## horizontal line



**NYC Taxi Trip Analysis**

04.17.2017

Video URL: <https://goo.gl/4r57Jc>

Taj Poovaiah

Mohit Mittal

Sneha Ravikumar

Final Project : INFO 7390 Advances In Data Science

# Overview

New York City, being the most populous city in the United States, has a vast and complex transportation system, including one of the largest subway systems in the world and a large fleet of more than 13,000 yellow and green taxis, that have become iconic subjects in photographs and movies.  
The subway system digests the lion share of NYC's public transport use, but the 54% of NYC's residents that don't own a car and therefore rely on public transportation still take almost 200 million taxi trips per year!

# Goals

To provide some helpful information - things that we wish we could know - to the general public.

How does the demand for taxi change during the day? Where are the taxis if you often have trouble getting one? And how much do people usually tip or am I over or under tipping? Data, Overview, Location, Time, Fare. Also, check if weather influences any of these and if it does, by how much?

1. To offer a simple web application to let people select pickup and dropoff locations, date and time they plan to take the trip, and then return a machine learning predicted cab fare and length of trip in minutes
2. Predict the density of taxi demand throughout New York City as it changes from day to day and hour to hour. So, given a specific location, date and time, and weather conditions can we predict the demand and number of pickups in that location to a reasonably high accuracy?

# 

# Use Cases

**Making transportation more efficient**  
Predictive models like these are interesting for many people  
**Daily User:** Know up-front pricing & plan ahead with behaviour & trend analysis  
**Taxi companies:** Companies can maximize their utilization by diverting the cabs into the locations during specific times

# Data

1. **NYC Taxi Data**: <http://www.nyc.gov/html/tlc/html/passenger/taxicab_rate.shtml>

We will work with the Yellow Cab Data for the year 2016 (Jan- June)

**Data Dictionary:** http://www.nyc.gov/html/tlc/downloads/pdf/data\_dictionary\_trip\_records\_yellow.pdf

1. **Weather Data:**

Collected from three stations:

* Central Park
* LaGuardia
* JFK

The ‘weatherData’ package in R provides us with the required information.

# Process Outline

1. Data Preprocessing

* Data Cleaning, handling missing values
* Join the taxi data with the weather data

2. Exploratory Data Analysis

3. Study of Supervised approaches and select the best model for prediction

4. Study of Unsupervised approaches (Clustering and Associative rule mining) for recommendation

5. Design of a pipeline and system to implement this approach and discussion on the system’s capabilities

6. Deploy the Model on Azure/AWS or Google Cloud Computing Platform

7. Build a web application to demonstrate the prediction and recommendation results.

# 

# 

# Milestones

|  |  |
| --- | --- |
| **Timeframe** | **Delivery** |
| Day 1-2 | Data Preprocessing and Exploratory Data Analysis |
| Day 3-6 | Model Building, Training, Selection |
| Day 7-8 | Deployment of models on cloud and build web application |
| Day 9-10 | System integration and documentation |

## 

## **Personas:**

1. Yellow Taxi Drivers:

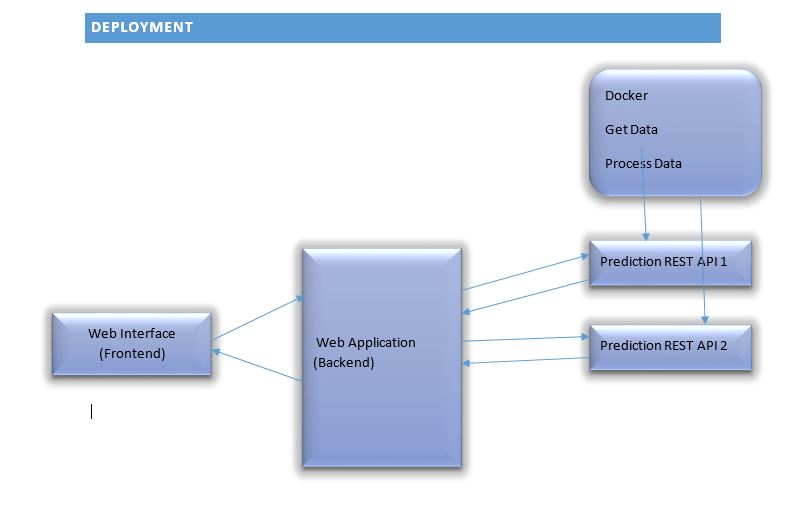
* To check for Recommendations of best areas to go to for more customers

1. Yellow Taxi Cab Users:

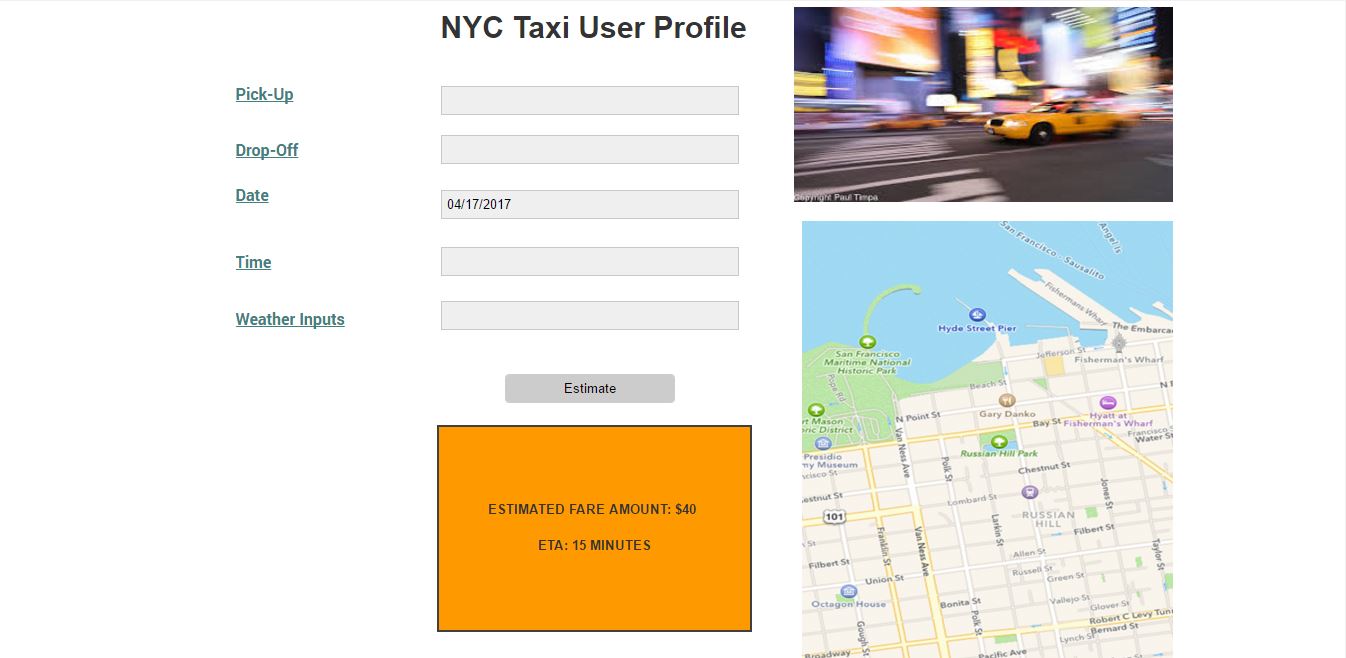
* To check the Estimated Fare Amount
* To check for the ETA (Estimated Time of Arrival)

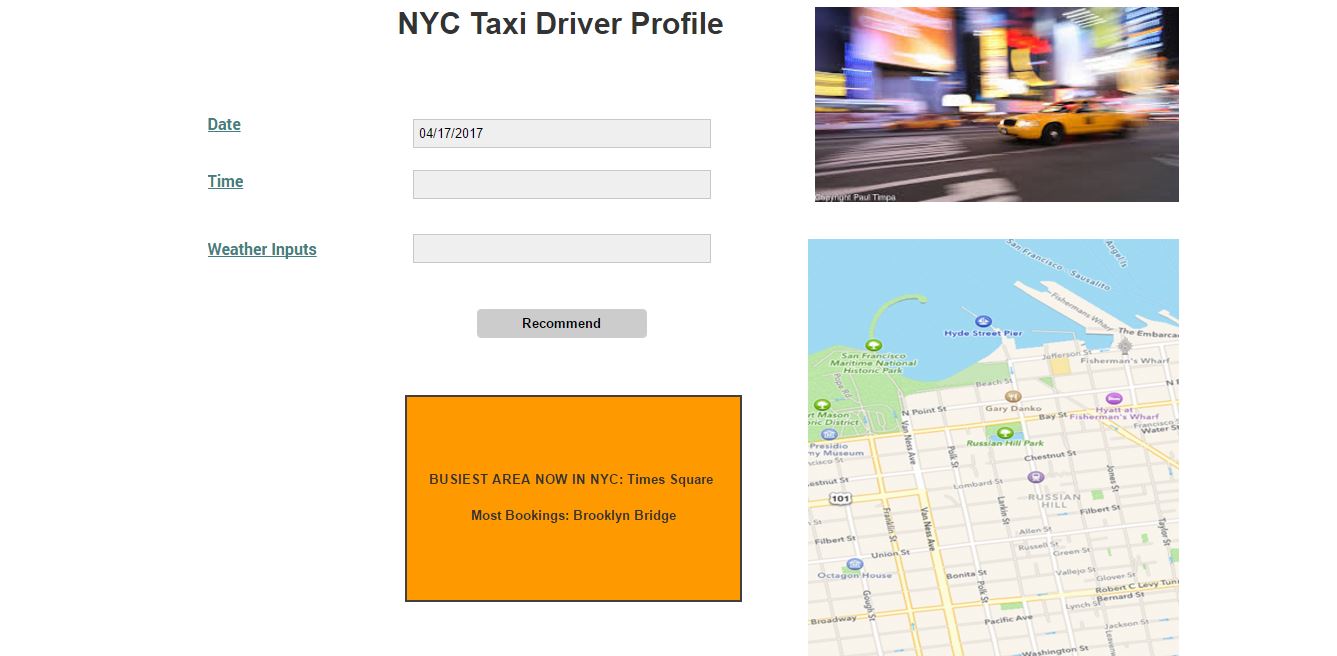
## **Deployment Details:**

1. Language: Python, Java
2. Pipeline: Luigi
3. Container: Docker
4. Cloud Tools/Platforms: Microsoft Azure Machine Learning Studio,AWS (Amazon WEb Services) EC2
5. Tools for Analysis: Microsoft Azure Visual Studio, ArchGIS
6. Other Considerations: Google Cloud Platform



## **USER INTERFACE DESIGN PLAN**





## **Reference and Sources:**

<http://nyc-taxi-trips.herokuapp.com/index>

https://www.ocf.berkeley.edu/~dlevitt/2015/12/13/final-project-nyc-taxi-and-uber-data/