Wen Sheng
Taylor Arnold
January 31, 2016

Problem Set 01 Data Analysis

Data Mining and Machine Learning

1. Factors that affect whether a hailed cab will leave Manhattan

According to visualization in the following sub part 3, both **Hour** and **Neighborhood** code affect whether a hailed cab will leave Manhattan, while Longitude/Latitude is not directly relevant with leaving probability. More importantly, these 2 factors seems to be **independent** with each other:

Within the same hour period, the probability of whether will leave Manhattan is according to neighbor hood: the peak around neighborhood around code 20 - 40, the bottom around neighborhood around code; Within the same Neighborhood, the probability of leaving is depended on the hour factor and shows high peak hour in 4 - 5 am in the morning and bottom in 6 pm and around 10 am.

2. Other models: Additive model

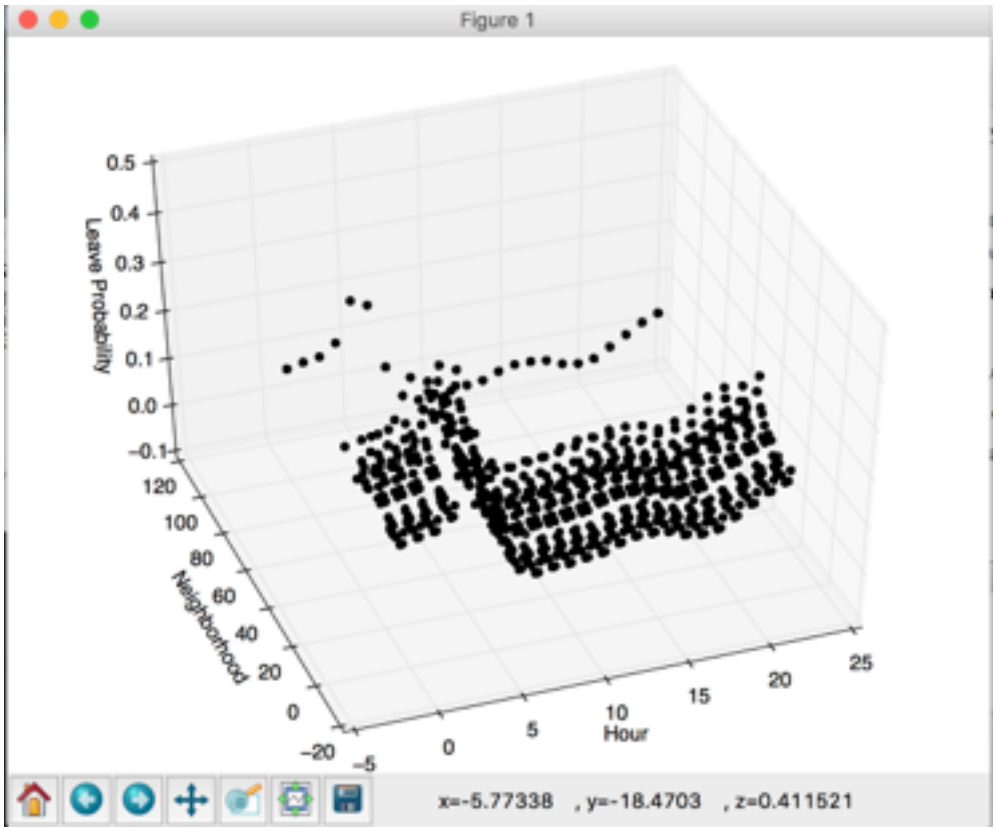
Additive model swill help to better illustrate the relationship between Hour, Neighborhood with leaving Manhunt's probability as they show an independent tendency. As the definition of additive model, one factor won't interfere with others.

$y = g1(hour) + g2(Neighborhood)$

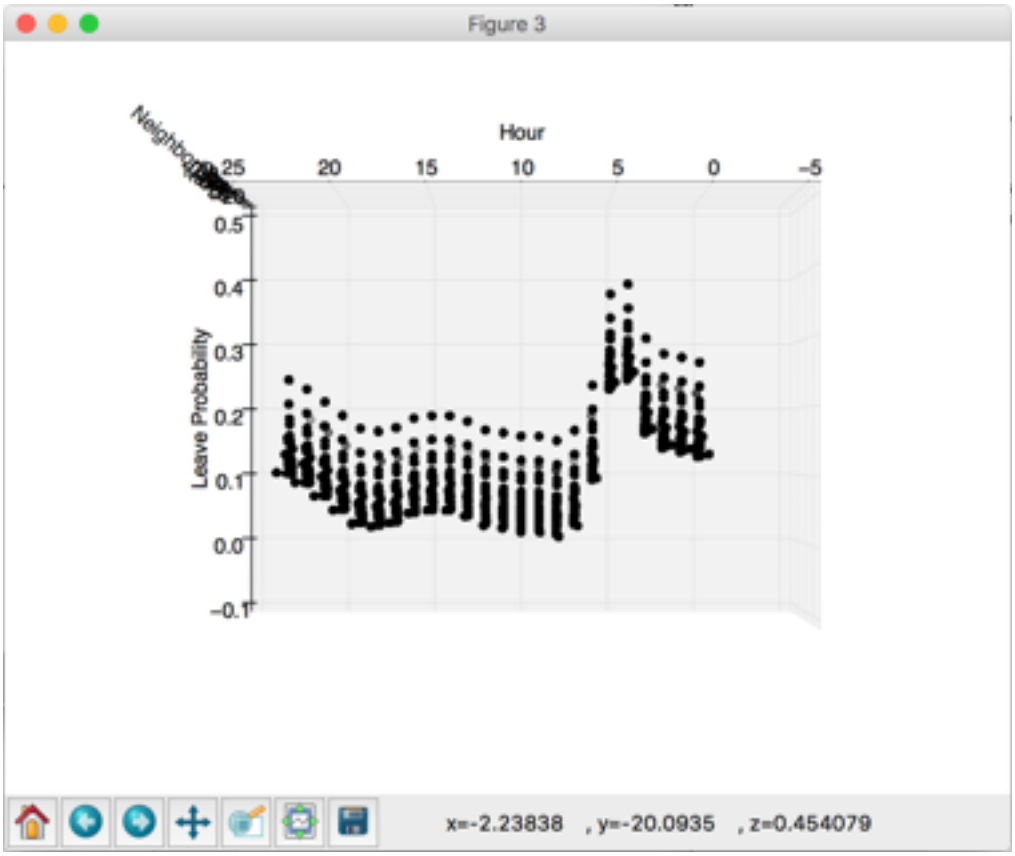
3. Linear Regression Model Visualization

I use 3D presentation to show the relation ship between X = [Hour, Neighborhood] with Y = leave Manhattan(0.0 - 1.0)

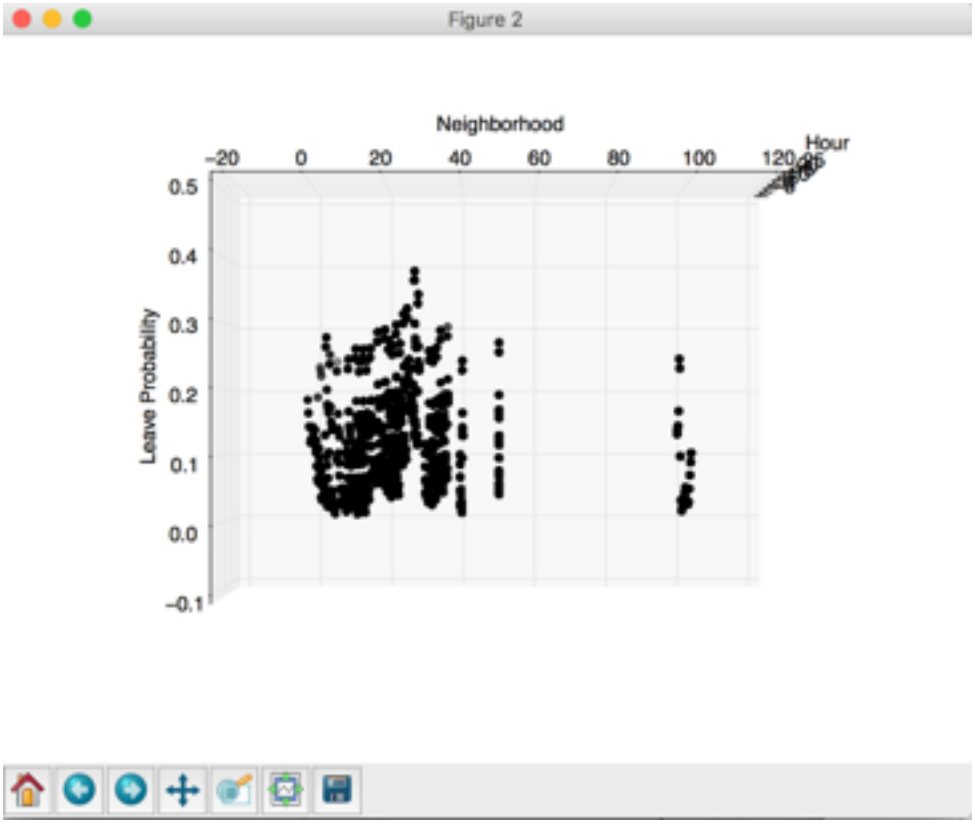
- **Overall view of Linear Regression Model**



• **2D view to show relationship between Hour and Y**



• **2D view to show relationship between Neighborhood and Y**



4. Summery from Visualization

A. For model1:

From Visualization, factors of **Longitude/Latitude** is not directly relevant with the probability of whether leaving Manhattan. But it still shows the peak value in specific longitude, latitude(around Manhattan).

B. For model 2 and 3:

Fig 1

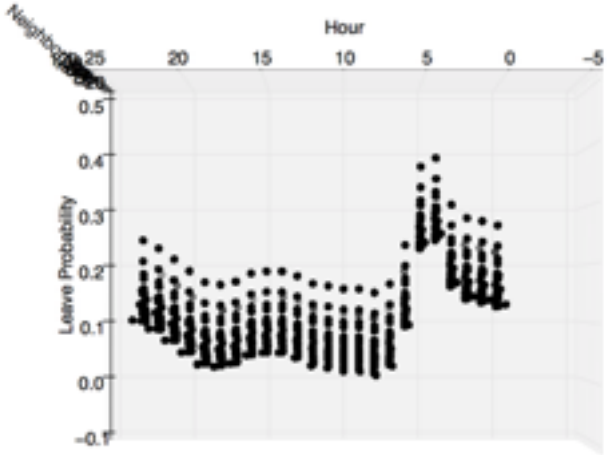
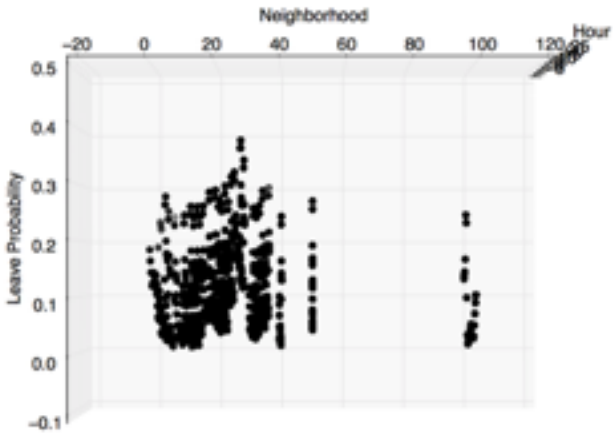


Fig2



Hour and **Neighborhood** both and **independently** affect the probability of leaving Manhattan.

Using **Univariate Analysis** approach:

a) Within the same Neighborhood (fig1), the leaving probability shows a similar tendency as hours is X. **4pm - 5pm is the peak hour.** I guess it's the time people feel not safe or (some subway lines is not operated) to take train/subways, so they have higher chance to choose taxi.

b) Within the same Hour, the Neighborhood shows a similar tendency as Neighborhood code is X. Code 35 - 36 (MN36 Manhattan, BK35 Brooklyn) . These are the location near Manhattan, so leaving probability is higher than farther neighborhood as taxi fees is not that high and people can accept it. Also if people leave in Manhattan, the reason they take cab should be it's not within walking distance so that leaving Manhattan is high. The result is conformed to real world analysis.