GA Individual Project

NYC Taxi Shareability & Fare Prediction

HANS HAN 02/03/2016

Contents

- What the project is about. (1 min)
- Why did I choose this project? (1 min)
- The results and demo. (2 mins)
- What steps did I go through? (3 mins)
- What I've learnt about the data? (1 min)
- Things that I'd like to continue to improve. (2 mins)

What is this Project about?

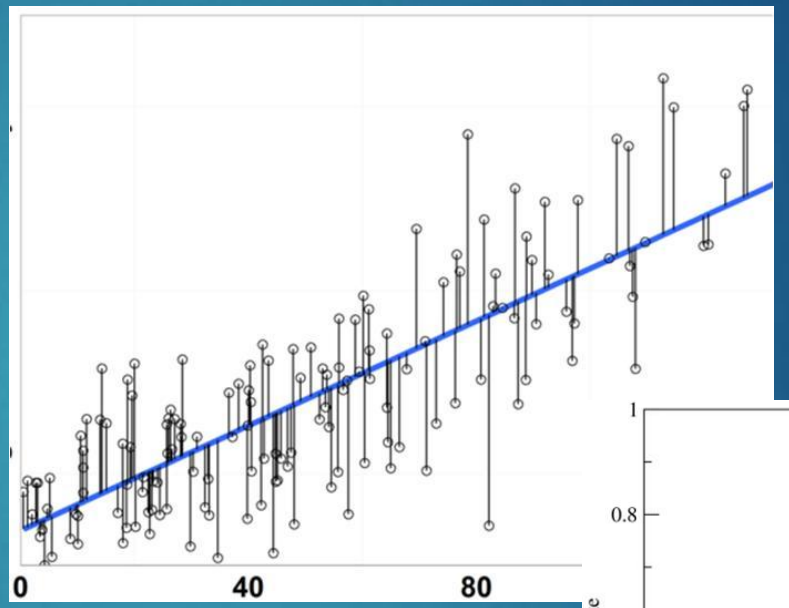
- To predict the taxi fare price based on historical data
- To classify taxi trips for their probability of shareability based on a average 10 minute waiting window using historical data
- To quantify the benefits of sharing based historical data (in- progress)
- Visualizations to help users better understand the complexity and volume of data

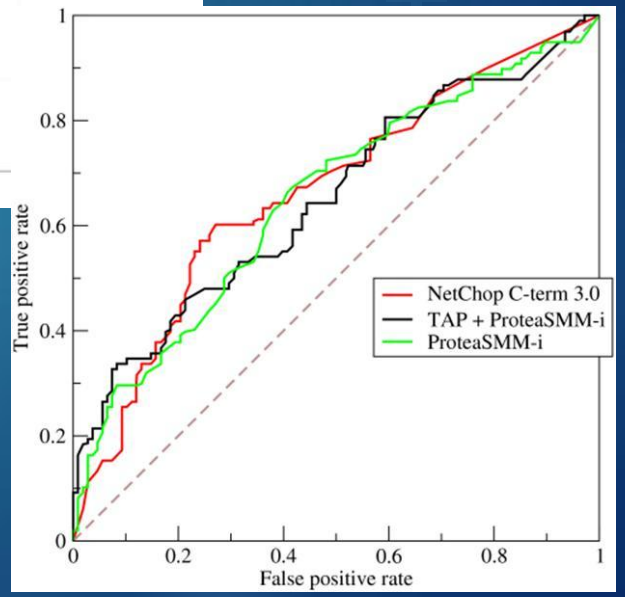
Why did I choose this project?

- Big publicly available, rich dataset with many visualizations done on it previously
- Not many interactive visualizations using this data
- Not many predictive models using this data
- Good practice with geo location data and mapping them
- Apply CRISM-DM methodology and combine machine learning and visualization

The Results and Demo

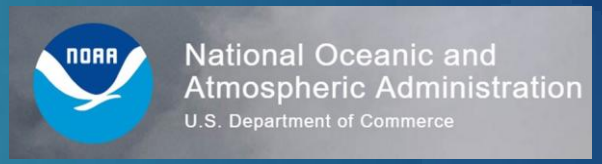
- Taxi fare prediction
  - R^2: Average 0.78
  - Mean Squared Error: $2.31
- Taxi shareability prediction
  - Accuracy: 0.77
  - Area Under the Curve: 0.76
- Demo of results





Steps to Get Through 1

- The data
  - New York City Taxi Trips for February 2013
    - Trip data and fare data
    - 14 million raw observations
  - National Oceanic and Atmosphere Administration
    - Hourly rain fall for February 2013

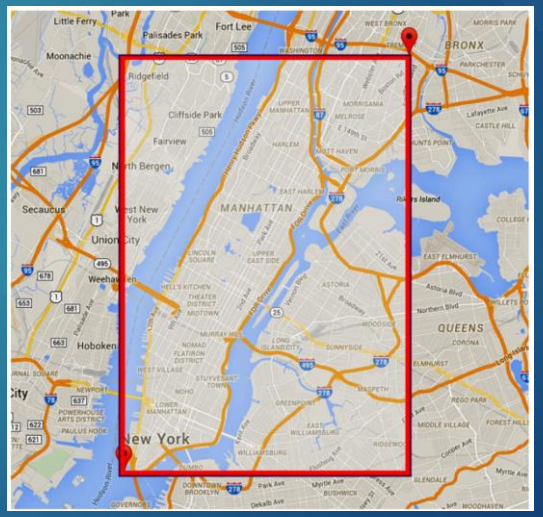


- The software stack
  - Python/Anaconda: Jupyter Notebook
  - Html, CSS for web contents and styling
  - JavaScript for user interactions with web
    - Leaflet (JavaScript library) for geo location data
    - D3 for slider control
  - Flask for linking python with web development

Steps to Go Through 2

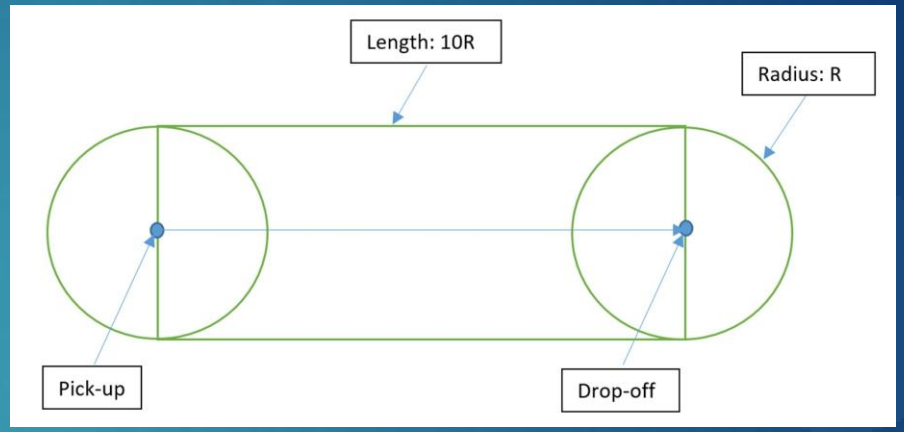
- Summarizing data
  - Data types
  - Statistical summaries
- Cleaning data
  - Erroneous entries in all dimensions
  - Null values in geo locations
- Rescoping project
  - Manhattan area (longitude and latitude)
  - Single passenger trips only (71%)
  - Trip time constraint
- Wrangle data
  - Merge on columns between 3 datasets

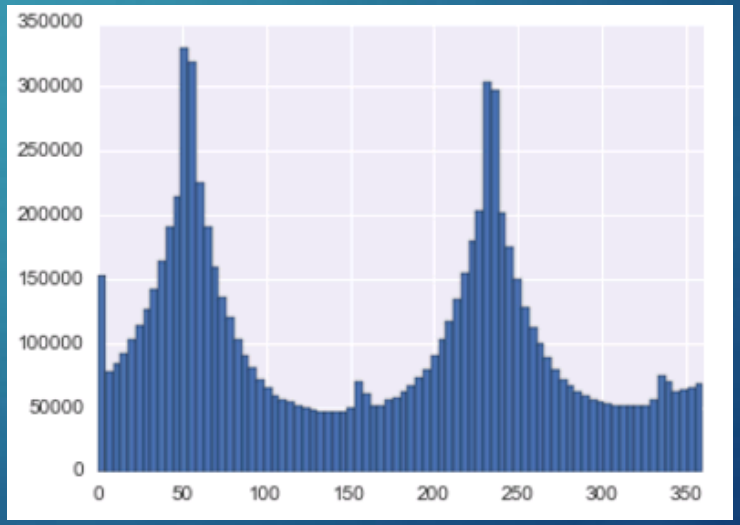
```
1      9942847
2      1863442
5       842936
3       555313
6       525664
4       259761
0          202
7            5
9            2
8            2
208          1
129          1
Name: passenger_count, dtype: int64
```



Things that I Tried or Learned

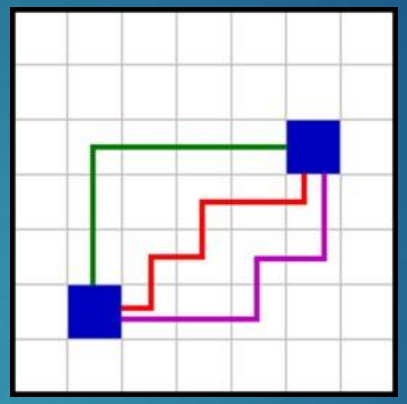
- Shareable trips for each trip
  - Too complex, simplified twice
  - Time and space domains
- Direction of travel
  - Two prominent ones
- Day of the week
  - Much less traffic on Sundays
- Hours of the day
  - Lowest at 4am
  - Peaks at 8am and 6pm
  - Obvious time-series cycles within

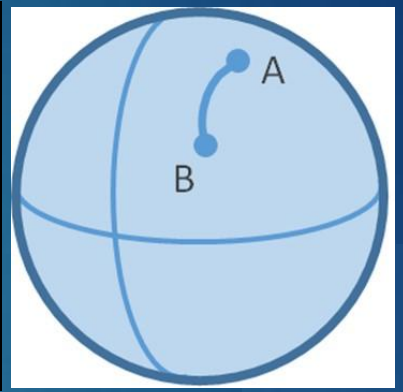


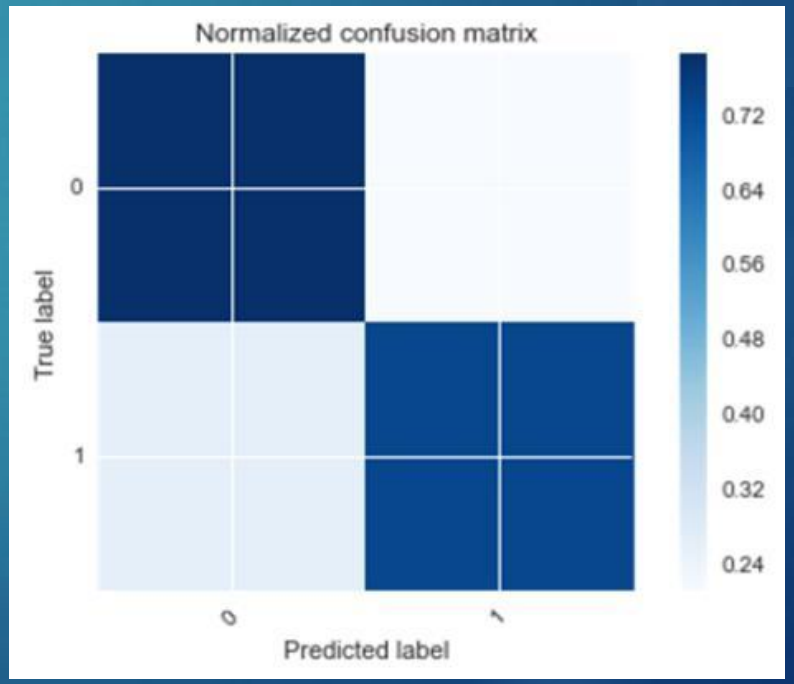


Steps to Go Through 3

- Feature engineering
  - Manhattan distance (pun!)
  - Haversine distance
  - Angle of travel direction
- Modelling
  - Regression
  - Classification
- Cross-validation
  - Train-test split
  - K-fold CV
- Model Performance
  - R^2 across 5-folds: SD of .001
  - Confusion Matrix







Future Improvements

- Run a year's worth of data to capture seasonality with Spark
- Combine the two web apps into one and improve user friendliness
- Include potential shareability savings based on a year's data
- Enrich with other datasets such as subway entry and exit locations and neighbourhood geojson boundaries on Open NYC Data website
- Better predictive model for probability to share within a acceptable time window
- Better look at the raw data and explore for more interesting insights and ideas to work on

Thank You!

Questions?