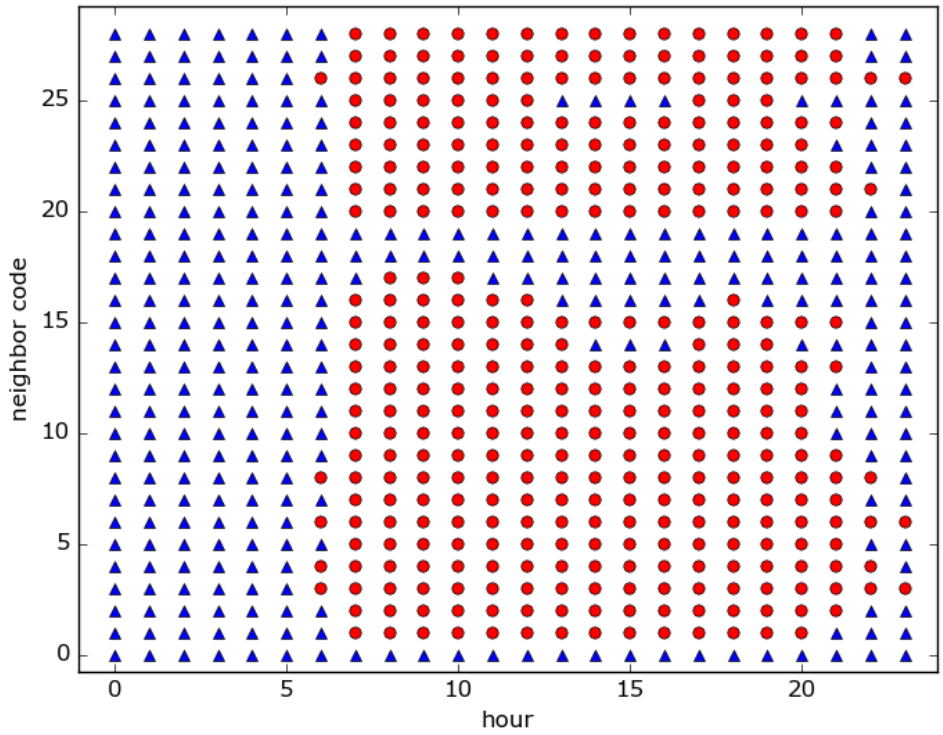
STAT665 Lab1

Bo Song

This algorithm uses category encoding for both variable 'hour' and variable 'neighborhood'. The mapping pattern of neighborhood code is demonstrated in following table.

NTACode	Name	Index
MN01	Marble	0
MN03	Central Harlem North-Polo Grounds	1
MN04	Hamilton Heights	2
MN06	Manhattanville	3
MN09	Morningside Heights	4
MN11	Central Harlem Sourth	5
MN12	Upper West Side	6
MN13	Hudson Yards-Chelsea-Flatiron-Union	7
MN14	Lincoln Square	8
MN15	Clinton	9
MN17	Midtown-Midtown South	10
MN19	Turtle Bay-East Midtown	11
MN20	Murray Hill-Kips Bay	12
MN21	Gramercy	13
MN22	East Village	14
MN23	West Village	15
MN24	SoHo-TriBeCa-Civic Center-Little Italy	16
MN25	Battery Park City-Lower Manhattan	17
MN27	Chinatown	18
MN28	Lower East Side	19
MN31	Lenox Hill-Roosevelt Island	20
MN32	Yorkville	21
MN33	East Harlem South	22
MN34	East Harlem North	23
MN35	Washington Heights North	24
MN36	Washington Heights South	25
MN40	Upper East Side-Carnegie Hill	26
MN50	Stuyvesant Town-Cooper Village	27
MN99	Park-cemetery-etc-Manhattan	28

Based on the linear regression model we built in question 2, we get the following prediction graph.



The red dot indicates the 1, the cab remains in Manhattan. The blue triangle indicates 0, the cab will leave Manhattan.

According to the prediction model I built, cab whose picking up time is between 7:00~20:00 and picking up location index is between 1~15 and 20~28 is more likely to drop off passengers within Manhattan.

It makes sense because in the midnight and dawn, passengers in Manhattan are more likely to go home out of Manhattan.