Optimizing Citi Bike

Aleksandra Vercauteren

Rebalancing Operations

- Avoid empty bike stations
- Avoid full bike stations
- Relocate bikes efficiently

 We need to know the flow

Data

- Citi Bike Trip histories
- Hourly weather data (NOAA)

More than 37 million data points...

Data

- June 2016
- Frequently visited stations (> 100)
- Registered users
- Weekdays

Total: 897 250

Targets: 350

Variables

Target	End station
Predictor	Start station coordinates Start time (epoch) Birthyear Weather data (wind speed, visibility, temperature, precipitation)

Results

- Best model: Decision tree (with parameter tuning)
- Accuracy: 7% (baseline: < 1%)
- Problem: massive kernel death when fine tuning

Data

- June 2016
- Only frequently visited stations (>100)
- Registered Users
- Only weekdays
- Female

Total: 58 063

Targets: 92

Results

- Weighted Voting Classifier with Decision Tree and Random Forest: accuracy of 13% (baseline: 1%)

- Feature importance:
  - Start time
  - Birth year

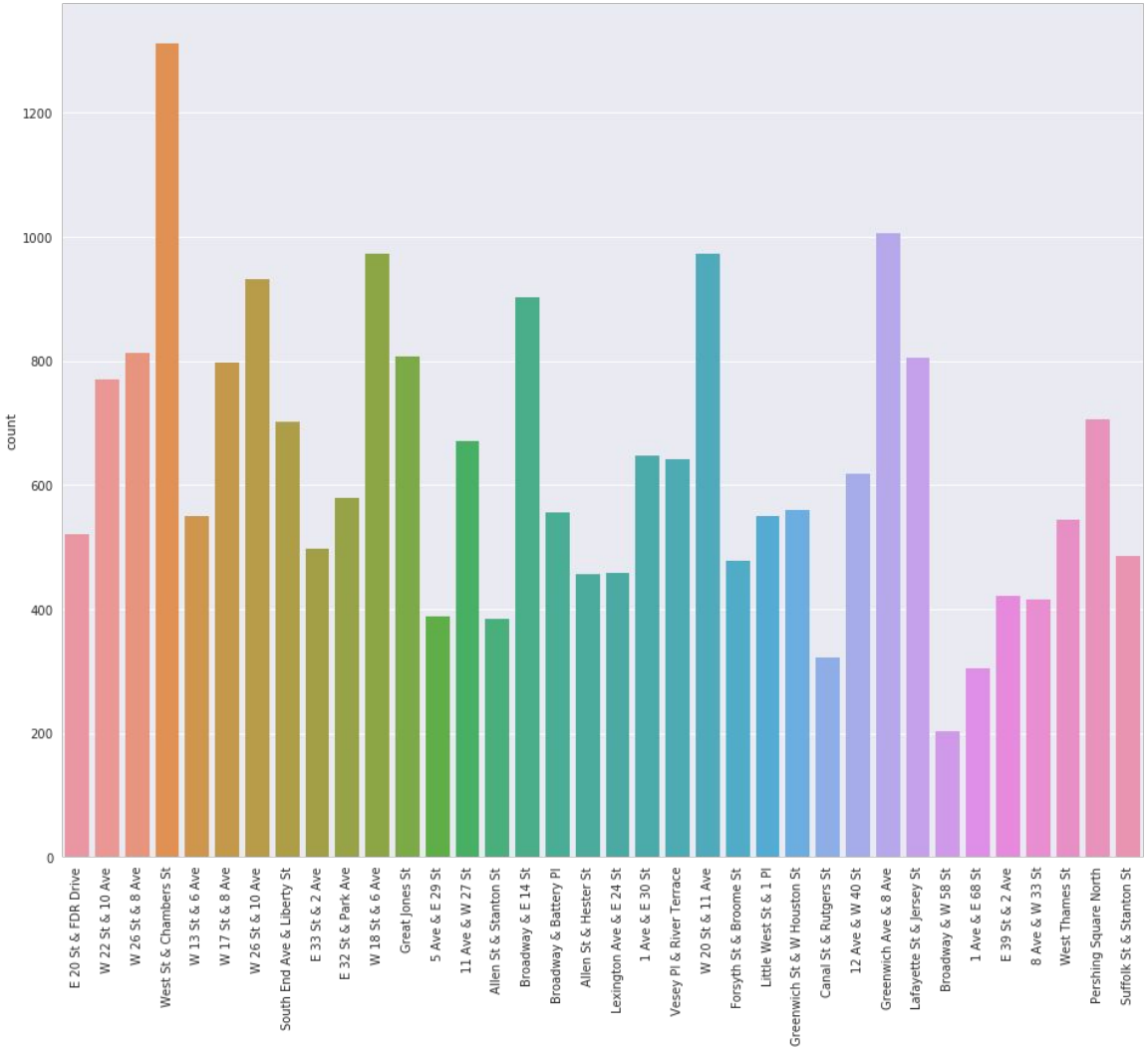
Startstation?

- Coordinates were used: latitude and longitude treated as independent features

  In the future: transform start station name to dummy variable.

- Well predicted end stations are simply more frequent?

Frequencies of well predicted endstations (f1-score > 10%)



Visualisation

This is just the beginning!

- Multi-output classification: end station and time
- More data (and CPU/GPU)
  - Compare with weekends
  - Other months

This is just the beginning!

- Multi-output classification: end station and time
- More data (and CPU/GPU)
  - Compare with weekends
  - Other months

Thank you!

github.com/alvercau