CurrencyFair NYC Taxi Challenge

James Lawlor

Brief Introduction

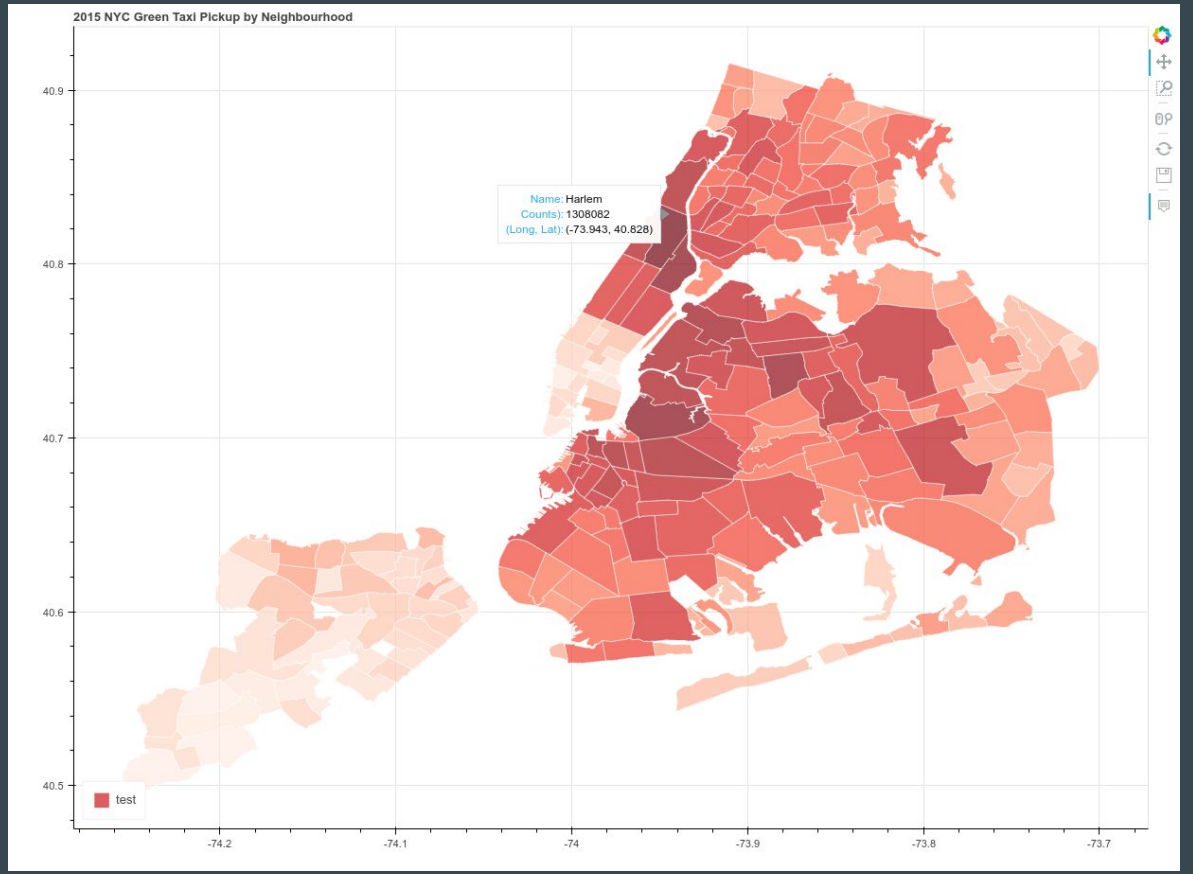
- “Green taxis” serve **outer NYC areas**
- Dataset contains **~20 million journeys** made in 2015
- Features include pickup/dropoff times, locations, tip amounts etc.
- **Goal** is to use the data to **find recommendations** for a driver

Suggestion 1 - Stick to busy areas

- Pickup Lat/Lon coordinates mapped to NYC neighbourhoods.

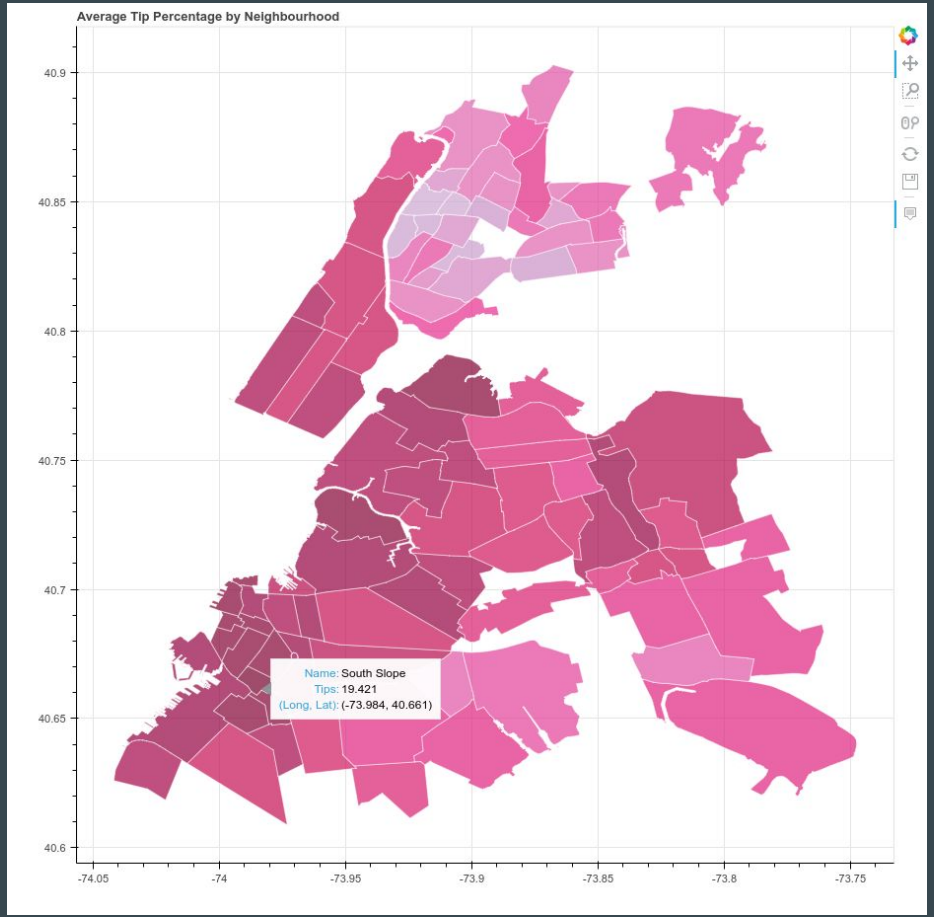
- Most popular is Harlem (~130k), followed by East Harlem and Williamsburg (~70k).

- Map generated using Geohashing, JSON shapefiles and the Python library Bokeh.



Suggestion 2 - Optimise tips

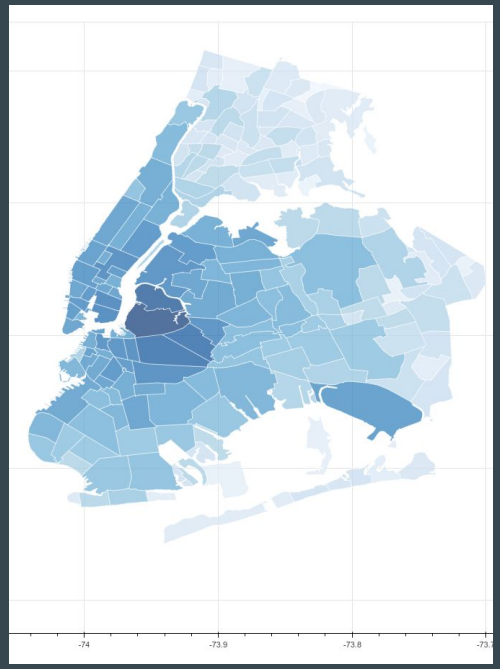
- **Higher tipping averages** in wealthier neighbourhoods (North Brooklyn, Park Slope, North Manhattan island)
- Tips are **~15% in Harlem** area, compared to **~18-19% in NW/W Brooklyn.**
- Assuming a tip culture, a driver will want to focus where tips are highest.
- Appears to be some **discrepancy between vendors** in the way tips are calculated.



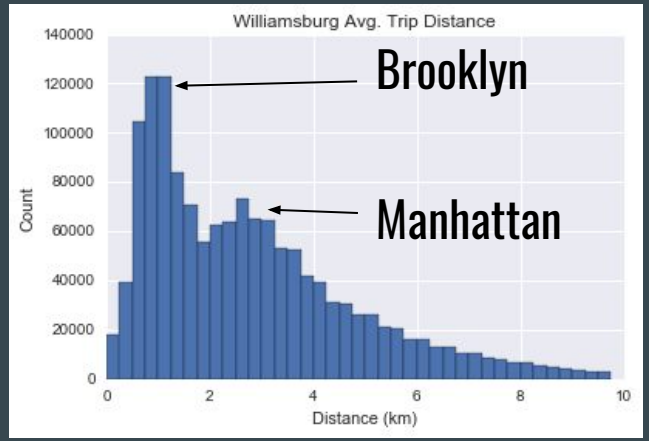
NB: Low frequency areas omitted

Williamsburg in detail

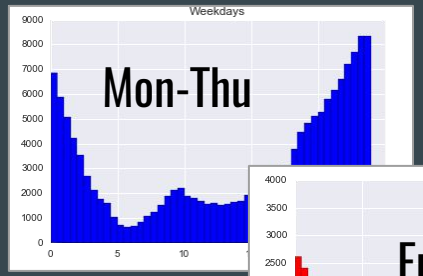
Popularity of destinations

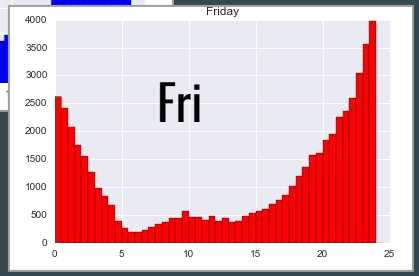


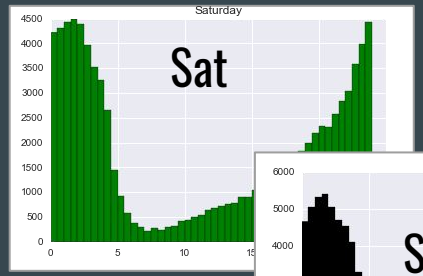
Trip distances

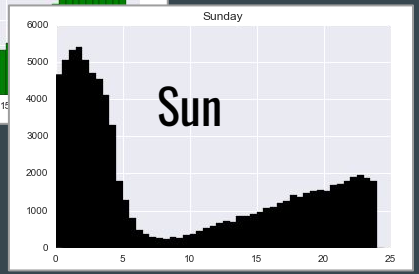


Frequency of pickups by time of day



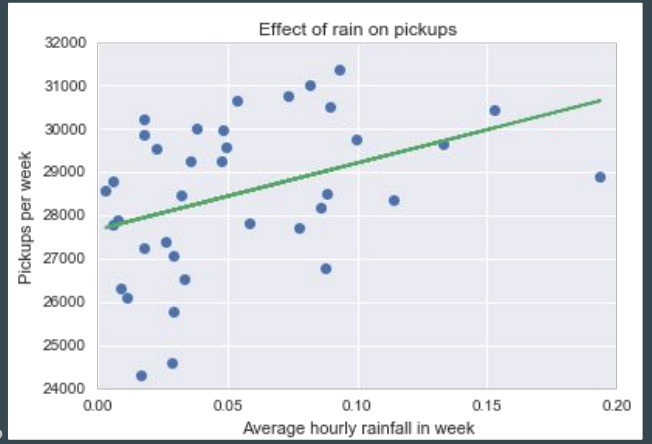




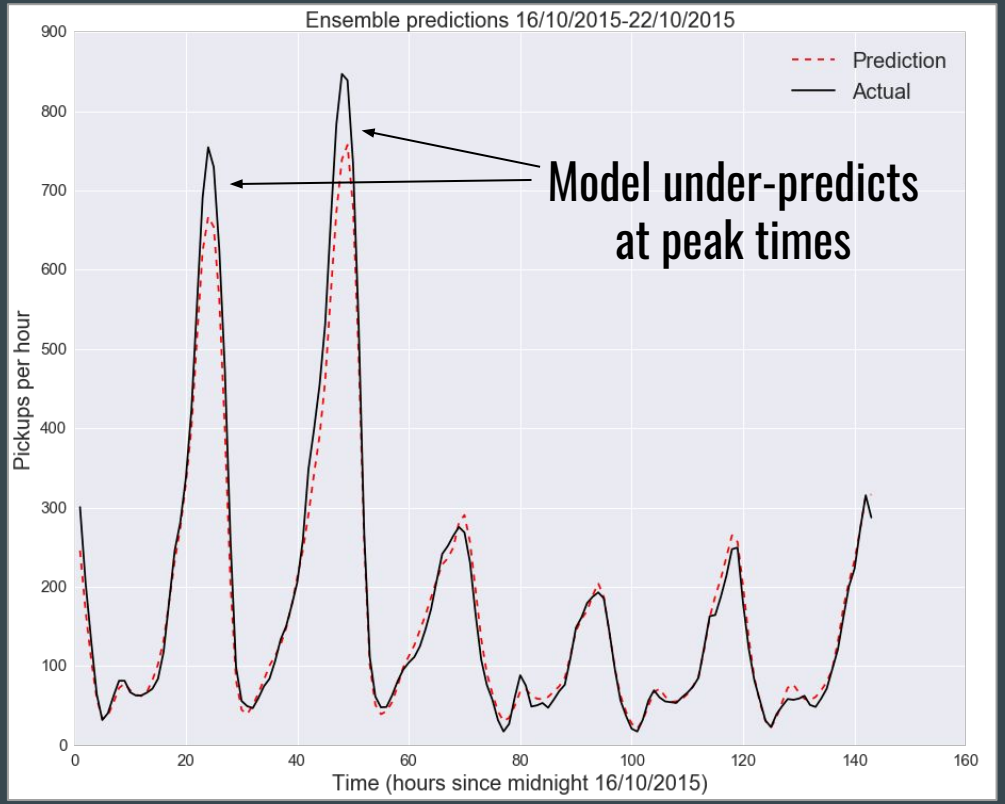


Suggestion 3 - Use forecasting

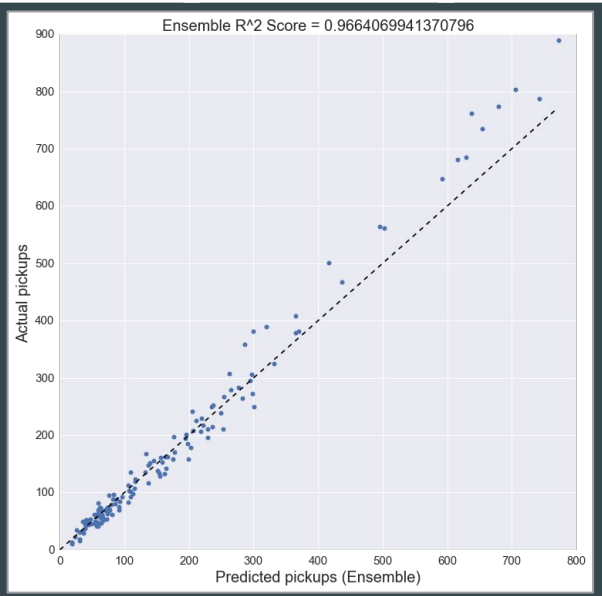
- Task - predicting hourly pickups in Williamsburg
- 3 very different models - Historical average; Random Forest; ElasticNet.
- Trained on Jan -> mid-October, tested on week 16th-22nd October (misses Halloween, Thanksgiving, Xmas season effects)
- Engineered time features, one-hot-encoding for ENET + hourly weather data from Central Park
- 5-fold cross-validation used for hyperparameters



Modelling results



Model	RMSE	R^2 Score
Benchmark	30.7	0.965
Random Forest	29.9	0.967
Elastic Net	30.4	0.966
Ensemble	30.1	0.967



A “Hot Zones” App

- Forecasting busy areas via Mobile App
- Subscription service @ $X/month
- Deploy using AWS Lambda & Amazon API Gateway - cheap and scalable
- Could integrate effect of sports events, concerts etc.
- Data available since 2013 for Green taxis, could further train model + account for seasonality

Thanks for listening!