Taxi Travel Time Prediction

Assignment 2 - Outcome Lecture

Sebastian Caldas and Nicholay Topin

Before we start: a survey!

- Who has done applied machine learning before?

Before we start: a survey!

- Who has done applied machine learning before?
- How much time did you spend on the implementation part of the assignment?

This lecture has 3 objectives:

Summarize the students' solutions to the assignment

Understand how the assignment relates to the **course's goals**

Provide the appropriate **context for the next assignment**

This lecture has 3 objectives:

Summarize the students' solutions to the assignment

Understand how the assignment relates to the **course's goals**

Provide the appropriate **context for the next assignment**

Ksenia Korovina

Zachary Wojtowicz

Global summary

“By 5pm on March 13, 2019, make a submission to Kaggle that beats the baseline.”

- Baseline was a simple “lookup table” approach
  - Calculate “hour block” for each data point: `int(pickup_hour/5)`
  - Features: hour block, PU location ID, DO location ID
  - At test-time, for a *(block, PU ID, DO ID)* tuple, predict average for matching training tuples

“By 5pm on March 13, 2019, make a submission to Kaggle that beats the baseline.”

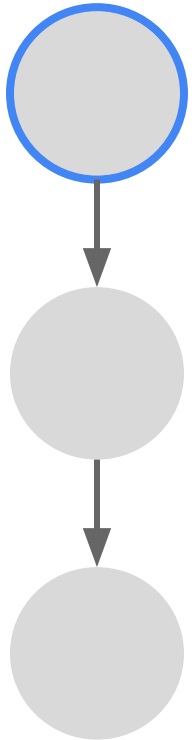
- Baseline was a simple “lookup table” approach
  - Calculate “hour block” for each data point: `int(pickup_hour/5)`
  - Features: hour block, PU location ID, DO location ID
  - At test-time, for a *(block, PU ID, DO ID)* tuple, predict average for matching training tuples

- Boosting and random forests with standard parameters outperform baseline

Any comments?

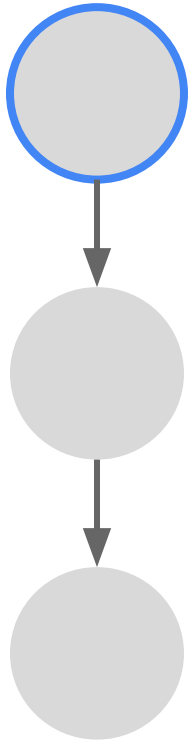
“Describe the pipeline used for your submission and present your results.”

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)



“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” (?)
   - Day of week and hour of day (categorical)
   - Month (?)
   - Minute/Hour of the week
   - Weekday vs. weekend
   - Distance between locations
   - Average time for pick-up/drop-off pair
   - Traffic estimates (count for pick-up/drop-off pair, sometimes hour)



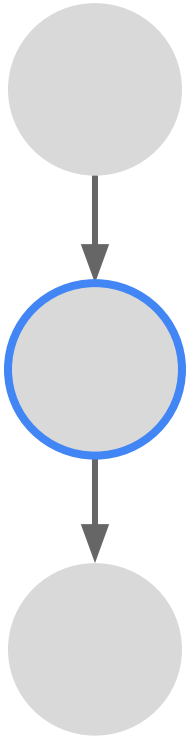

PIT STOP

How can we handle categorical features?

Why did the average time work?

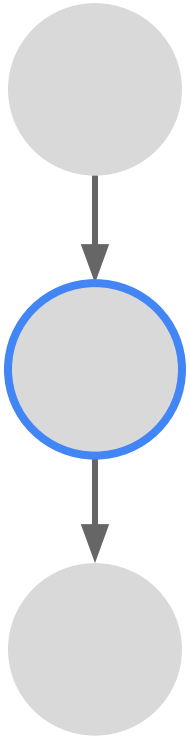
“Describe the pipeline used for your submission and present your results.”

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



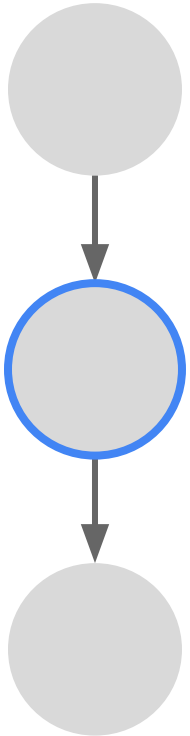
“Describe the pipeline used for your submission and present your results.”

3. Split into train/val sets
    - Test set was given
    - Best estimates if train happened before val
4. Method Selection
    - Random forests (most popular)
    - Boosted trees
    - Nearest neighbors
    - Shallow feed-forward neural network (quite unpopular?)
    - Classifier per pick-up/drop-off pair (sometimes band of day)
        - Requires handling sparsity



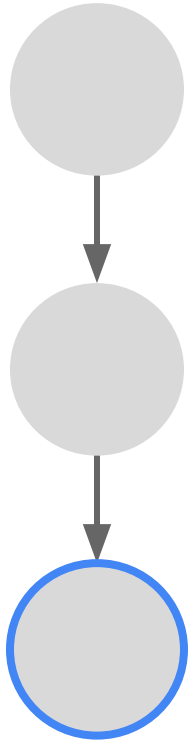
“Describe the pipeline used for your submission and present your results.”

3. Split into train/val sets
    - Test set was given
    - Best estimates if train happened before val
4. Method Selection
    - Random forests (most popular)
    - Boosted trees
    - Nearest neighbors
    - Shallow feed-forward neural network (quite unpopular?)
    - Classifier per pick-up/drop-off pair (sometimes band of day)
        - Requires handling sparsity
    - **Few students had their own baselines.**



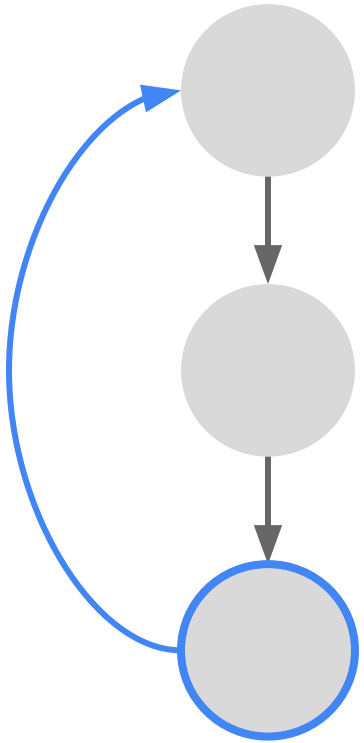
“Describe the pipeline used for your submission and present your results.”

5. Tuning
    - Tune on a developer set (different from train/val)
    - Cross-validation (?)
    - Different hyperparameters per pick-up/drop-off pair (MTL)
    - Pick an extreme value of the grid search (?)
6. Evaluate
    - 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 (?)
    - Different hyperparameters per pick-up/drop-off pair (MTL)
    - Pick an extreme value of the grid search (?)
6. Evaluate
    - Convert back from log-space
    - Evaluate on val set (before submitting to Kaggle)
7. Iterate



Any comments?

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

- The first step is to understand / diagnose your current approach

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

- The first step is to understand / diagnose your current approach

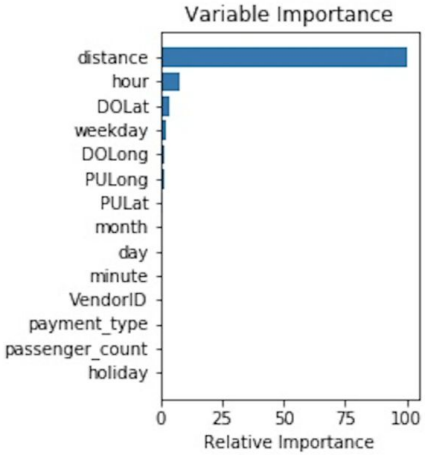


Figure 2: GBRT Feature Importances

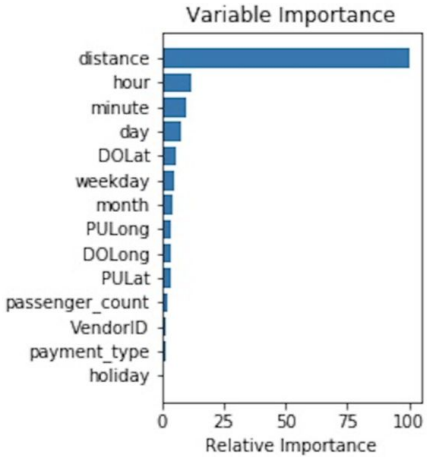


Figure 3: Random Forests Feature Importances

Figures by Jie Xie

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

- The first step is to understand / diagnose your current approach

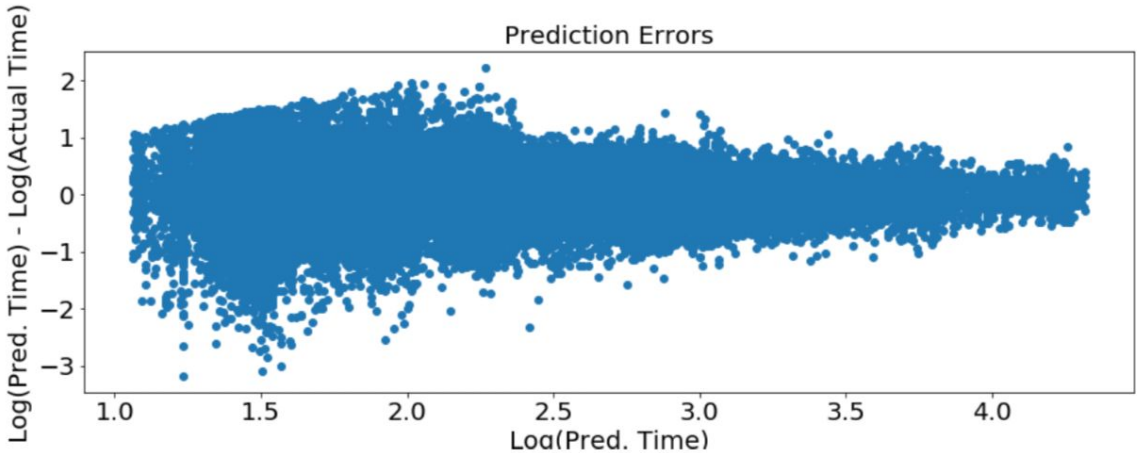


Figure 3: Log prediction errors as a function of predicted travel time. The lack of intercept and slope indiciates the lack of systematic bias in prediction errors.

Figure by Zachary Wojtowicz

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

- The first step is to understand / diagnose your current approach

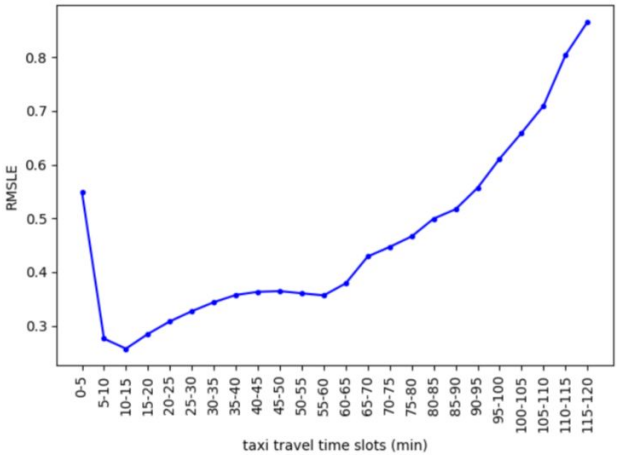


Figure 3: Performance for each travel time slot

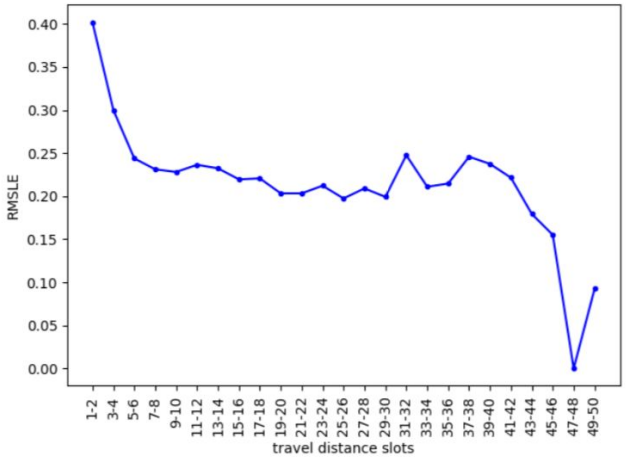
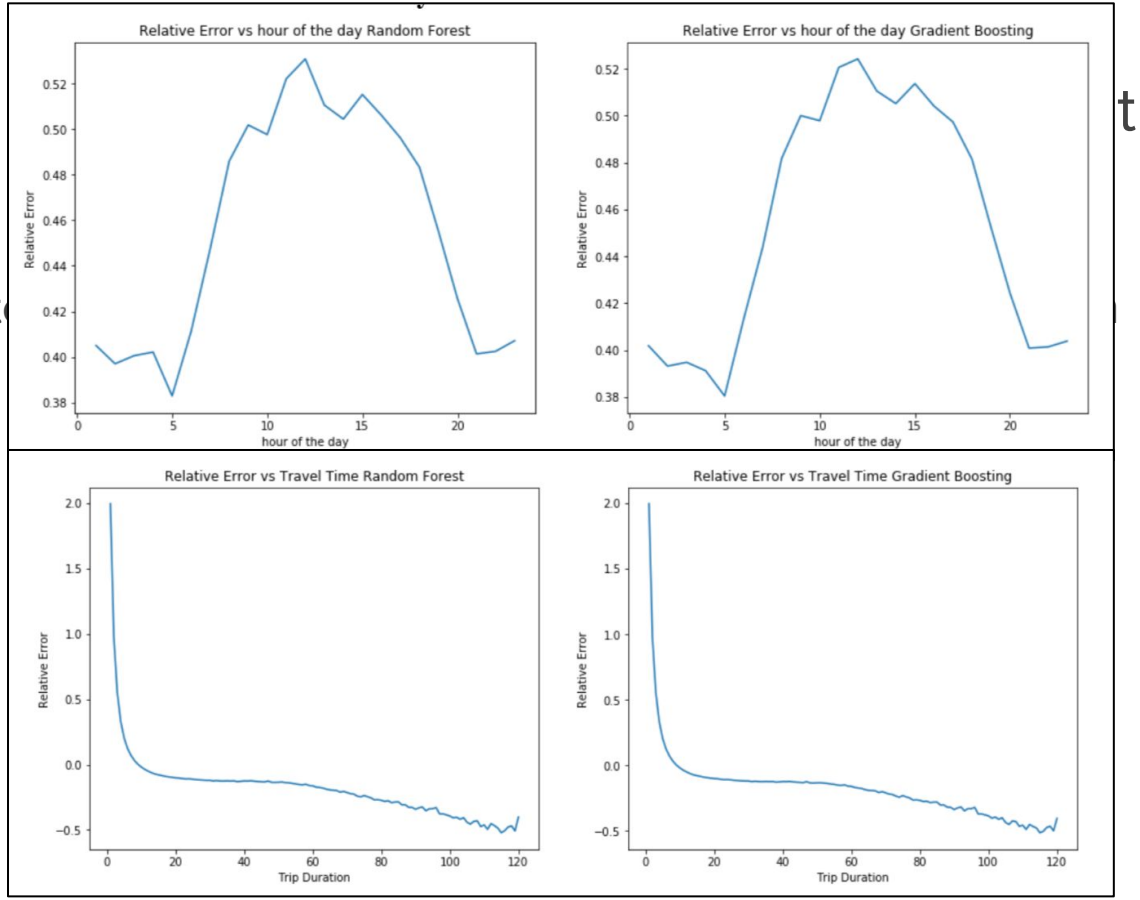


Figure 5: Performance for each distance slot

Figures by Vignesh Kannan

“Propose tions or extensions

- The first st



Figures by Aditya Galada

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

- The first step is to understand / diagnose your current approach

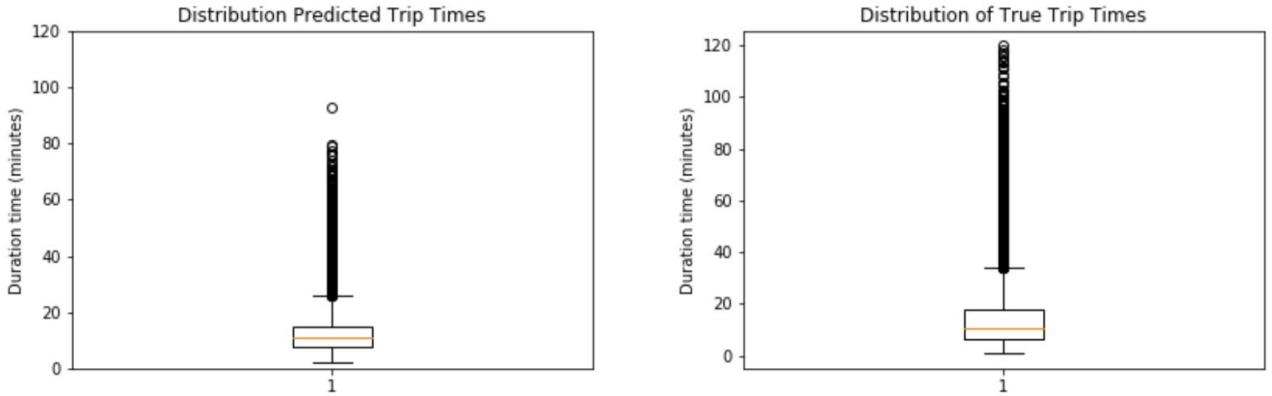


Figure 6: Box plot distribution comparisons for predicted and true trip times

Figure by Neel Guha

Now, how can we do better?

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

- Better features
  - Make sure to include spatio-temporal features
  - Distance and average travel seem powerful but could be redundant

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

- Better features
  - Make sure to include spatio-temporal features
  - Distance and average travel seem powerful but could be redundant

- Better models
  - Properly tuning your current models

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

- Better features
  - Make sure to include spatio-temporal features
  - Distance and average travel seem powerful but could be redundant
- Better models
  - Properly tuning your current models
- More data
  - Subsample more data
  - Random forests seems to plateau after a while
  - External data sources
    - Weather data
    - Traffic data
    - Holidays

Any comments?

This lecture has 3 objectives:

Summarize the students' solutions to the assignment

Understand how the assignment relates to the **course's goals**

Provide the appropriate **context for the next assignment**

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

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

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

This lecture has 3 objectives:

Summarize the students' solutions to the assignment

Understand how the assignment relates to the **course's goals**

Provide the appropriate **context for the next assignment**

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

Assignment 3 will focus on iterating upon your preliminary pipeline

- We will provide you with a **new preprocessed version of the data**.
- We will not impose any restrictions on which pipeline you decide to implement and **you can use external sources of data**.
- **We will provide a set of baselines which you should beat**


kaggle
again!

Just like Assignment 2, Assignment 3 will have two deadlines:

- By the first deadline, you should have a Kaggle submission that beats our proposed baselines
  - Failing to do so will impact your grade
- By the second deadline, you should improve your model and write your report
  - This second deadline is the one previously specified in the course's calendar
- The first deadline will be one week before the second

Just like Assignment 2, Assignment 3 will have two deadlines:

- By the first deadline, you should have a Kaggle submission that beats our proposed baselines
  - Failing to do so will impact your grade
- By the second deadline, you should improve your model and write your report
  - This second deadline is the one previously specified in the course's calendar
- The first deadline will be one week before the second
- **The Kaggle competition is meant to incentivize you**
  - Your grade will not be negatively affected based on your ranking
  - The only exception is failing to beat the given baselines

Any questions?