / Iterate
Present to collaborators

Any comments?

We are done!