Paper 11716 - 2016

# Arrest Prediction and Analysis based on Data Mining Approach

Karan Rudra, Oklahoma State University; Maitreya Kadiyala, Oklahoma State University;
Dr. Goutam Chakraborty, Oklahoma State University

## ABSTRACT

In lieu of past indiscretions by the police departments leading to riots in major U.S cities, it is important to assess factors leading to arrest of a citizen. The police department should understand the arrest situation before they can take any decision and diffuse any possibility of a riot or similar incidents. Many such incidents in the past are a result of police department failing to make right decisions in the case of emergencies.

The primary objective of this study is to understand the key factors that impact arrest of a person in New York City and understanding various crimes that have taken place in the region in the last two years. The study explores different regions of New York City where the crimes have taken place and also the timing of incidents leading to them. The study explores different regions of New York City like Manhattan, Brooklyn, Queens, Bronx and Staten Island where the crimes have taken place and also the timing of incidents leading to it. This study explores the reasons for arrest, reason for suspicions, timed arrest, identifying threats, basis of search and if arrest required a physical intervention. The study analyzes the reasons for arrest which are suspicion, timing of frisk activities, identifying threats, basis of search, if arrest required a physical intervention, location of the arrest and if frisk officer needs any support. Decision tree, Logistic Regression, Neural Network and Polynomial Regression models were built, of which Polynomial Regression model turned out to be the best model with validation misclassification rate of 6.54% and model sensitivity of 49.84%. The results from the decision tree model showed that suspicion count, search basis count, summons leading to violence, ethnicity, and use of force are some of the important factors influencing arrest of a person.

## INTRODUCTION

Recent riots in Ferguson, Baltimore and Dallas has led to many questions with regard to the decisions made by the police officers. it was important that we assess factors leading to arrest of a citizen, know their indiscretions better, understand the situation in hand and come with a viable solution to discern the aforementioned activities which might disrupt the proceedings. These riots are classic examples of police department not having required help in the case of emergencies to help them make correct decisions. These decisions taken by the police in a split second can make or break a person's life and might led to unprecedented wave of anger if their decisions are incorrect. As the protectors of law and in light of greater good, police sometimes make calculated decisions which might cause backlash. In the last couple of years due to the use of social media, the news is spreading like wildfire causing unrest among the citizens and also the police department.

## DATA DESCRIPTION

The data used in this project was collected from NYPD stop and frisk report database. The data used for exploration and analysis are for the years 2013 and 2014. It consists of 234,527 observations having 56 variables; it is a multivariate dataset consisting of variables related to citizens being stopped, questioned and frisked by the police officers. Variables such as suspicion count, search basis count, force use, pct. (precinct location 1,5,6,7…. 123) as well as others are used in the models. The target variable Arrest made is a binary variable indicating if a citizen was arrested or not when stopped and frisked by the police official. It has 9.6% of 1's which indicates arrest was made and 90.4% of 0's which indicates no arrests were made.

A brief description of the important variables is given below:

**Suspicion Count** – The police have categorized various suspicious activities as nominal categorical variables. (Example – Suspicion of doing unlawful activities, suspicion of harassing someone, suspicion of drinking in unlawful areas)

**Suspicion** – This variable explains whether a person has been frisked or not on suspicion. 0 for someone who was not frisked 1 for person who was frisked.

**Search basis count**- The count of reasons or basis for which police have observed for frisking.

**Search basis** – Whether the person has been frisked or not based search basis category.

**Force Use** - Was physical force used when frisked

**Offrcid** – Whether the officer showed his ID when frisked

**Pct** – Precinct location of nearest police station

**ofcshld** – Was someone assisting or shielding the officer

**City** – The different districts in New York (Brooklyn, Manhattan, Bronx, Queens, Staten Island)

**Offunv** – Was the officer in Uniform

**Sumissue**- Was summon issued against the person

Categorizing individual attributes into groups to simplify analysis was another major task. Various exploratory techniques such as one way frequencies and cross tab analysis was used on the data to explore the importance and variance explained by various categorical variables.

## EXPLORATORY ANALYSIS

The study shows that Brooklyn and Manhattan are the areas in New York where there is utmost criminal activity but highest number of arrests in a single location has been made in Bronx near 110 Church intersections. Males are the most arrested among different sexes.

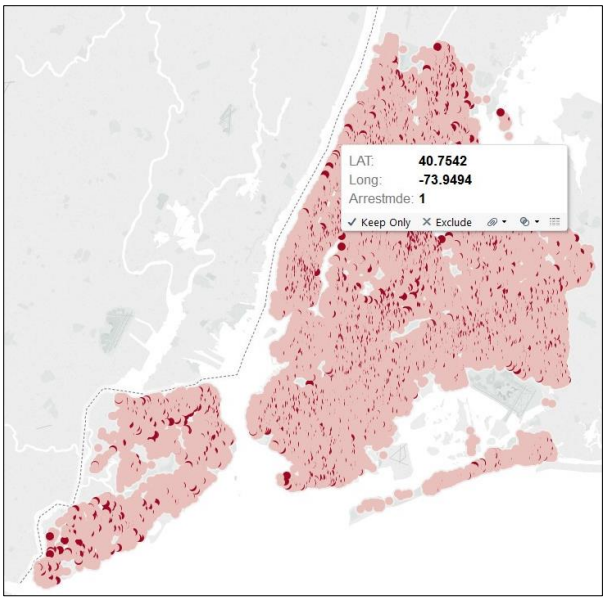


**Figure 1. Map Chart Representation of Arrest Made in NY City**

Figure 1 represents a Map Chart of number of arrests made in New York City over a period of 2 years starting from January, 2013 to December, 2014. The dark red spots on the map represent concentration of areas where arrests have been made. It can be observed that the number of arrests is more in southern part of Staten Island.

Arrests made and persons frisked are highest in the month of January and decrease gradually over the years. Most crimes are reported in Brooklyn and Manhattan. Stop to arrests percentage is the highest in Bronx area of New York.

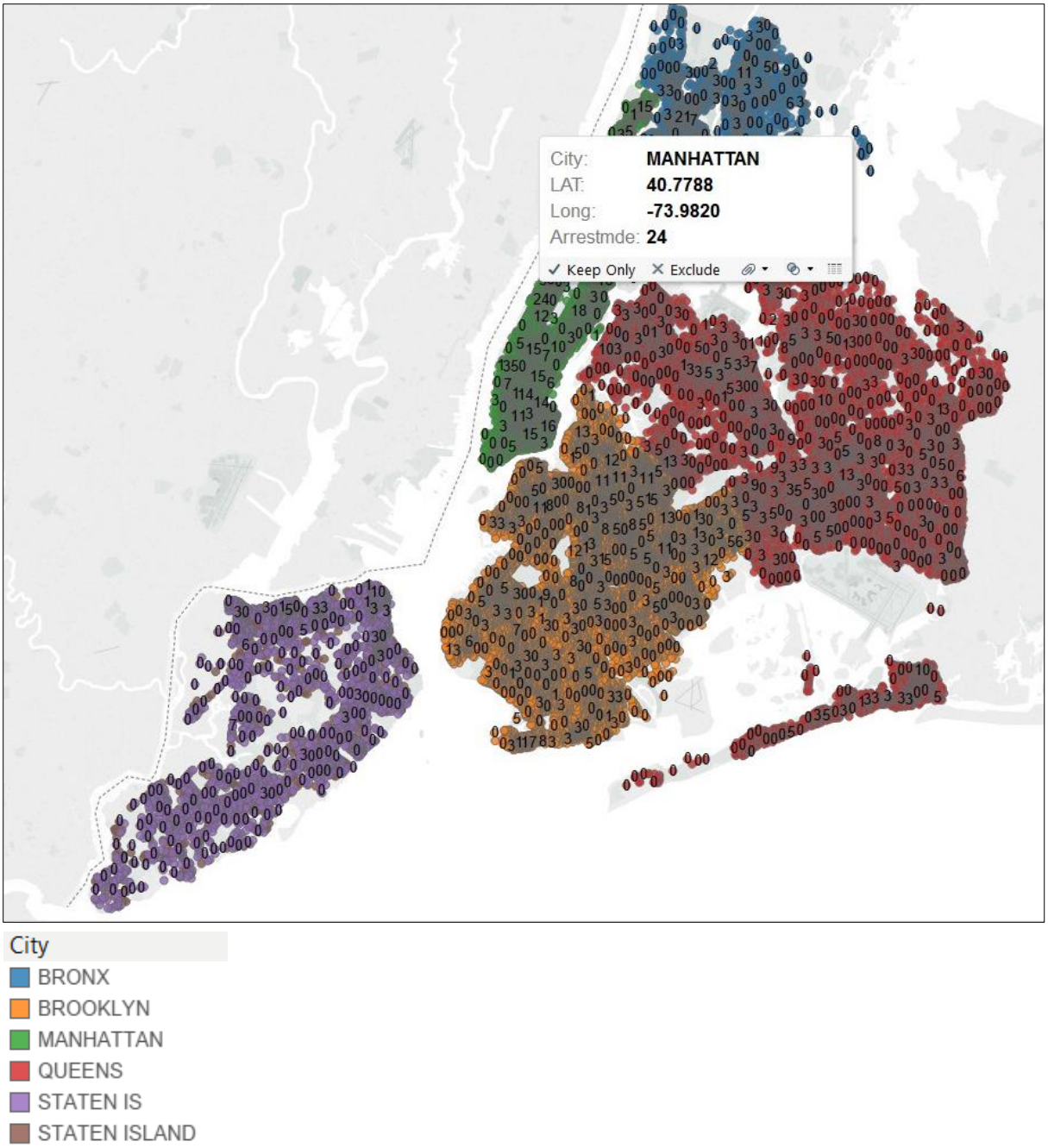


**Figure 2. Count of Arrests made in New York**

The above graph shown in Fig. 2 illustrates the number of people who were arrested in different locations in New York. 0 indicates the location where people were frisked for activities but not arrested. Bronx region has the highest number of arrests. Efforts are to be made by police to have more personnel around that area and have round the clock security to curb such activities.

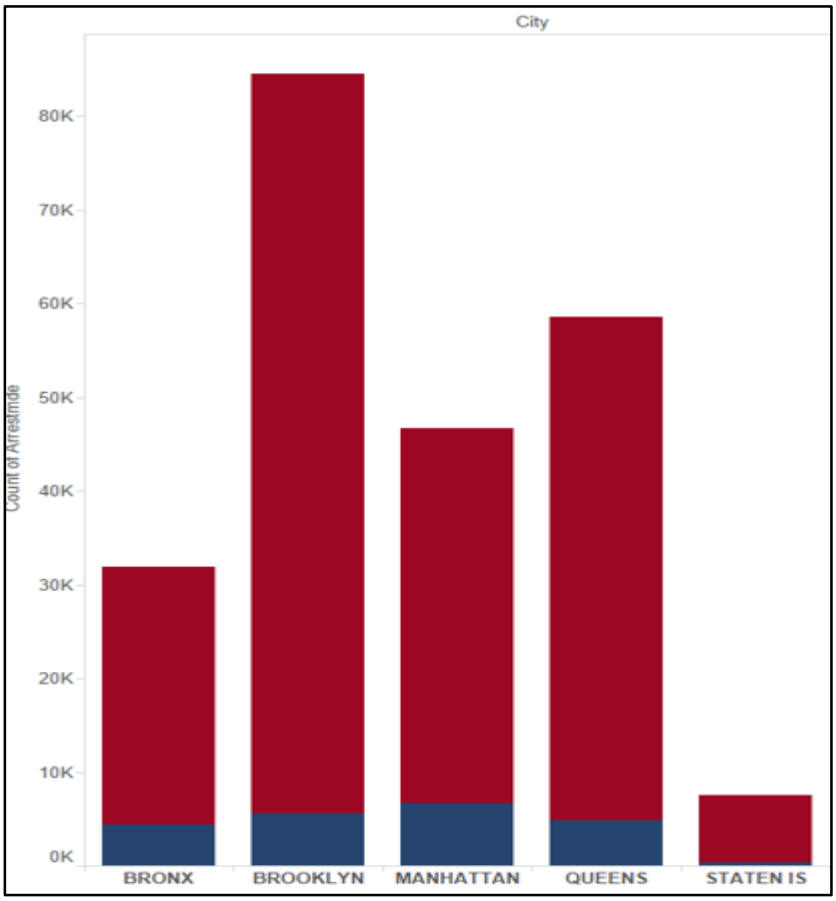


**Figure 3. Frequency distribution of frisk made and arrest made by region**

Fig. 3 shows the distribution of no. of persons frisked and number of persons arrested. The number of people who were frisked is represented in Red and the number of people who were arrested after being frisked is represented in Blue.

Manhattan has the highest number of suspicious activities and also higher number of people being frisked. This region also witnesses high number of people being arrested.

Bronx and Manhattan are the regions Police force should be increased to curb unlawful activities which are very high in these regions.

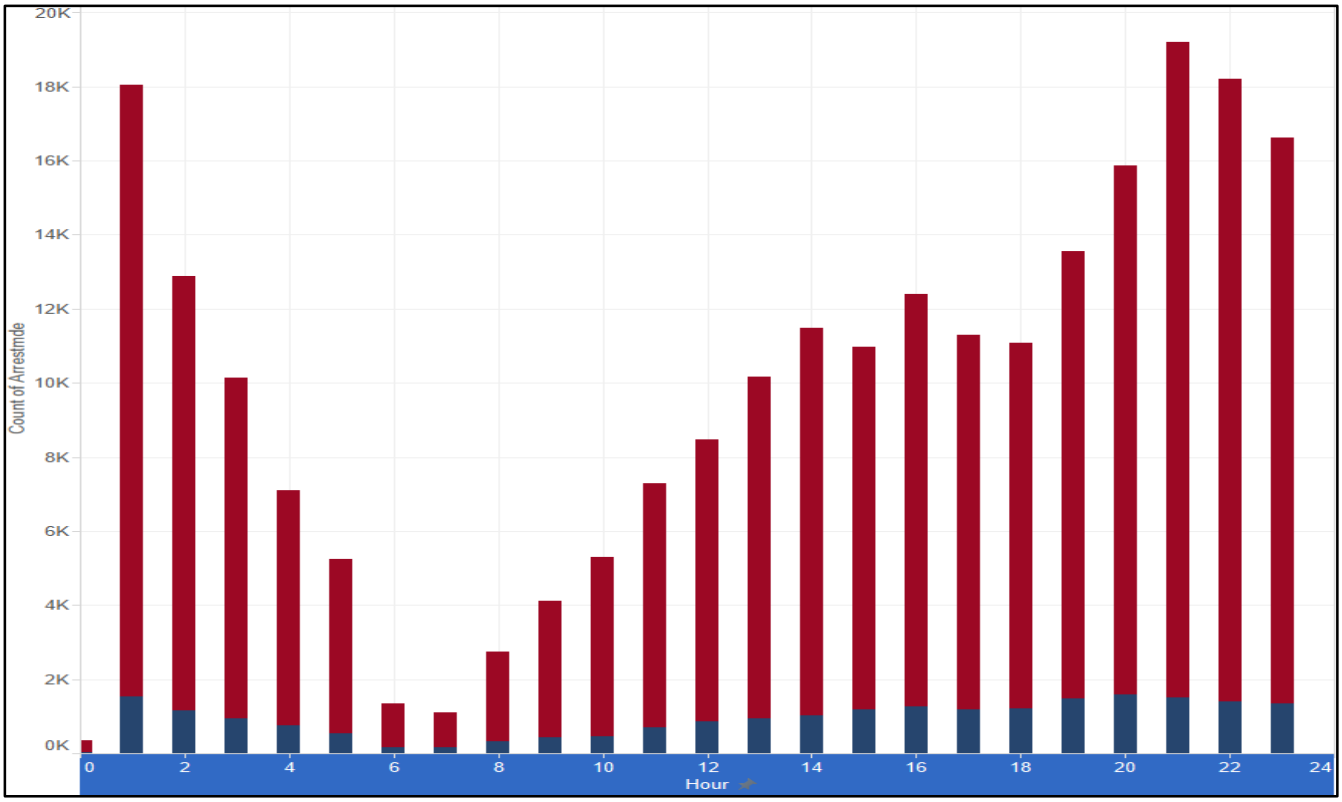


**Figure 4. Bar Chart of Time of the day vs arrest made**

Most of the frisk activities happen in the night during the peak hours between 10 PM and 1 AM. But surprisingly highest number of frisk activities and arrests are happening around 4PM during the day. People getting frisked and arrested around this time are between age groups 18 to 25.

But number of arrests is increasing as the day progresses implying the persons being arrested have tangible links with time. The police should find ways to reduce these activities by being vigilant and create awareness.

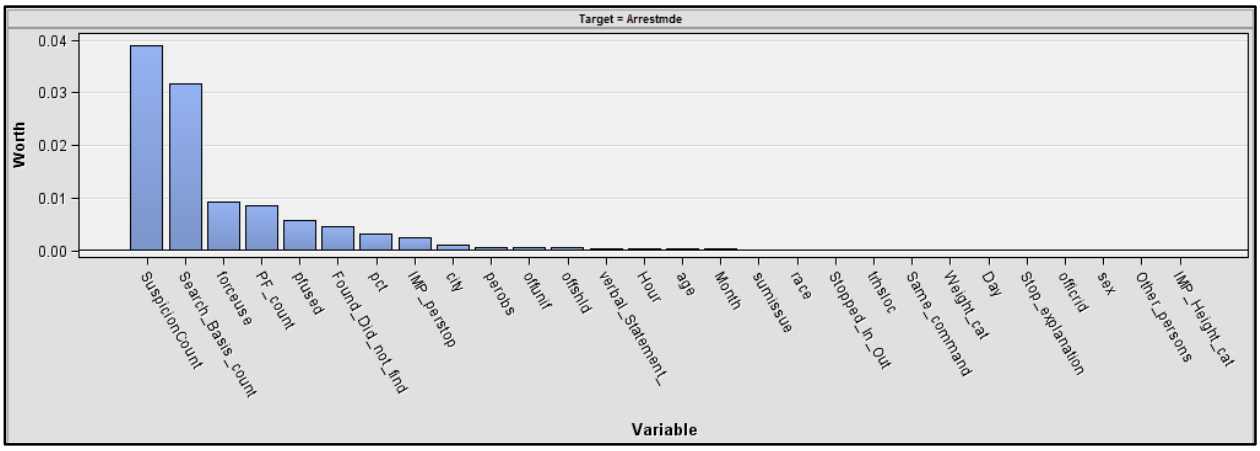


**Figure 5. Variable worth estimate**

## MODEL BUILDING

Data is partitioned into training and validation datasets by a ratio of 60:40. Data mining models such as Decision tree, Neural Network and Logistic regression and Polynomial regression with stepwise model selection are built using SAS® Enterprise Miner 12.3. These models were later compared using Model comparison node in order to evaluate the best model using validation misclassification rate as the selection criteria. The model diagram is shown in Figure 6.

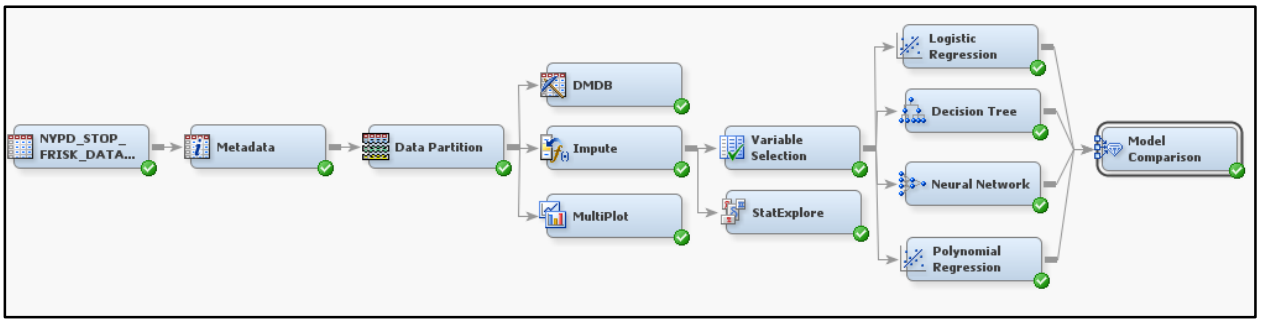


**Figure 6. Model Diagram (Nodes)**

Using Model Comparison Node in SAS® Enterprise Miner 12.3, competing models were diagnosed and compared with each other. Validation Misclassification rate is used as a selection criterion for selecting the best model. Table 2 shows the Fit statistics of Model comparison node with Polynomial regression model outperforming all the other models with validation misclassification rate of 6.54%.

| Selected Model | Model Node | Model Description | Target Variable | Selection Criterion: Valid: Misclassification Rate |
|---|---|---|---|---|
| Y | Reg4 | Polynomial ... | Arrestmde | 0.065375 |
| | Neural | Neural Net... | Arrestmde | 0.065546 |
| | Reg | Logistic Re... | Arrestmde | 0.069426 |
| | Tree2 | Decision Tr... | Arrestmde | 0.06593 |

**Table 1. Model Selection Fit Statistics**

The significant variables in the model are persons stopped, precinct number, location of the activity, force use, stop activity inside the house or outside the house, month of the criminal activity, number of reasons for physical force being used, count of suspicions, number of reasons for search basis. Interaction effect is seen between physical force and number of persons observed. Seasonality effect can be seen because of interaction between month and suspicion count variables. These variables and their interactions are based at 95% confidence interval.

| | | Predicted | |
|---|---|---|---|
| | | Arrest not made(0) | Arrest made(1) |
| Actual | Arrest not made(0) | 83568 | 4706 |
| | Arrest made(1) | 1388 | 4150 |

**Table 2. Confusion Matrix in validation data**

Confusion matrix given in Table 2 shows the actual by predicted values for the binary target variable. Sensitivity or the arrests made that are correctly identified value from the confusion matrix is 47% and specificity or the arrests not made that are correctly identified is 98.4%.

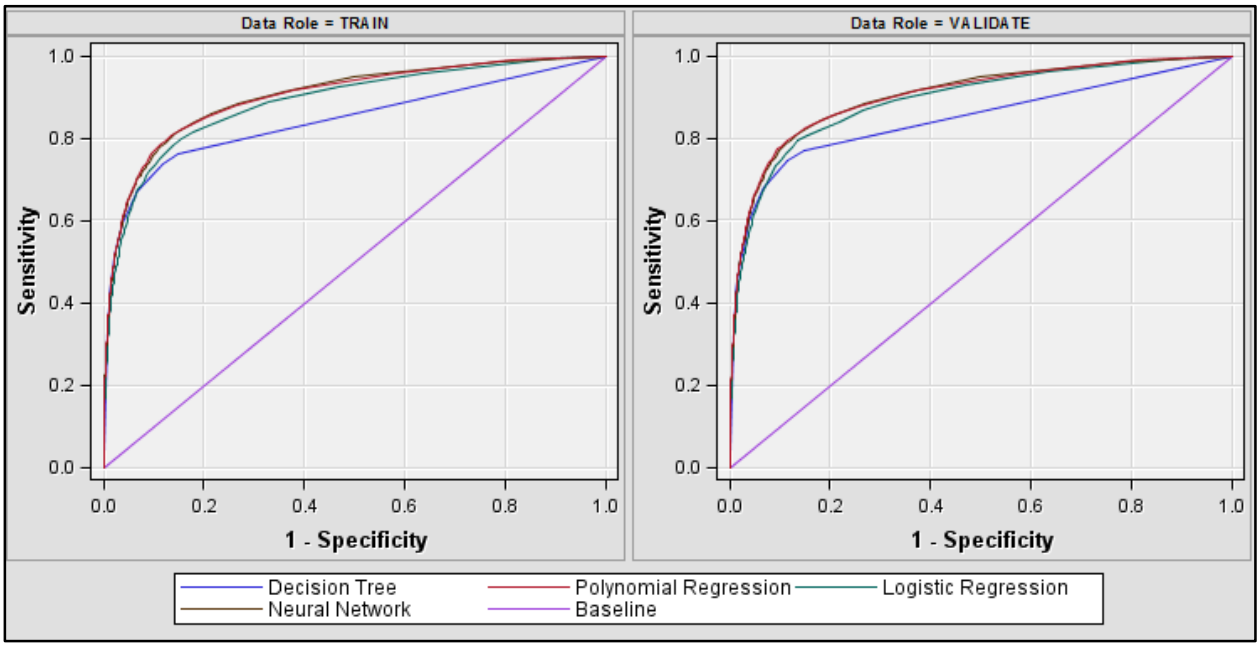


**Figure 7. ROC Chart**

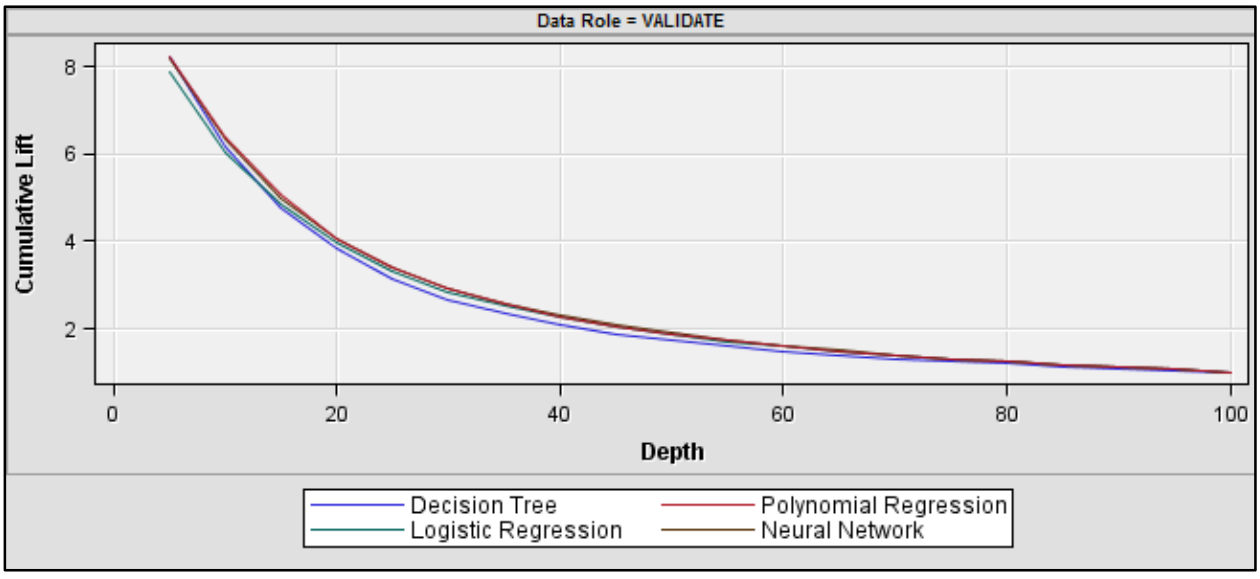


**Figure 8. Cumulative Lift graph of all models**

The model is able to distinguish whether the arrests were made or not with a validation misclassification rate of 6.54%.

## CONCLUSION

The results from the analysis shows that Bronx and Manhattan are the areas with high percentage of arrest. In Staten Island even though the stop density or the number of stops per area is high, the total number of arrests is comparatively low. Police personnel can be reduced in these areas and redirected to other regions where unlawful activities are more. The variables such as suspicion count, search basis count, searched basis, force used, pf count, pf used are the important variables that determine if an arrest should be made or not.

The number of persons frisked to number of persons arrested is highest in the region of Manhattan. Police should make extra efforts in this region to curb unlawful activities; force used in this region should be increased by at least 10% to curb these unlawful activities.

Location related variables seems to be one the most important variables. There are always high risk regions within the city where criminal activities are the highest so police are concentrating their efforts only in these regions but they also need to be more vigilant in other low and medium risk regions which do not lie in the vicinity of the precinct locations. One proposed solution to overcome this problem can be to introduce surveillance cameras and monitor those regions effectively.

There is interaction effect observed in the model between physical force and persons observed. Generally, police tend to stop and arrest people moving in groups because of suspicion. But since criminal activities are committed by individuals or small groups, police should not resort to use unwarranted physical force on large groups. Unreasonable usage of physical force may lead to aggravation of the situation and can cause riots, such situations can be dealt with verbally or by requesting more backup instead of using physical force.

Seasonality effect can be seen because of interaction between month and suspicion count variables. Certain months have a high number of suspicious activities going on in the city. So if police department can come up with new surveillance methods to curb these suspicious criminal activities, number of offenses can be reduced.

## REFERENCES


Book Predictive Modeling with SAS Enterprise Miner. 2nd ed. Kattamuri S. Sarma

Book Agresti, A. 2013. Categorical Data Analysis. 3rd ed. Hoboken, NJ: John Wiley & Sons.

Journal Article Davidian M,2001 “Nonlinear Models for Univariate and Multivariate Response”

Journal Article Derr, R.E. (2000), “Performing Exact Logistic Regression with the SAS System,“ Proceedings of the 25th Annual SAS Users Group International Conference (SUGI 25), 254-25.

Web Article Catilin Dempsey, 2012, “Criminal Mapping and Analysis”, URL: - “https://www.gislounge.com/crime-mapping-and-analysis”

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Karan Rudra
Oklahoma State University
405-612-5103
krudra@okstate.edu

Maitreya Kadiyala
Oklahoma State University
405-612-9309
maitreya.kadiyala@okstate.edu