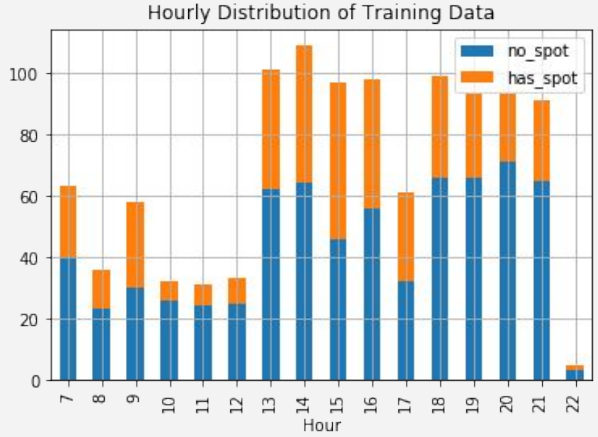
InstaPark

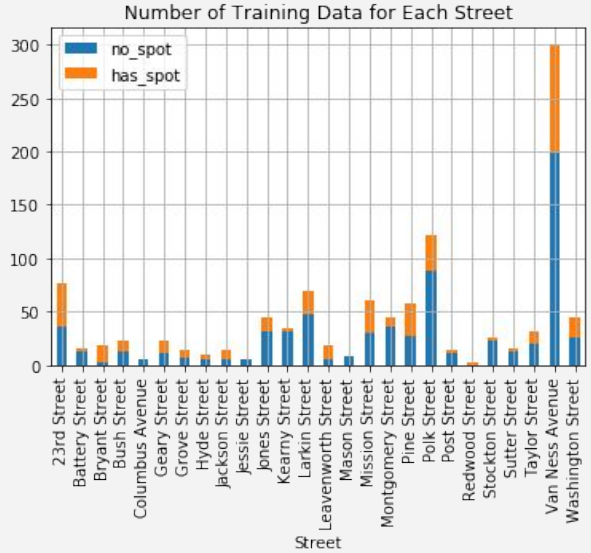
Parknav Parking Prediction

Guoqiang Liang, Jingjue Wang

Exploratory Data Analysis

- Training Data
  - 1,100 data entries from Jan 18, 2014 to Mar 9, 2014
  - No data from 11pm to 7am
  - 26 Streets in total (1/3 data goes to Van Ness Avenue)
- Meter Data
  - All data points are in first half of day originally
  - Shift timestamps by 7 hours
  - Few data on Sunday (might be free parking)
- Sensor Data
  - Different street coverage from training set
  - Didn't use this dataset in the end





Feature Engineering

- Concatenated “Street”, “From”, and “To” columns as a feature named “address” .
- Target mean encoding address features.
- Extracted day of a week from “Date” variable.
  - Monday was encoded as 0, and Saturday was encoded as 6
- Convert “Time” from a timestamp to numerical variables by using the following equation:
  - Converted_time = hour * 2 + round(minute / 30.), treat 30 minutes as a unit.
  - Ex: “13:05” will be converted as 13 * 2 + round(5 / 30.) = 26.0

Other Data Sources

- Google Geocoding Public API
  - Collected geographic coordinates for the two endpoints of a street
  - Calculated the middle point's coordinates by using the two endpoints' coordinates of a street
  - Performed K-means Clustering on the middle point's coordinates to cluster streets into 6 clusters.
- Yelp Fusion Public API
  - Collected the number of restaurants within 322 meters from the middle point of a street.
  - Calculated the average restaurant ratings by using the following equation:

$$AvgRating_s = \frac{\sum_{i=1}^{N} rating_i \times NumReview_i}{\sum_{i=1}^{N} NumReview_i}$$

where s: a given street, i: a restaurant, N: number of restaurants around the given street.

- Parking Record Data Set
  - Matched the meters in the parking record data set with the streets in the training data set.
  - Extracted the hourly parking occupied percentage for seven days in a week as new features.

Model & Evaluation

- No Dedicated Validation Set
  - Hard to split the train and validate
  - Highly unstable results depending on how we split the data
  - Use 5 fold cross-validation with customized metric (F0.5) instead
- Gradient Boosting
  - Bad results due to limitation of data size
  - F0.5 score of 0.4 ~ 0.5
- Random Forest
  - Cross-validated F0.5 score of 0.58 (Kaggle Public LB: 0.61)
  - Feature importance
    - Time, Address (various encoded values), Day of Week, Street Length
    - Number of Restaurant, Average Ratings
    - Parking records have little importance ...