Business Case

San Francisco taxi cabs dataset analysis

The questions

CO2 EMISSION REDUCTION

Calculate the potential for yearly reduction on CO2 emissions, caused by the taxi cabs roaming without passengers.

NEXT PICKUP PREDICTION

Build a predictor for taxi drivers, predicting the next place a passenger will hail a cab.

CLUSTERING

Identify clusters of taxi cabs.

CO2 EMISSION REDUCTION POTENTIAL

Assumptions

- The taxi cab fleet is changing at the rate of 10% per month (from combustion engine-powered vehicles to electric vehicles).
- the average passenger vehicle emits about 404 grams of CO2 per mile.

Raw data

	latitude	longitude	occupancy	time	taxi_id
0	37.78615	-122.40628	1	2008-05-17 10:00:42	new_auctjir
1	37.78620	-122.40181	1	2008-05-17 10:01:43	new_auctjir
2	37.78629	-122.39917	1	2008-05-17 10:02:43	new_auctjir
3	37.78618	-122.40140	1	2008-05-17 10:03:53	new_auctjir
4	37.78655	-122.40231	1	2008-05-17 10:04:54	new_auctjir

Processed data

	latitude	longitude	occupancy	time	taxi_id	trip_id
0	37.75153	-122.39447	0	2008-05-17 14:12:10	new_abboip	3bc085c8-feb5-476a-8f22-19560c0e1349
1	37.75149	-122.39447	0	2008-05-17 14:13:34	new_abboip	3bc085c8-feb5-476a-8f22-19560c0e1349
2	37.75149	-122.39447	0	2008-05-17 14:14:34	new_abboip	3bc085c8-feb5-476a-8f22-19560c0e1349
3	37.75149	-122.39446	0	2008-05-17 14:15:35	new_abboip	3bc085c8-feb5-476a-8f22-19560c0e1349
4	37.75144	-122.39449	0	2008-05-17 14:41:43	new_abboip	3bc085c8-feb5-476a-8f22-19560c0e1349

Aggregated data

trip_id	taxi_id	distance	duration	start_longitude	start_latitude	end_longitude	end_latitude	occupancy	start_time	end_time
000a7016-95c0-4a6f-afbd-4eb6a0e0f4e3	new_abgibo	0.543472	3.750000	-122.44625	37.79566	-122.44964	37.79824	0	2008-05-21 00:18:05	2008-05-21 00:21:50
001403cc-e3ca-4b79-91d2-ecf59f160905	new_abdremlu	22.396868	24.266667	-122.39922	37.61946	-122.42014	37.80612	1	2008-05-22 20:58:38	2008-05-22 21:22:54
001b0894-cd7c-4de2-8b9e-65b230ca05de	new_abdremlu	5.189942	15.133333	-122.42340	37.79338	-122.39837	37.77192	1	2008-05-22 23:21:56	2008-05-22 23:37:04
0021089d-39f0-47df-a99c-7b47b9442d7c	new_udwadla	1.116952	3.800000	-122.43212	37.79689	-122.42184	37.79624	0	2008-06-02 23:36:46	2008-06-02 23:40:34
00363a3a-2962-4b70-a7a6-f5011d954bbd	new_abgibo	1.986649	5.666667	-122.40230	37.76155	-122.39509	37.75129	0	2008-06-06 08:01:46	2008-06-06 08:07:26

Calculation of Geographical Distance

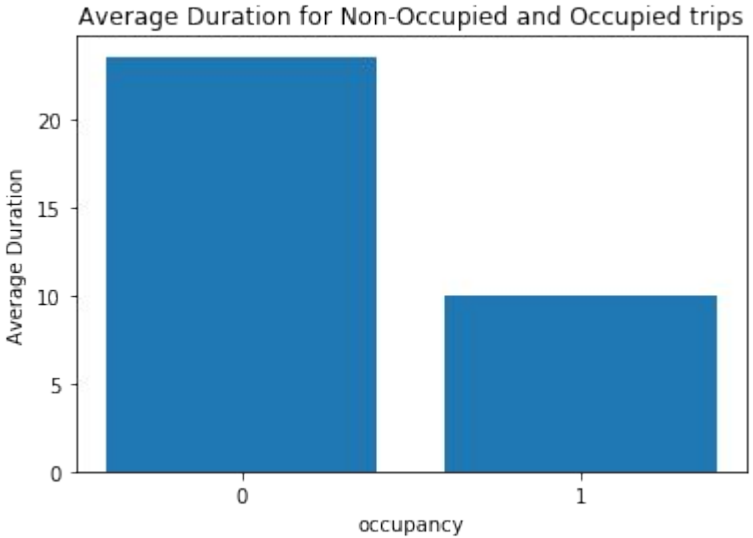
Equirectangular approximation

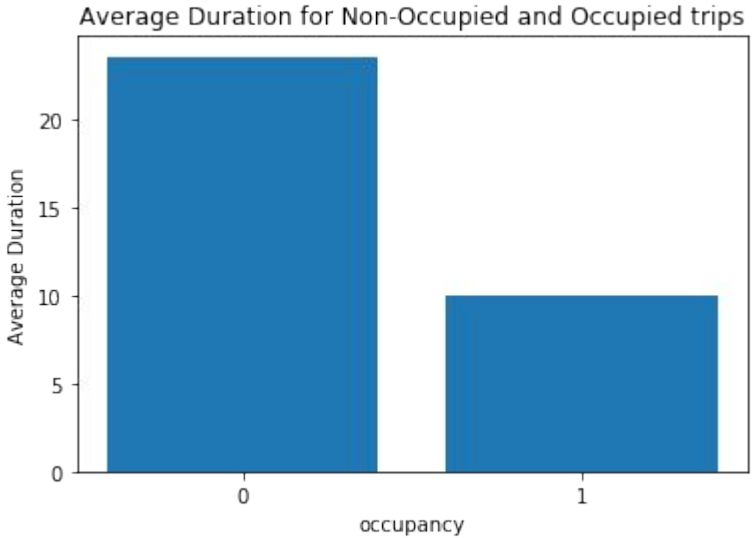
$$x = \Delta lon. \cos(lat)$$

$$y = \Delta lat$$

$$d = R. \sqrt{x^2 + y^2}$$

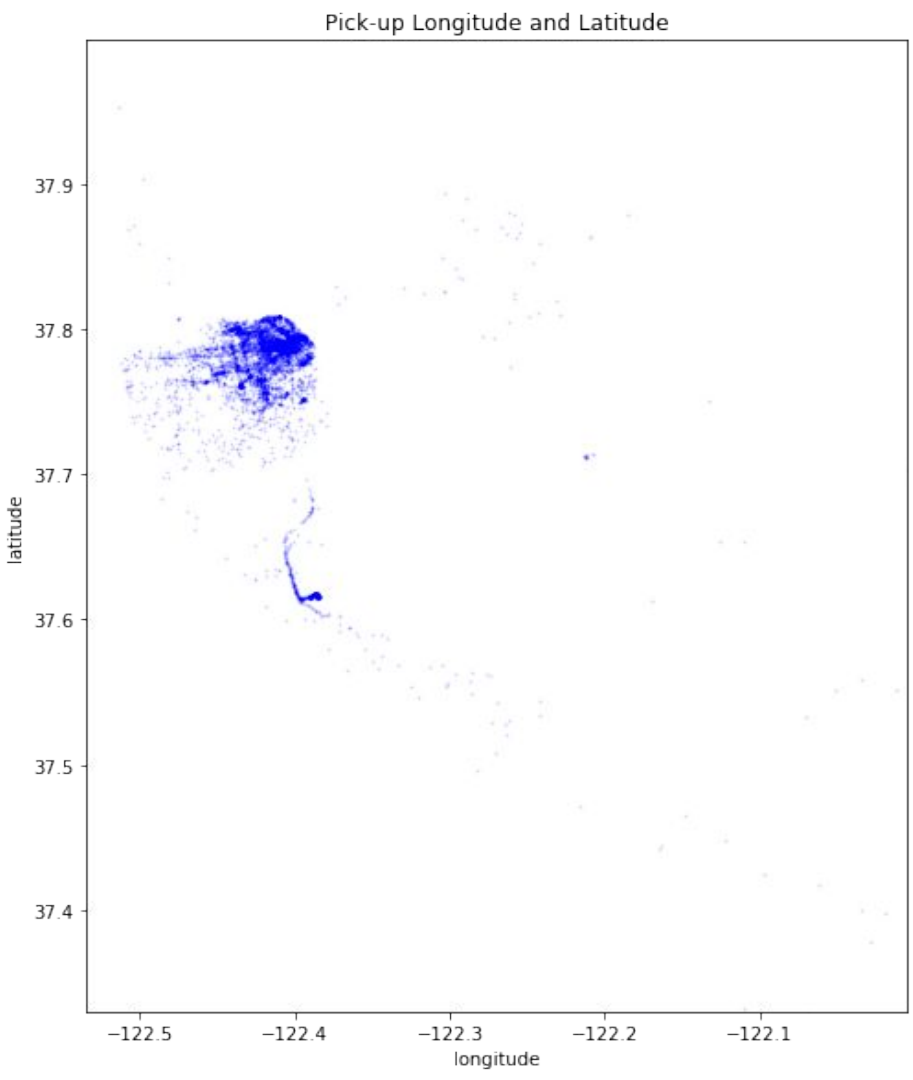
Statistics on duration and distance of free rides vs with passenger rides



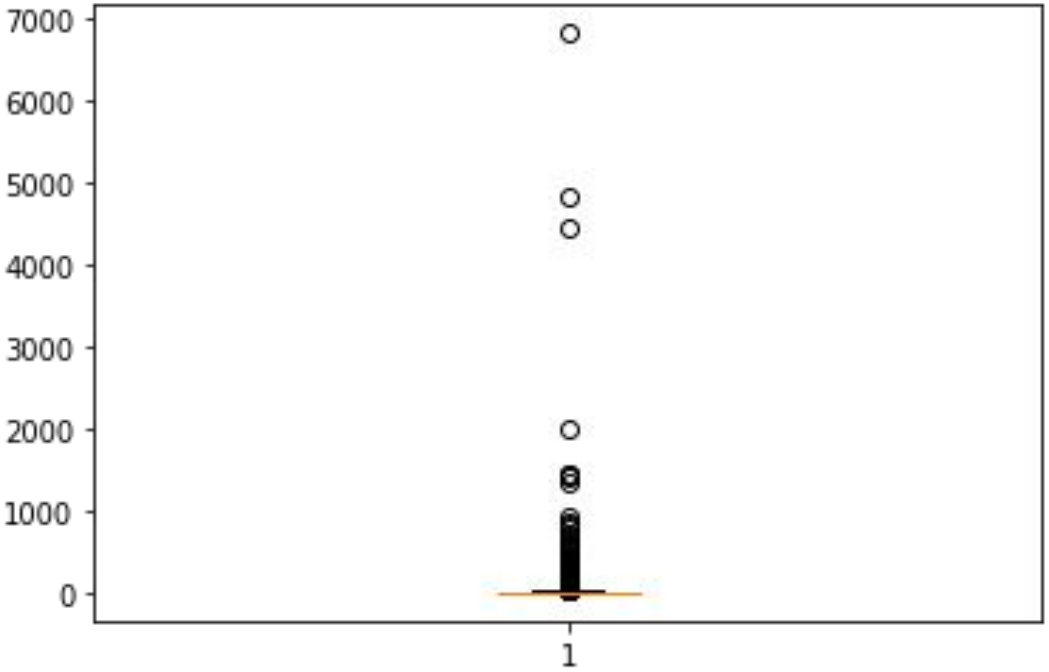


Missing values

No missing values or invalid observations were detected


Pick-up Longitude and Latitude
37.9
37.8
37.7
37.6
37.5
37.4
latitude
−122.5
−122.4
−122.3
−122.2
−122.1
longitude

The boxplot of the duration



Feature addition

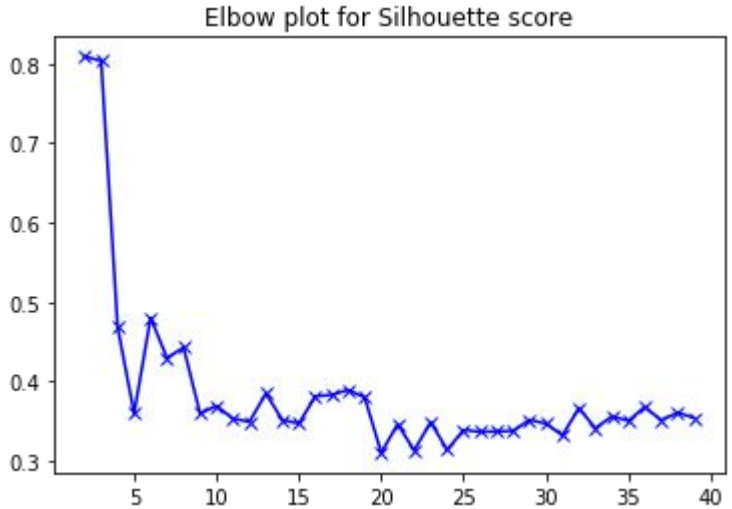
- Hour
- Day
- Weekday
- Speed

```
hour_bins = [-1, 5, 7, 10, 16, 21, 23]
bin_names = ['late_night', 'morning', 'morning_peak',
'afternoon', 'evening', 'night']
```

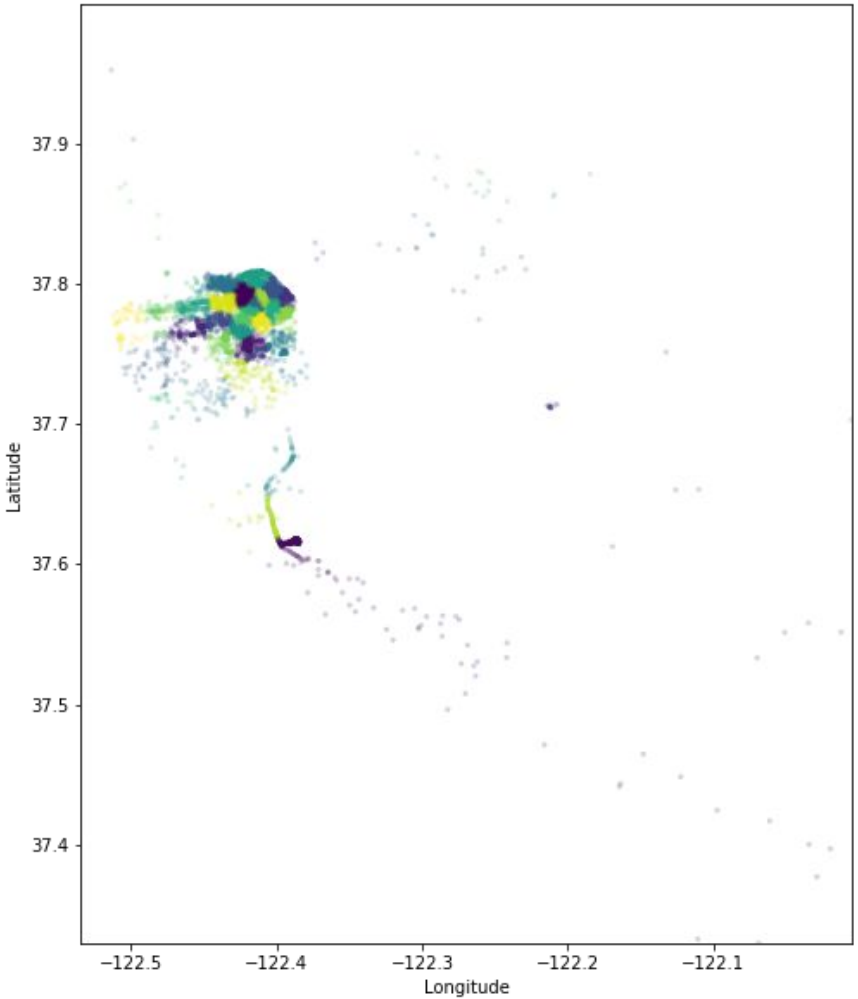
One more feature... region clusters

K means clustering for longitude and latitude

Silhouette score for different values of K



Silhouette score is remaining the same after k = 35



Map based on clusters

Opportunities of external data sources

- Weather data
- Airport trip flag
- etc.

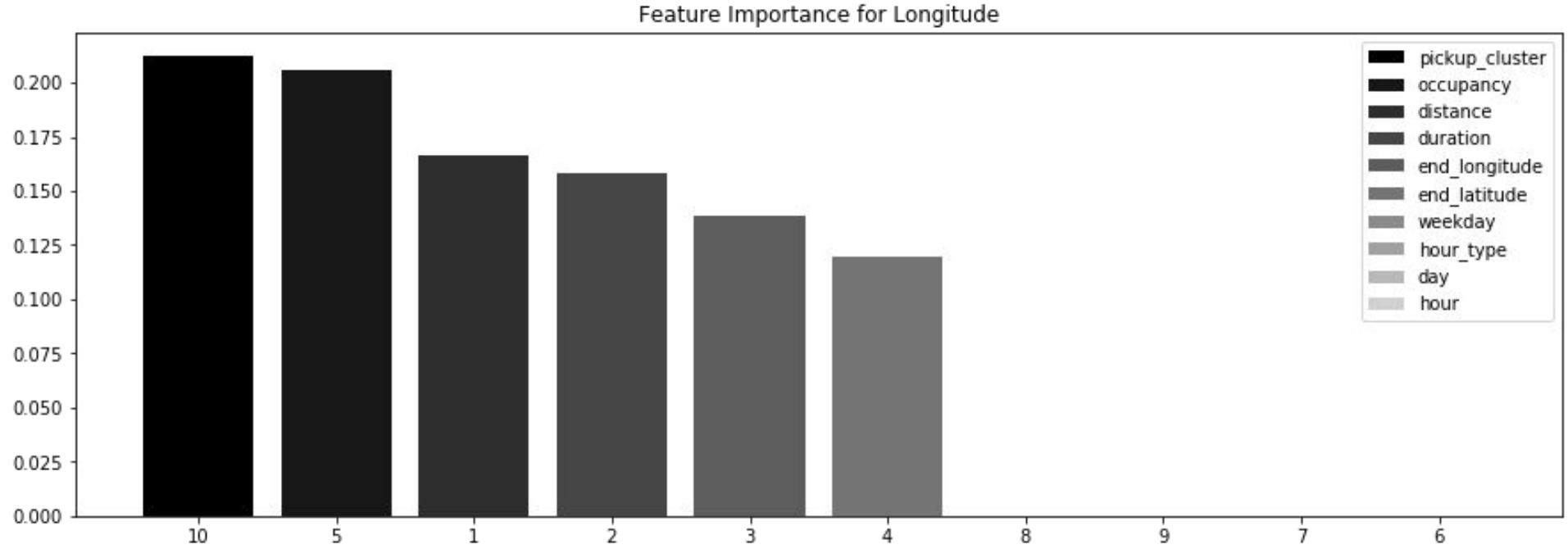
Modeling

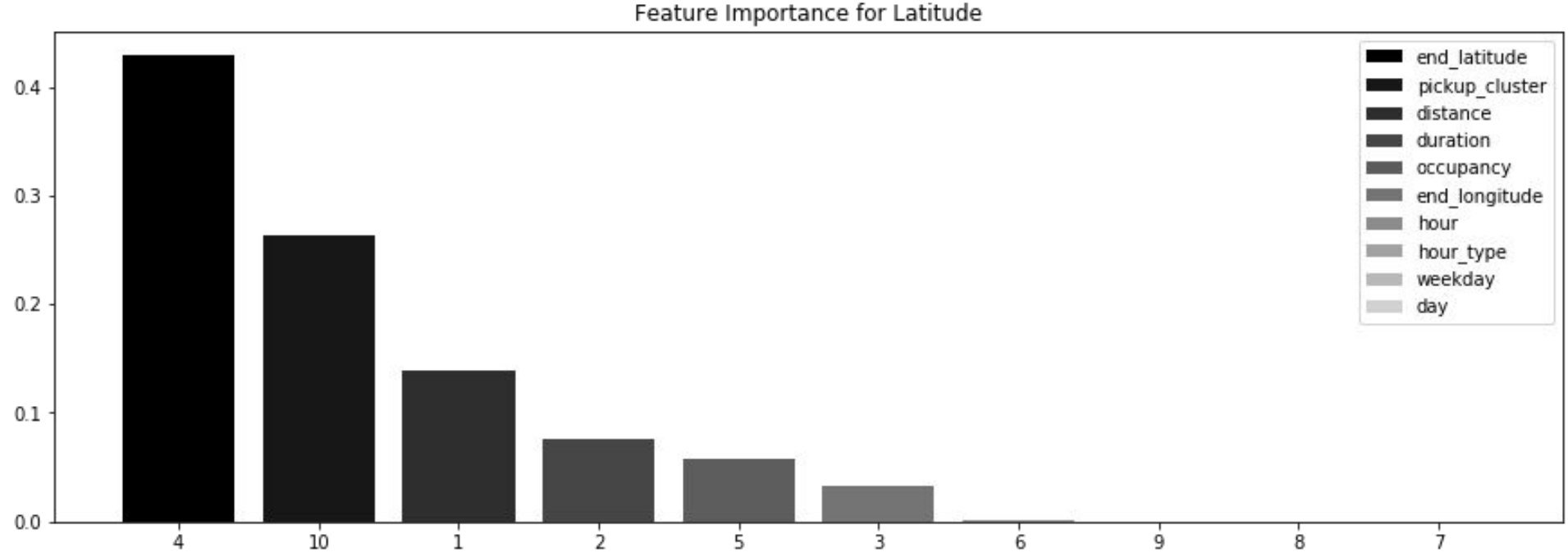
Multivariate XgBoost

Mean squared error = 0.000333

Multivariate Linear Regression

Mean squared error = 0.0012


Feature Importance for Longitude
pickup_cluster
occupancy
distance
duration
end_longitude
end_latitude
weekday
hour_type
day
hour
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
10
5
1
2
3
4
8
9
7
6


Feature Importance for Latitude
end_latitude
pickup_cluster
distance
duration
occupancy
end_longitude
hour
hour_type
weekday
day
0.4
0.3
0.2
0.1
0.0
4
10
1
2
5
3
6
9
8
7

Pitfalls of the use of the model

Independent variable

We are forcing model to learn to go to the pick up locations even when taxis were wrongmost of the time in their search

Showing concrete location

We can show the demand on the region instead of showing the concrete location so as taxis can decide to which region they want to ride.

Smoothing time series

We have not used time series data potential and characteristics

Solution

Predicting demand given the region and the time bins

Implementation

Time Series Prediction

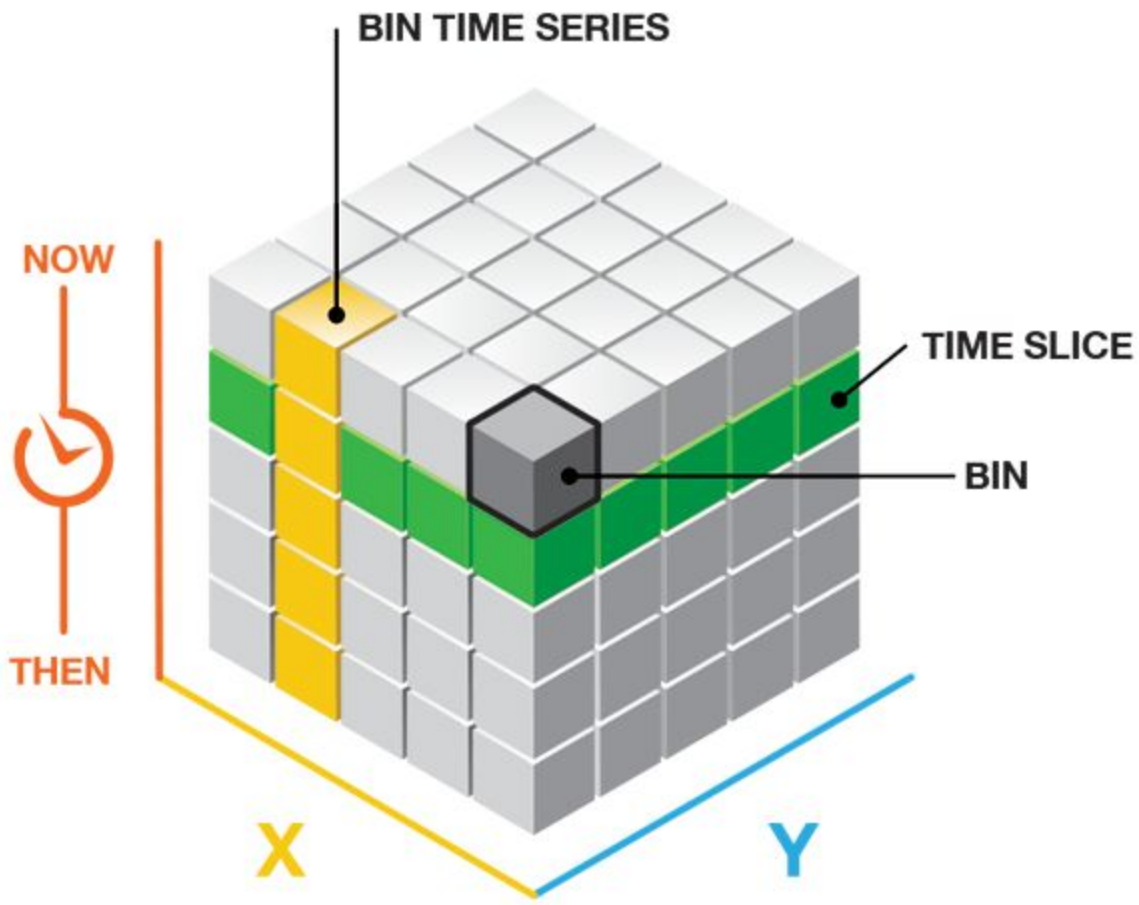
Given a region and a 10 min interval, we have to predict pickups.

(a): How to break up the San Francisco into regions?

(b): Every region of San Francisco has to be broken up into 10 min interval.

We have used K means clustering to break up San francisco into regions

We already know, about the pickup at time ‘t’, we will predict the pickup at time ‘t+1’ in the same region. Hence, this problem can be thought of as a Time Series Prediction problem. It is a special case of regression problems. In short, we will use the data at time ‘t’ to predict for time ‘t+1’.


BIN TIME SERIES
NOW
THEN
TIME SLICE
BIN
X
Y

Thanks!