Location Based Demand Prediction for the NYC Ride Sharing And Taxi Market

Xiaoying Ji, Andrew Jo, Marcel Gwerder, Feng Xiao

Agenda

* Background
* Goals
* Proposed Concept
* Datasets
* Data Processing
* Modeling
* Evaluation of Results
* Architecture
* User Interface
* Q&A

Background

* Ride Sharing and Taxi Market

World’s Taxi Market Value currently valued >$100 billion and is highly fragmented

Ride Sharing, leading as a major building block of the “sharing economy”

- Exponential growth in recent years:

1. Over 1.5 million drivers and 42 million passengers, and 2 billion annual rides.
2. Uber’s expected IPO of $120B in 2019 (more than Ford, GM, and Chrysler combined)

* Pain points
    - Unpredictable wait time and fare income
    - Nonoptimal supply and demand

Goals and Plan of Activity

* Innovations
* Provide an option for the companies to be proactive
* Risks
* Internal goals and distribution for the project:

Levels	Details	Note/Deadline
Level 0	Data Visualization/UI	11/11/2018
Level 1	Prediction	11/29/2018
Level 2	Utilize additional variables	Realistic Goal
Level 3	Additional predictions	Reach Goal

Proposed Concept

- Grid-based Model Training
  - Dividing the minimum square containing NYC into 30*30 grids
    - Mapping coordinates of the data to the corresponding region
    - Aggregating data in the same region
  - Train and store the model in Cloud
- Prediction
  - Input: timestamp, weather, location
  - Output: predicted number of customers in each region
- Heatmap
  - Heatmap color depending on demands prediction in the selected day
  - The user pick up a day in the future

Datasets

- Uber Pickups in NYC
  - Source: Kaggle
  - 4.5 million Uber pickups from 2014
  - Date/time, Latitude, Longitude(Pick up & drop off)
- Taxi & Limousine Trip Data in NYC
  - Source: NYC Taxi & Limousine Commission
  - 165 million data from 2014
- (Level 2) Weather Data
  - Source: Weather.com
  - Temperature(in hour), precipitation(rain, snow)...

Data Processing

- Used Amazon Athena to handle size of the datasets (~24GB, 165 mio rows)
- Combine rides with historical hourly weather data
- Remove erroneous data points (e.g. out of range lat/long)

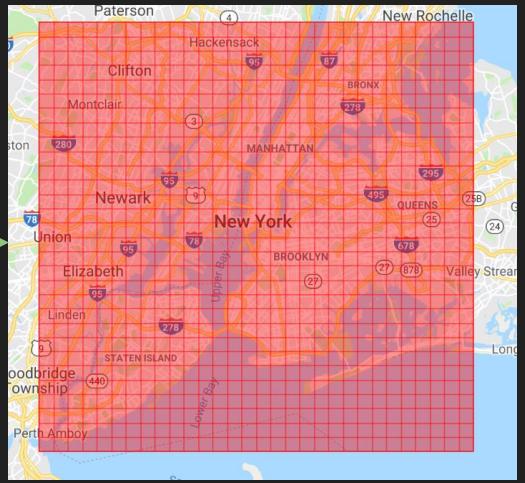
	pickup_datetime	pickup_latitude	pickup_longitude
1	2014-06-29 00:14:00.000	40.732315	-74.007685
2	2014-06-29 00:14:00.000	40.691362	-73.993780
3	2014-06-29 00:14:00.000	40.738440	-73.986000
4	2014-06-29 00:14:00.000	40.721337	-73.997502
5	2014-06-29 00:14:00.000	40.720995	-73.986895
6	2014-06-29 00:14:00.000	40.719027	-73.984927

	pickup_datetime	pickup_latitude	pickup_longitude	temp	precip_hrly
1	2014-08-11 13:40:08.000	40.737787	-73.988175	55.0	10.0
2	2014-08-11 13:40:08.000	40.756623	-73.989976	55.0	10.0
3	2014-08-11 13:40:08.000	40.742605	-73.986720	55.0	10.0
4	2014-08-11 13:40:10.000	40.757687	-73.997051	55.0	10.0
5	2014-08-11 13:40:10.000	40.774054	-73.872902	55.0	10.0
6	2014-08-11 13:40:10.000	40.760001	-73.981041	55.0	10.0

Data Processing

- Map the pickup location to the respective tile in the grid
- Aggregate data to get the demand for every tile and every hour of the year

	demand	day	day_of_week	hour	temp	precip	x	y
1	12	274	Wednesday	5	56.0	10.0	17	16
2	6	274	Wednesday	5	56.0	10.0	16	13
3	24	354	Saturday	18	19.0	10.0	18	19
4	1	354	Saturday	18	19.0	10.0	16	23
5	4	358	Wednesday	20	55.8	2.2	22	18
6	1	337	Wednesday	15	42.5	8.0	16	11



Modelling

- Next step in the project and currently work in progress
- Train model that predicts demand for specific tile at given time in the future
- Alternative is the classification into low/medium/high demand
- Derive reasonable predictors from limited input given e.g. day of week from date/time.

Promising options include (continuous response / non-linear relationship)

➔ Regression Trees / Random Forest
➔ Deep Neural Network

Keras

Modelling

- A first test using a random forest with regression trees gives an R^2 of **~0.985**
- Adding predictors one by one shows that the **x,y** combination has by far the biggest impact followed by the **day_of_week** dummy variables and **hour**

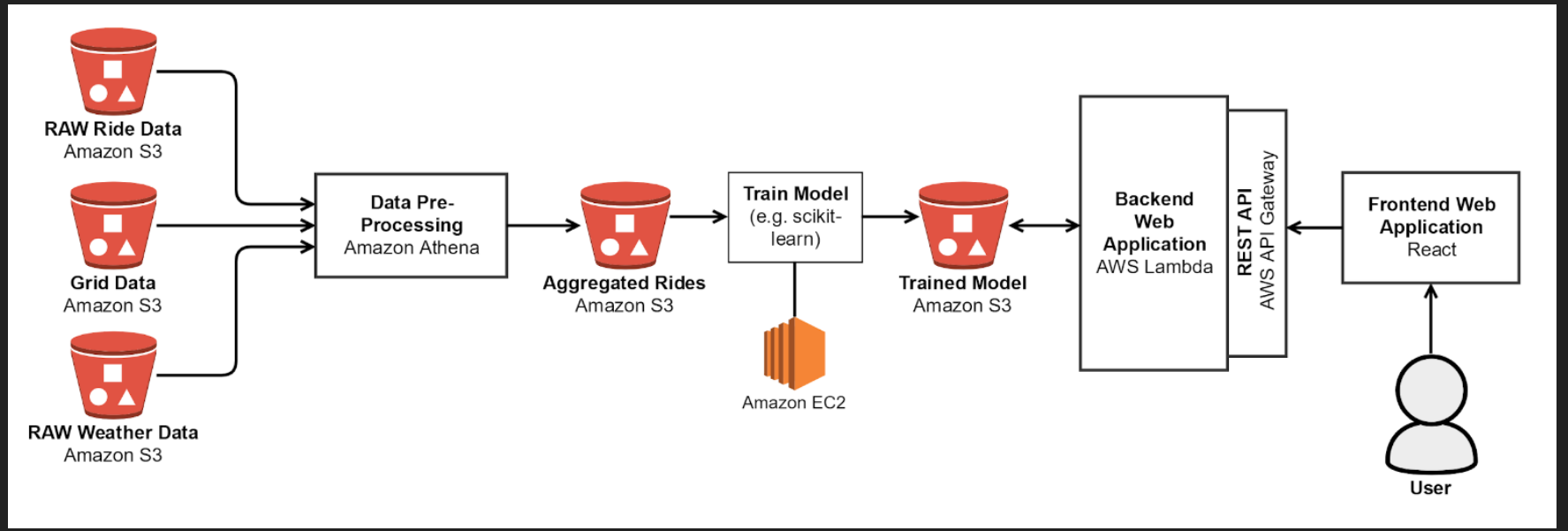
	demand	day	hour	temp	precip	x	y	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
0	6650	306	1	34.5	10.0	15	17	0	0	0	1	0	0	0
1	6395	62	19	0.0	10.0	16	19	0	1	0	0	0	0	0
2	6321	84	19	14.0	10.0	16	19	0	0	0	0	0	1	0
3	6290	105	19	41.5	7.5	16	19	0	0	0	0	0	1	0
4	6267	83	18	2.0	10.0	16	19	0	1	0	0	0	0	0

```
(project-RKpx-M55) dev@laptop7:/mnt/c/Development/CS6220/project$ python model.py
R^2: 0.9859169914729624
Mean Absolute Error: 15.83829224188525
Mean Squared Error: 3369.475871446246
Root Mean Squared Error: 58.04718659372085
```

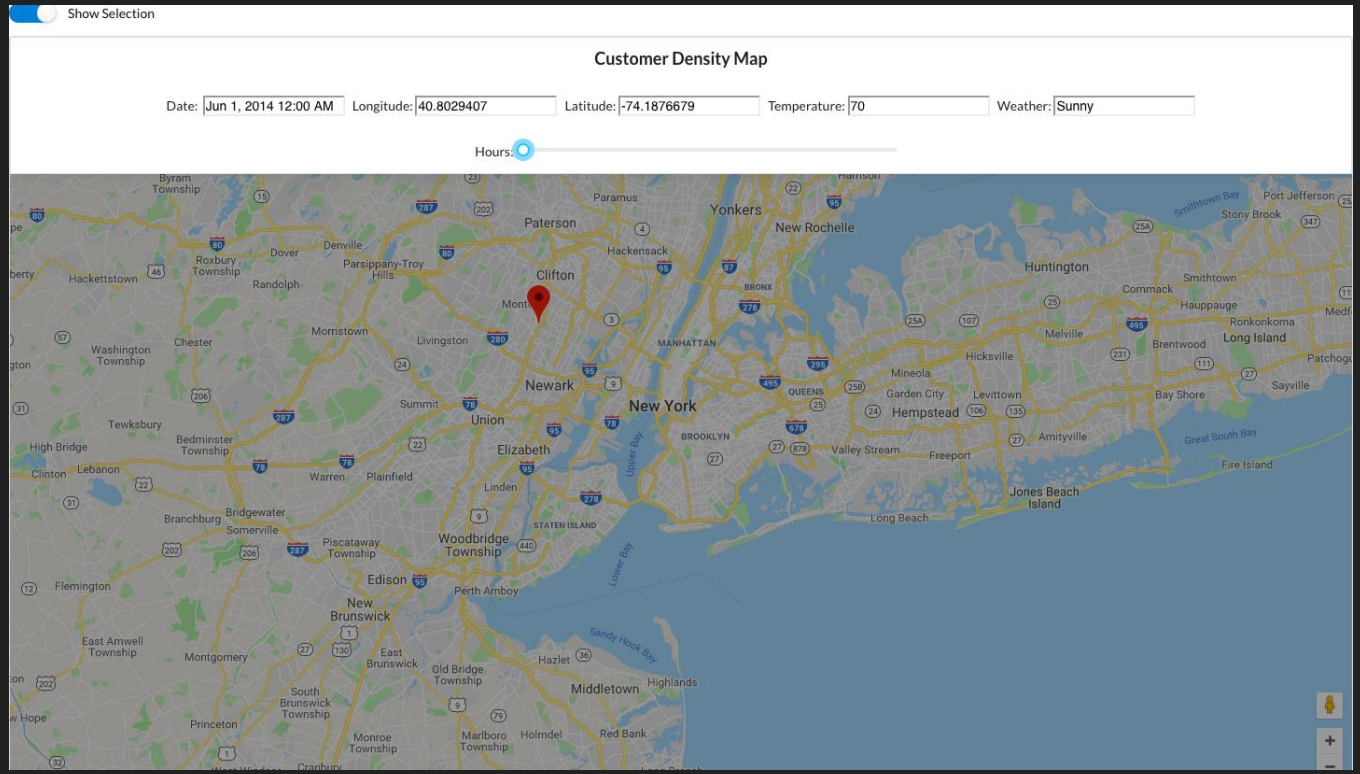
Evaluation of Results

- Aggregated data still consists of ~1 mio data points which provides flexibility when spitting into training, validation and test datasets
- High R^2 indicates a high prediction accuracy but there could also be a problem in the model (needs further investigation)
- Required prediction accuracy is subject to end-user perception

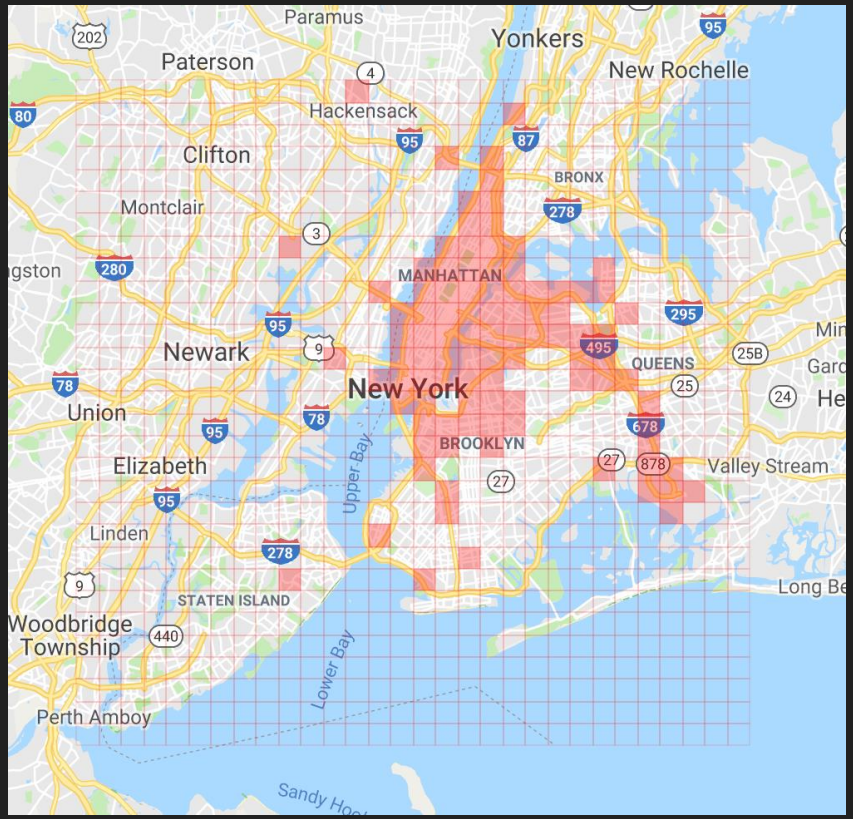
Architecture

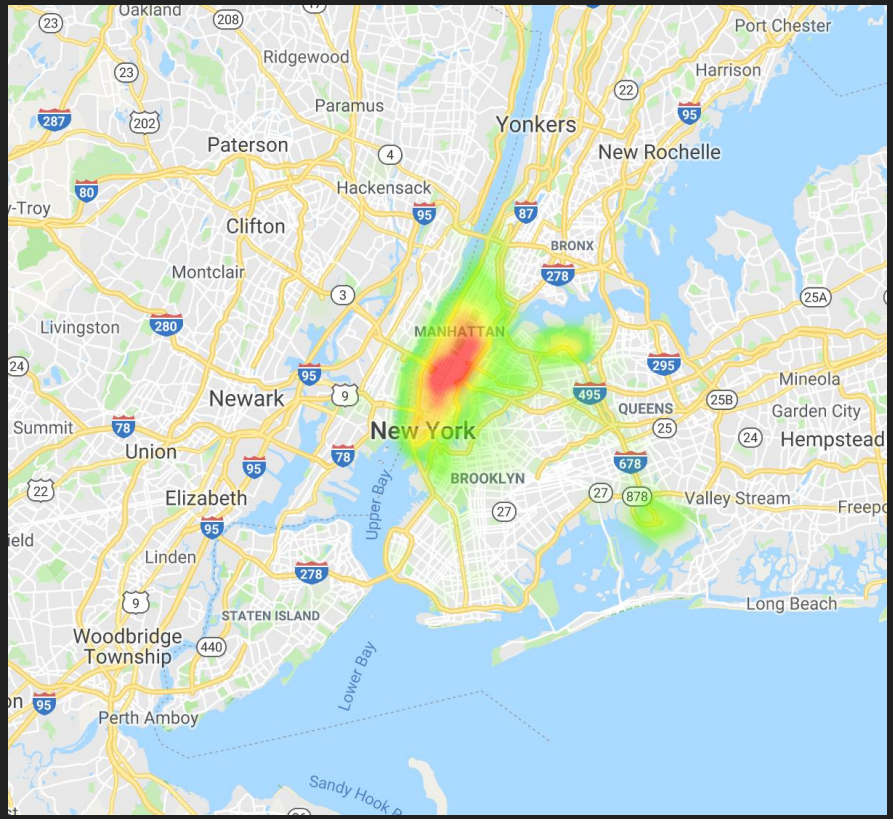


UI

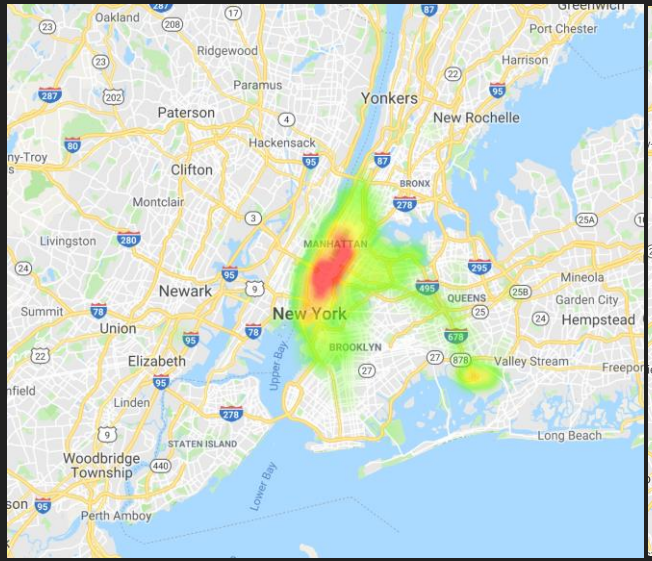

Show Selection
Customer Density Map
Date: Jun 1, 2014 12:00 AM
Longitude: 40.8029407
Latitude: -74.1876679
Temperature: 70
Weather: Sunny
Hours:
Paterson
Yonkers
New Rochelle
Clifton
Hackensack
Paramus
Huntington
Smithtown
Commack
Hauppauge
Long Island
Ronkonkoma
Brentwood
Melville
Hicksville
Mineola
Garden City
Levittown
Hempstead
Bay Shore
Sayville
Amityville
Freeport
Valley Stream
Jones Beach Island
Long Beach
Fire Island
Great South Bay
Newark
New York
Union
Elizabeth
Linden
Woodbridge Township
Edison
Perth Amboy
New Brunswick
Middletown
Highlands
Hazlet
Sandy Hook Bay
Lower Bay
Upper Bay
Morristown
Livingston
Summit
Warren
Plainfield
Piscataway Township
Somerville
Bridgewater
Branchburg
Flemington
East Amwell Township
Montgomery
East Brunswick
Old Bridge Township
South Brunswick Township
Princeton
Monroe Township
Marlboro Township
Holmdel
Red Bank
Washington Township
Chester
Roxbury Township
Dover
Denville
Parsippany-Troy Hills
Randolph
Hackettstown
Byram Township
Tewksbury
Bedminster Township
High Bridge
Clinton
Lebanon
MANHATTAN
BRONX
QUEENS
BROOKLYN
STATEN ISLAND

Heatmap of NYC taxi pickup At 12:00pm-1:00pm

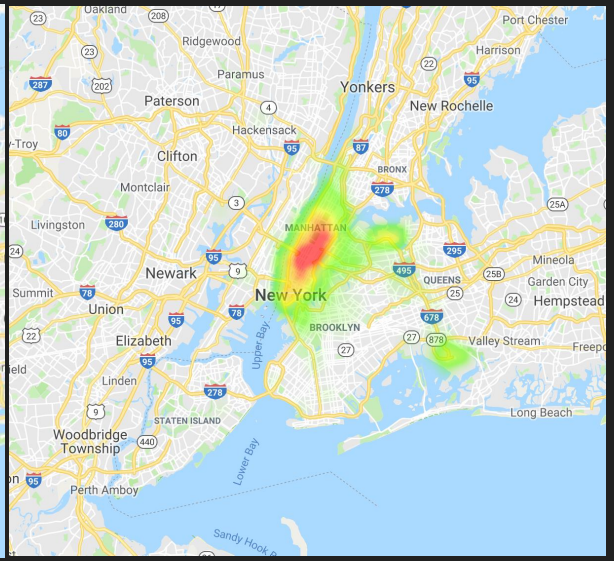




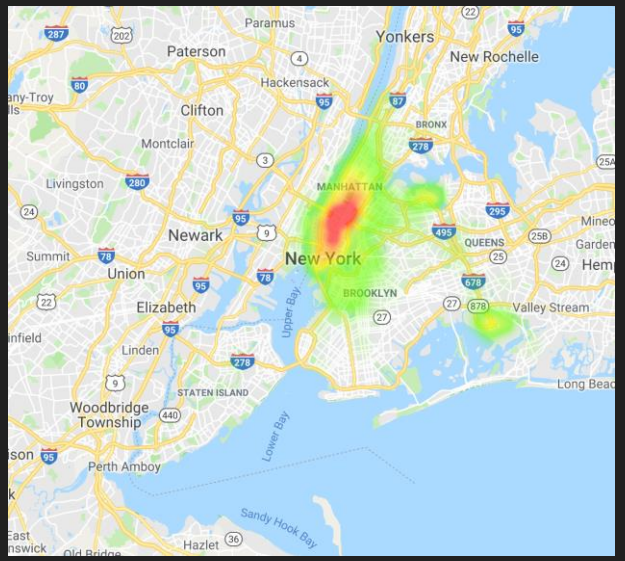
Comparison of Taxi Pickups at Different Time



6am-7am



10pm-11pm



12pm-1pm

Q&A