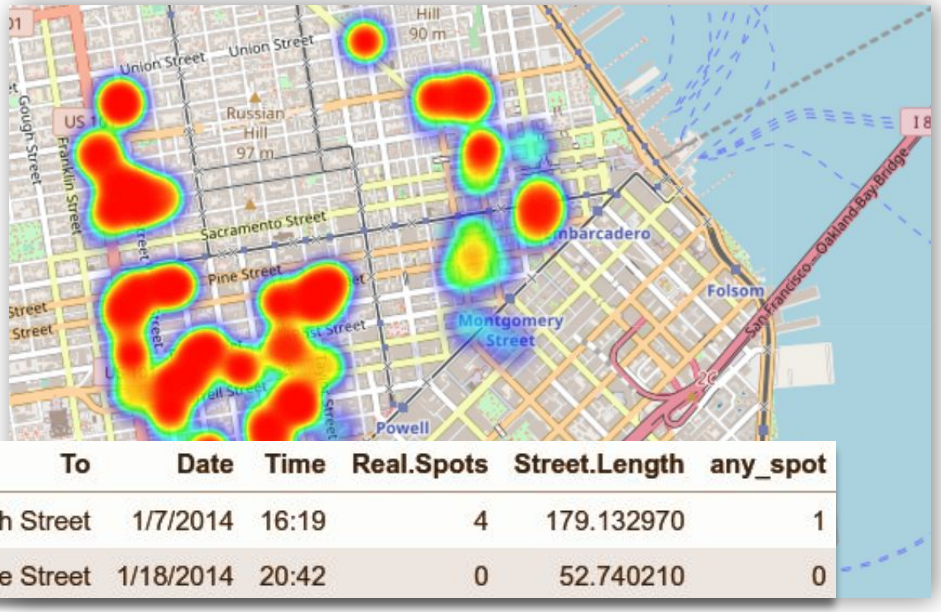
San Francisco Parking Spot Prediction

- Course project of advanced machine learning -

Team TensorFlowBOYS:
Yiqiang Zhao & Fang Wang

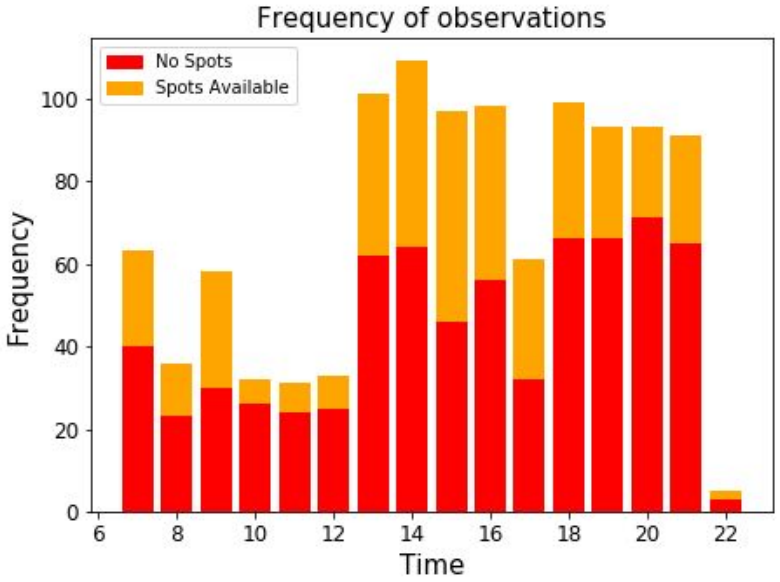
Data

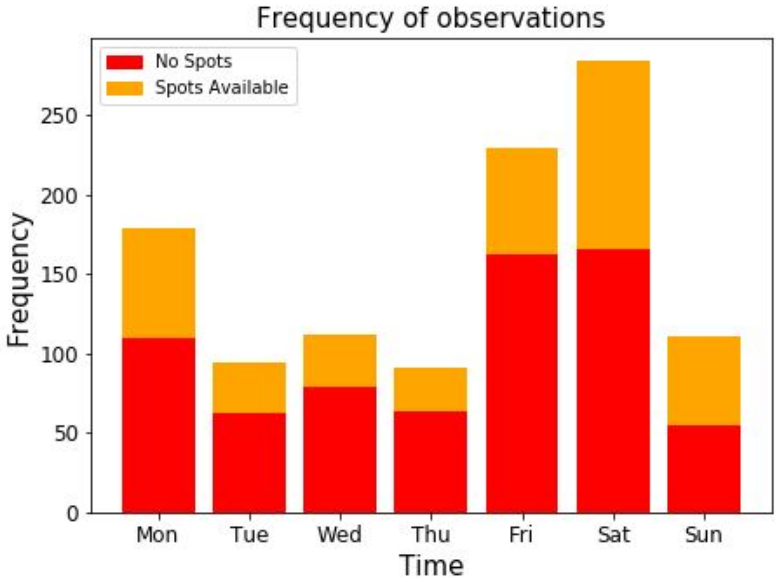
- **Training set 1100 rows**
  - Date range: 1/18/2014 - 3/9/2014
  - 70 identified streets
- **Test set 726 rows**
  - Date range: 3/28/2014 - 11/4/2016
  - Same streets as training set



Street	From	To	Date	Time	Real.Spots	Street.Length	any_spot
Mission Street	25th Street	26th Street	1/7/2014	16:19	4	179.132970	1
Polk Street	Ellis Street	Olive Street	1/18/2014	20:42	0	52.740210	0

EDA


Frequency of observations
No Spots
Spots Available
Frequency
0
20
40
60
80
100
6
8
10
12
14
16
18
20
22
Time


Frequency of observations
No Spots
Spots Available
Frequency
0
50
100
150
200
250
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Time

Feature Engineering

- Standard process:
  - Concatenate “Street”, “From”, “To”
  - Parse Time variable to Day of Week, Weekdays/Weekends, Hour
- Mean encoding:
  - Street
  - Day of Week
  - Weekdays/Weekends
  - Hour

Hour → Half Hour → 15 min

Feature Engineering

- External data: SF parking meters data (number of spots on a street)

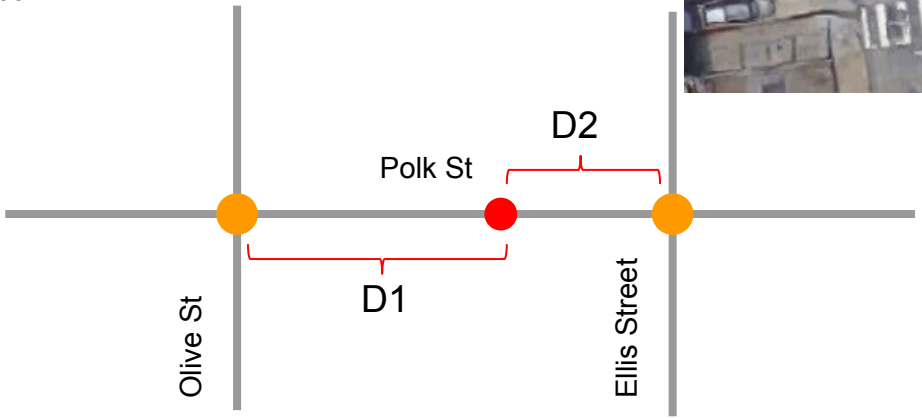
Lat	Lon	DateTime
37.771637	-122.437112	2016-08-23 15:53:57
37.777229	-122.465370	2016-04-05 01:56:17
37.771172	-122.437683	2016-04-22 18:25:54
37.777328	-122.465012	2016-05-03 05:10:29

Street	From	To
Mission Street	25th Street	26th Street
Polk Street	Ellis Street	Olive Street

(Polk St & Ellis St) → (37.784039, -122.419383)

Feature Engineering

- Parking meter
- Street intersection



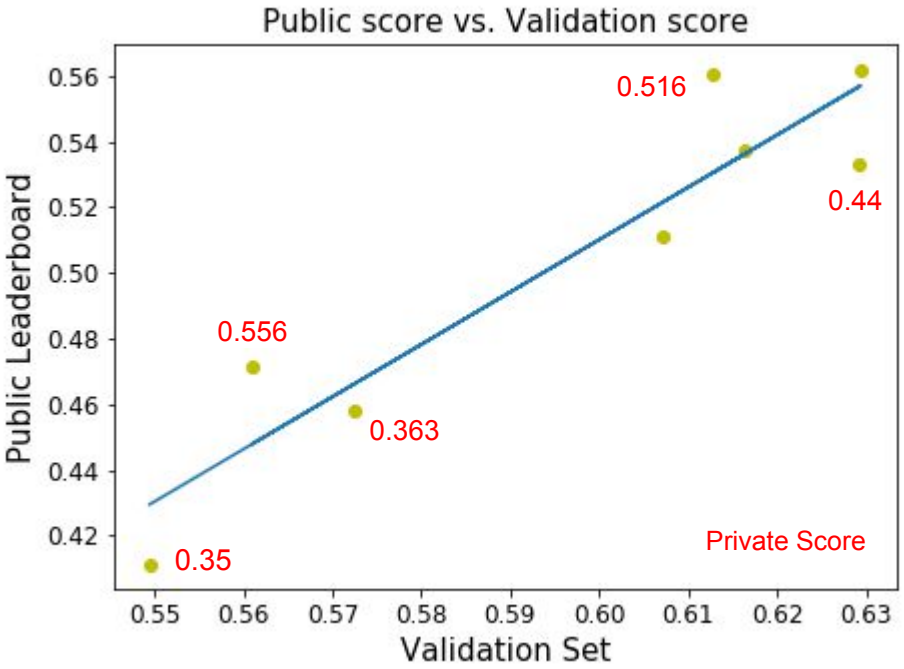
D1 + D2 ~= Street Length

Validation Set

Split the data based on temporal order

Training set: 800 rows (36.5% ‘1’s)

Validation set: 300 rows (36.3% ‘1’s)



Validation Set

Random Forest:
n_estimators=200
min_samples_split=50
max_features=0.5
max_depth=5

oob score: 0.741
train data accuracy: 0.773
val data accuracy: 0.740
train data F0.5 score 0.699
val data F0.5 score 0.645

Random Forest Scores

Public Board Score	Private Board Score	Difference
0.566	0.550	-0.016
0.560	0.541	-0.019
0.577	0.551	-0.026
0.556	0.556	0
0.565	0.586	+0.021
0.540	0.526	-0.014
0.560	0.554	-0.005

Model Selection

Public	Private
0.57831	0.58080

Models trained: XGBoost, logistic regression, and random forest

Ensemble: majority voting -- 5/10 votes count (only ~140 ‘1’s out of 726)

Model	Publicboard F0.5 Score	Validation F0.5 Score	Validation Precision	Validation Recall
Random Forest	0.556	0.682	0.532	0.646
Logistic Regression	0.578	0.610	0.688	0.624
XGBoost	0.547	0.670	0.504	0.629
Ensemble	0.578	/	/	/

Takeaways

- Use a simple model for a small dataset.
- Have a reliable validation set.
- Trust your validation score more than your public score.
- Be more cautious to predict positive as F0.5 punishes heavily on false positive.
- Communicate with other teams more and you learn more.

Thanks!

Yiqiang Zhao: yzhao83@usfca.edu

Fang Wang: fwang18@usfca.edu