Taxi Travel Time Prediction

Assignment 1 - Outcome Lecture

Sebastian Caldas and Nicholay Topin

This lecture has 3 objectives:

Socialize 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:

Socialize the students' solutions to the assignment

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

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

Ifigeneia Apostolopoulou

Ian Char

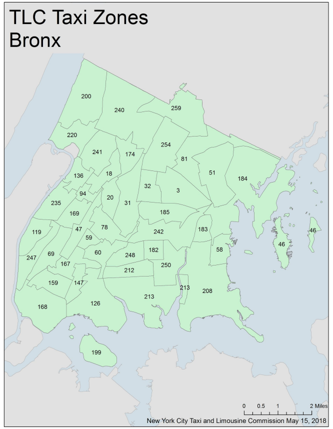
Global summary

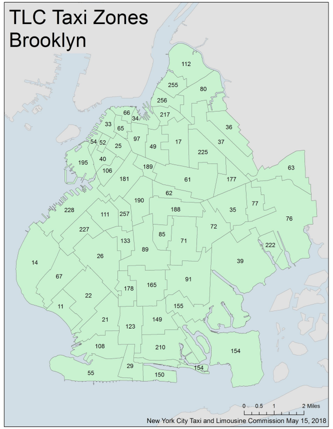
“Familiarize yourself with the data; identify potential difficulties and required cleaning/pre-processing”

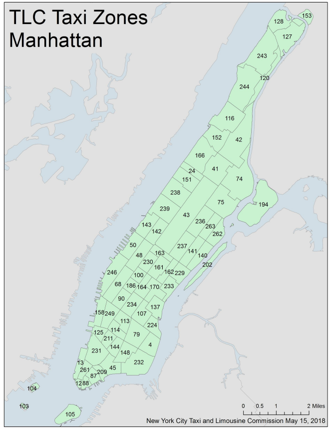
- > 67 million samples
- 7 features
  - Vendor ID {1,2}
  - Tpep_pickup_datetime (date-time format)
  - Tpep_dropoff_datetime (date-time format)
  - Passenger count [1,9]
  - PULocationID [1,265]
  - DOLocationID[1,265]
  - Payment_type [1,5]

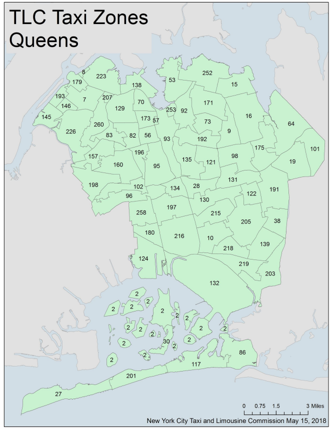
“Familiarize yourself with the data; identify potential difficulties and required cleaning/pre-processing”

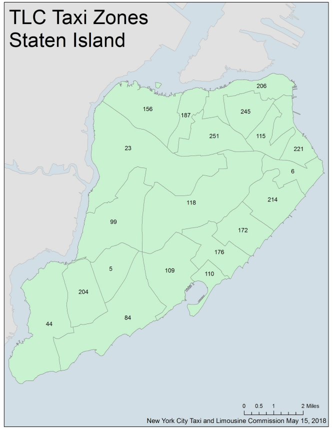
- Locations:
  - Most trips start in Manhattan (61m), Queens (4m), Unknown (1m), and Brooklyn (1m)
  - Most trips end in Manhattan (59m), Queens (3m), Brooklyn (3m), and Unknown (1m)
  - 20 most common locations are in Manhattan (all except LaGuardia and JFK Airport)











“Familiarize yourself with the data; identify potential difficulties and required cleaning/pre-processing”

- Data is given only for certain months (Jan-July) of 2017
  - Any data from another year or month should be removed
- Outliers:
  - Trips with time less than 0 (some students suggested trips under X minutes were outliers)
  - Trips with time more than 60 or 120 or 720 minutes (no trip across NYC is more than 6h)
  - Trips before 2017 / Trips outside of expected month range
  - Trips with 0 passengers (maybe trips with >7 passengers)

“Familiarize yourself with the data; identify potential difficulties and required cleaning/pre-processing”

- Students found that trip time correlated with features such as day of the week, pick up hour and (to a lesser extent) passenger count.

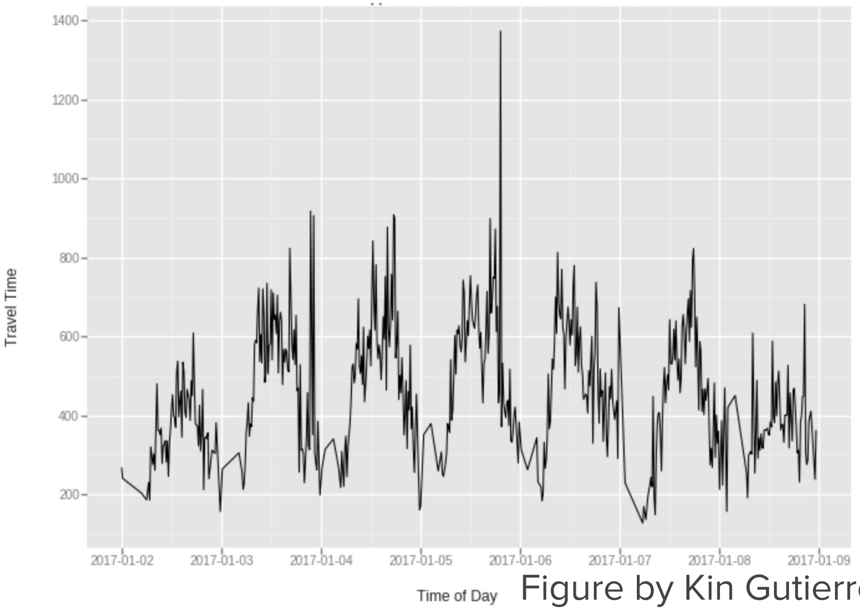


Figure by Kin Gutierrez

“Familiarize yourself with the data; identify potential difficulties and required cleaning/pre-processing”

- 99.95% of trips are between a pair of zones which has at least 5 occurrences.

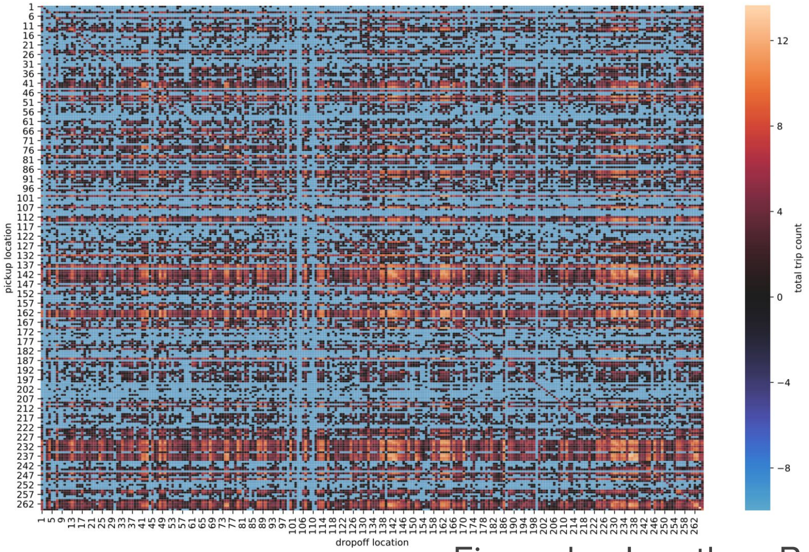


Figure by Jonathon Byrd

Any other interesting findings?

“Formulate a machine learning problem that will help the domain expert achieve their goal”

- Regression problem
  - MSE
  - Root Mean Squared Log Error
    - Avoids large travel times having too large an impact
    - Penalizes underestimates more than overestimates
  - MSE weighted with an underestimate loss
  - MAE, MAPE
  - Huber loss
  - Discretized accuracy (e.g., % within some ‘d’ of actual time)
- To avoid over-penalizing some samples, we can cap the loss.

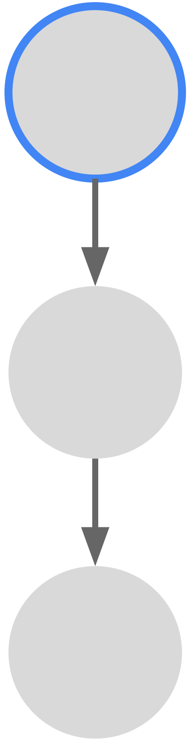
“Formulate a machine learning problem that will help the domain expert achieve their goal”

- Dealing with the dataset’s size:
  - Subsample plus ensembling
  - Divide into distinct tasks (e.g., split 6am-10am predictions into own task, task per pick up location)
  - Use methods with low overhead (data + method fit in memory)
  - Use online methods (e.g., gradient descent)
- External factors:
  - Add external information about weather and holidays
- Train/Val/Test splits:
  - Strangely, people suggested random splits
  - Some suggested withholding last part only (correct!)

Any comments?

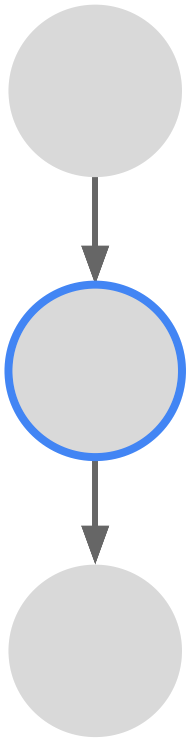
"Propose a detailed analytical pipeline to solve the machine learning problem"

1. Preprocessing
   - Remove outliers
   - Extract travel time from "datetime" columns
2. Feature engineering
   - Distance between locations
   - Split "datetime" columns into day of week and hour of day.
   - Treat "vendor ID" and "payment type" columns as categorical
   - Treat "passenger count" as continuous
   - Remove "payment type" and "vendor ID"



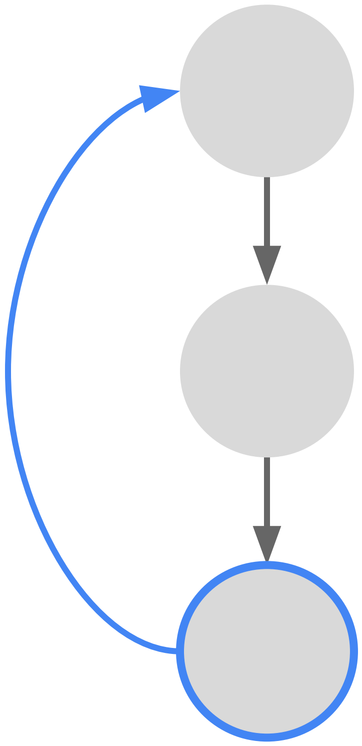
“Propose a detailed analytical pipeline to solve the machine learning problem”

3. Split into train/val/test sets
    - Normalize within each set
4. Potential methods:
    - Linear regression / polynomial regression
    - LASSO
    - Random forests
    - Gradient boosting
    - Nearest neighbor matching
    - Shallow feed-forward neural network
    - ARIMA
    - Bayesian regression (assume log-normal distribution)



“Propose a detailed analytical pipeline to solve the machine learning problem”

5. Evaluate
6. Diagnose
   - For which locations does your pipeline work well?
   - Use different stratifications
7. Iterate!



Any comments?

“Design an experiment to evaluate the effectiveness of your approach”

- Baselines:
  - Most previous methods need finer location information
  - The baselines should be run on the same data
  - A common suggested approach was to use the average trip duration for each pair of pick up and drop off destinations
    - Use a global average for pairs with too little data
- Ultimately, a practitioner will have a real business need that needs to be addressed and should evaluate how the overall solution addresses these needs

Any comments?

This lecture has 3 objectives:

Socialize 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

This lecture has 3 objectives:

Socialize 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 2 will focus on the implementation of a preliminary pipeline

- We will provide you with a **preprocessed version of the data**
- We will not impose any restrictions on which pipeline you decide to implement but **you can only use the given data**
    - Any engineered features must come from this data
    - You should not use any external data (e.g., from other years)

Assignment 2 will focus on the implementation of a preliminary pipeline

- We will provide you with a **preprocessed version of the data**
- We will not impose any restrictions on which pipeline you decide to implement but **you can only use the given data**
    - Any engineered features must come from this data
    - You should not use any external data (e.g., from other years)
- **We will provide a baseline which you should beat**

kaggle

Assignment 2 will have two deadlines

- By the first deadline, you should have a Kaggle submission that beats our proposed baseline
  - 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

Assignment 2 will have two deadlines

- By the first deadline, you should have a Kaggle submission that beats our proposed baseline
  - 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 baseline

We want you guys to do great on Assignment 2!

- We will provide you with **sample submissions** from last semester
  - Different problem
  - Different assignment
  - Still, they give a rough idea of what we are expecting

We want you guys to do great on Assignment 2!

- We will provide you with **sample submissions** from last semester
  - Different problem
  - Different assignment
  - Still, they give a rough idea of what we are expecting
- For the students that didn't do so well on Assignment 1:
  - Look at the sample submissions and come to office hours

We want you guys to do great on Assignment 2!

- We will provide you with **sample submissions** from last semester
  - Different problem
  - Different assignment
  - Still, they give a rough idea of what we are expecting
- For the students that didn’t do so well on Assignment 1:
  - Look at the sample submissions and come to office hours
- For the students that did well:
  - Keep up the good work!

Any questions?