Data Mining & Predictive Analytics (BUDT758T)

Project Title: Arrests Predictor – Predict and classify NYPD Dataset

Team Members: Sagar Maniar
Sarthak Potnis
Sneha Saxena
Zhengzhou Wang

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

SAGAR MANIAR
SARTHAK POTNIS
SNEHA SAXENA
ZHENGZHOU WANG

Contact Author

Contents

I. Executive Summary

The project aims to provide a hands-on experience in solving a real world problem based on a large dataset. This entailed obtaining the dataset, analyzing and cleaning the data, building models to solve the problem at hand and arrive at a recommendation based on comparison of models.

Arresting suspects in public places can be challenging at times. Several factors are in play and the officers have no more than a few minutes to decide if a suspect is to be arrested or not. Of all the arrests made, a sizeable percentage is of wrongful arrests. Through this project we aim to build a predictive model that can aid in making the decision to arrest a certain suspect, more importantly to help reduce the incidence of wrongful arrests.

The dataset being a public dataset it required extensive cleaning. After intense data cleansing we built five different models namely linear regression, logistic regression, knn, naive bayes and random forests. Among these we found the logistic regression model to be the better model with respect to values of accuracy, sensitivity and specificity on the validation data set. We also found that the ROC curve for this model was better than those of the other models.

Though the study limits to the transportation area, the models developed can be scaled up/ down and tailored according to their dataset. The study could be used in other areas (e.g. court, public security).

II. Data Description

The data set was downloaded from the New York Police Website.

http://www.nyc.gov/html/nypd/html/analysis_and_planning/stop_question_and_frisk_report.shtml

The most updated one when the project started was the 2014 dataset.

Since the dataset is taken from a public resource. The data requires a significant amount of pre-processing/cleaning. The original dataset contains 112 variables. The whole dataset was loaded into IBM Watson, the team shrinked the dataset by selective choosing the most contributing variables from IBM Watson report.

There are a number of missing values in the variables, several of the variables are not interpretable. After removing those variables, the team build a dataset using all the variables, based on the P-value of them, following 55 variables are chosen.

```
 [1] "arstmade" "pct"      "ser_num"  "timestop" "trhsloc"  "perobs"   "perstop"  "typeofid" "othpers"  "sumissue"
[11] "offunif"  "frisked"  "searched" "adtlrept" "pistol"   "riflshot" "pf_hands" "pf_wall"  "pf_grnd"  "pf_drwep"
[21] "pf_ptwep" "pf_baton" "pf_hcuff" "pf_pepsp" "pf_other" "radio"    "rf_vcrim" "rf_othsw" "rf_attir" "cs_objcs"
[31] "cs_descr" "cs_casng" "cs_lkout" "rf_vcact" "cs_cloth" "cs_drgtr" "ac_evasv" "ac_assoc" "cs_other" "ac_incid"
[41] "ac_time"  "sb_other" "repcmd"   "revcmd"   "sex"      "age"      "ht_feet"  "ht_inch"  "weight"   "build"
[51] "city"     "addrpct"  "xcoord"   "ycoord"   "detailCM"
```

The categorical value the study tries to predict is the “arstmade”. 1 being unarrested, 2 being arrested. Most of the variables are being treated as categorical expect: pct, ser_num, timestop, perobs, perstop, repcmd, revcmd, age, ht_feet, ht_inch, weight, addrpct, xcoord, ycoord, detailCM. Details regarding the variables can be found in data dictionary attached in Appendix.

Due to the limitation of R program and computer capability, the sample size the study choose is 1000 records. Below is a small sample observation of the data we used. For the implementation of different models (e.g. Naive Bayes), several variables are normalized to fit the model.

The dataset is interesting, as the police officer actually input the data, recording the suspect’s physical features, the kind of physical forces that were used by police officer, any weapon carried by the suspect, what kind of weapon were carried. Since the data are entered by police officers, there are some human errors in the process of record making. For example, a person’s age can’t be 300 year. Since there is no way to verify the real data the officer trying to record, the data was taken out from the dataset to ensure the accuracy of the prediction model.

The dataset contains the information regarding people who were stopped by New York police officer, some of the variables are suspect's gender, age, race, where the person was stopped, at what time, how long is the stop action. Specific type of physical forces that were used by the police officer, location of the stopped etc. The study tries to classify given a person's stopped time, location and physical feature (e.g. Build, height, race, gender, age), would this person be arrested or not.

III. Methodology

After cleaning and analyzed the dataset, the team initiated with the process of modelling and creating predictive and analytical models to formulate a model towards the prediction of a person's arrest.

Due to the capability of R, the team cannot take half of the dataset as training data and use the others as validation. After multiple trails, the training dataset is set at 1000 records level, the performance is measured on randomly selected validation dataset of 700 records.

The team started by implementing certain models such as Linear Regression model, Logistic Model, Random Forest. The team attempted more complicated models such as KNN, Bayes Model and Clustering. The aim is to create models and compare different models according to their performance on the validation data det.

The above listed models were created and compared on the basis of their ROC curves, Sensitivity, Specificity and Accuracy of the models. The main focus was made on the Accuracy and Specificity of the model. Because according to the specificity the people who should not be accused will not be arrested if the prediction specificity along with the accuracy of the model is high. This was the reason the team mainly focused on the specificity and the accuracy of the models.

Following are the values of the Accuracy, Specificity and Sensitivity of the models along with the comparative ROC Curve.

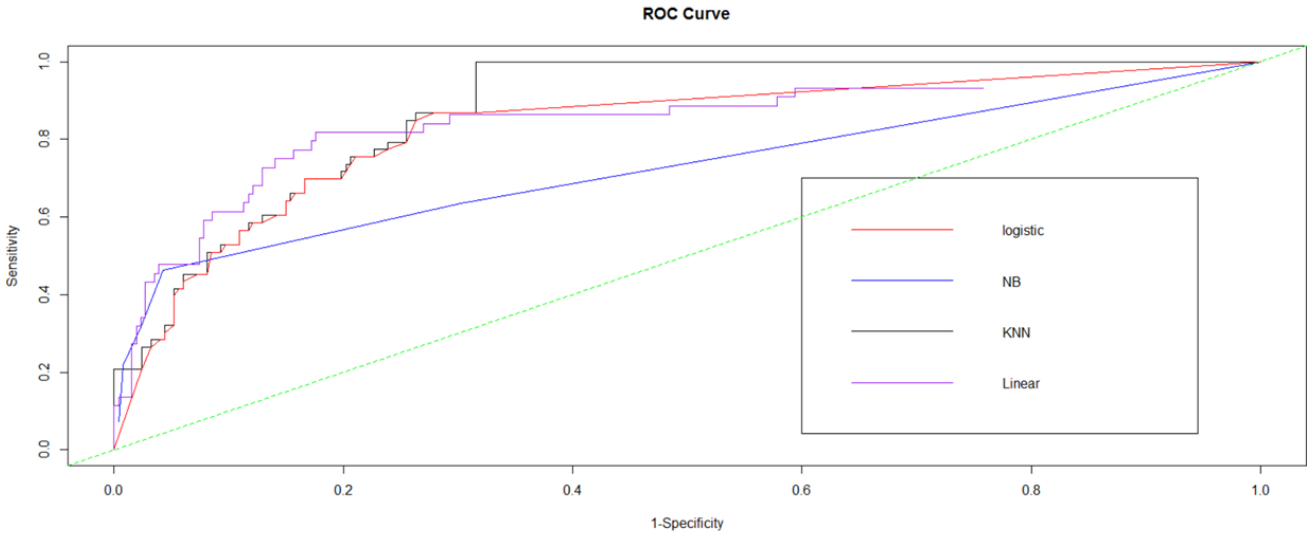
IV. Results and Finding

The model is built to identify of a person would get arrest or not. In the application perspective, the study would help the police officer to decrease the occurrences of wrongly arrest. Below is a brief summary of the best-performed models:

	Accuracy	Sensitivity	Specificity
• Linear Regression	0.90	0.35	0.97
• Logistic Regression	0.91	0.47	0.96
• K nearest Neighbor	0.86	0.76	0.87
• Naïve Bayes	0.90	0.80	0.88

To choose the best model, the models were compiled into one ROC curve.

ROC Curve



Among the various models implemented, the logistic model outperforms other with a 91% accuracy rate and it is also the one that is most close to the top left corner of the ROC graph. In additional to that, since the purpose of the study is to lower the rate of wrongful arrest, (false negative), specificity plays a critical role in the model selection, logistic model has a 96% specificity and is the model the study recommend.

The team also used other method, such as random forest, clustering, association rules. With pruned tree, the accuracy rate is 89%. (Appendix1)

For KNN model, after computing the accuracy and specificity, error rates of different K are compared (Appendix2, 3). The optimal K is nine. Naïve Bayes' result is attached in Appendix4. For random forest model, error rate is plotted against each K size, the optimal K size is nine. (Appendix 5). A screenshot of pruned tree is attached in Appendix 1.Due to the nature of the dataset (Appendix 6). Clustering model gives the team two clusters, one defined as arrested, the other not arrest. Since the cluster number in this study is low, clustering does not have a good performance. The team attempted association rules, due to the limitation of computer capability, the model cannot be implemented successfully. The code for association rule is included in the electronic submission.

In this study, complicated models do not outperform the simplified ones. With more data, the simpler solution (estimating a distribution with a histogram) actually becomes more accurate than the sophisticated solution (estimating the parameters of a model using a linear regression).

V. Conclusion

The project was aimed at creating predictive models to serve the main aim of predicting if a person who is stopped should be arrested by the NYPD officials or not. Our findings and analysis resulted into removal of certain unwanted and less related variables. The model classified people who get arrested into get arrested and not arrested based on suspect's race, age, build, location of stop, and stop.

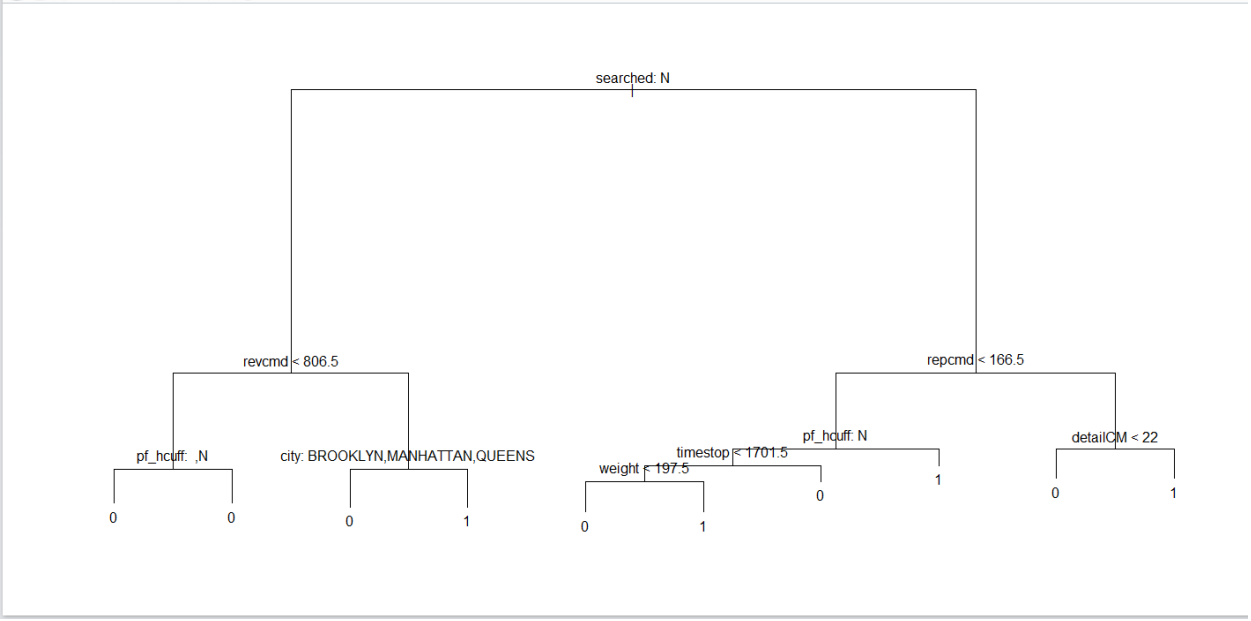
The best model is Logistic Regression Model which gives out an accuracy of 91% and specificity of 96%. We have used this criteria for the selection of model as it justifies the prediction of people who should not be arrested to the best as compared to the other models and this will be very useful to reduce the number of wrongful arrests.

This model can be used to create a check list for the police officers so that they can realize based on the prediction of the model if the person who is stopped should be arrested by them or not. The model can be implemented also for different years and will be very helpful as we will be able to compare the predictions for different years which will suggest a formulation of a new rule or new behavior which may be leading to a certain variation in the trend.

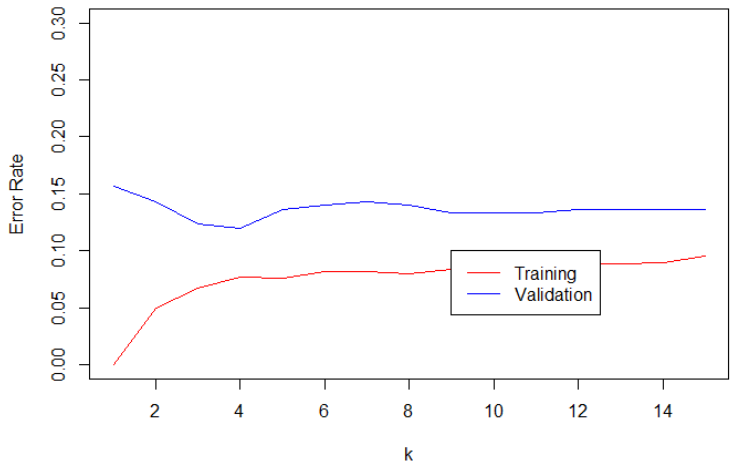
The model may also be used to understand and analyze the effects of the person's appearance, age, past record and other factors and the extent to which these factors affect the decision of an officer to leave or arrest the person. The study could be further applied in other safety departments, with modification on the parameters.

VI. Appendix

1. Pruned tree.

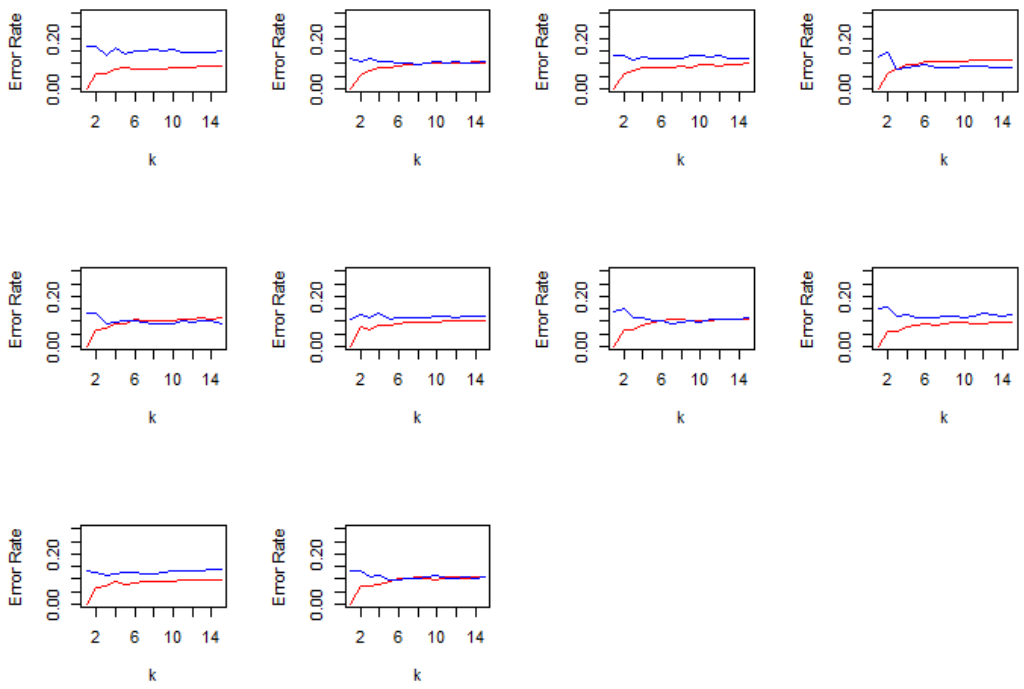


2. Knn




```
> CM1
   prediction
      0   1
  0 603   2
  1  45  50
> CM2
   prediction2
      0   1
  0 246   9
  1  31  14
```

3. Knn



Minimum Validation Error k: 3Minimum Validation Error k: 8Minimum Validation Error k: 3Minimum Validation Error k: 3Minimum Validation Error k: 9Minimum Validation Error k: 1Minimum Validation Error k: 7Minimum Validation Error k: 6Minimum Validation Error k: 3Minimum Validation Error k: 5

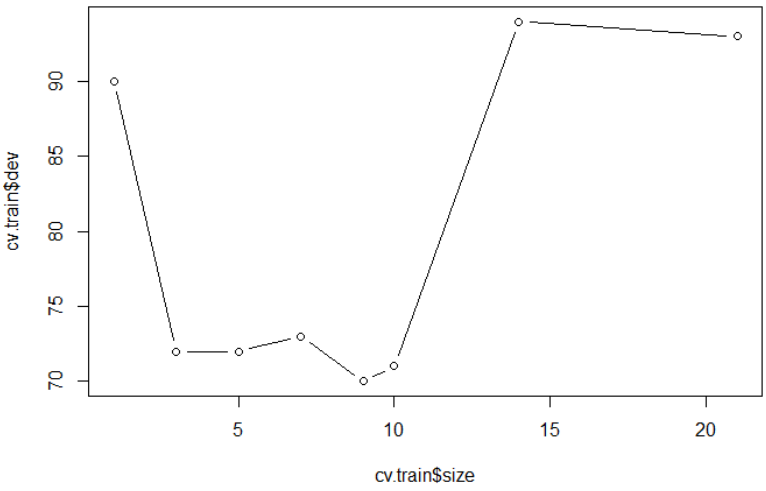
4. Naïve Bayes Table

```
      predicted
actual   0   1
     0 234  22
     1  20  24
```

NaïveBayes Apriori

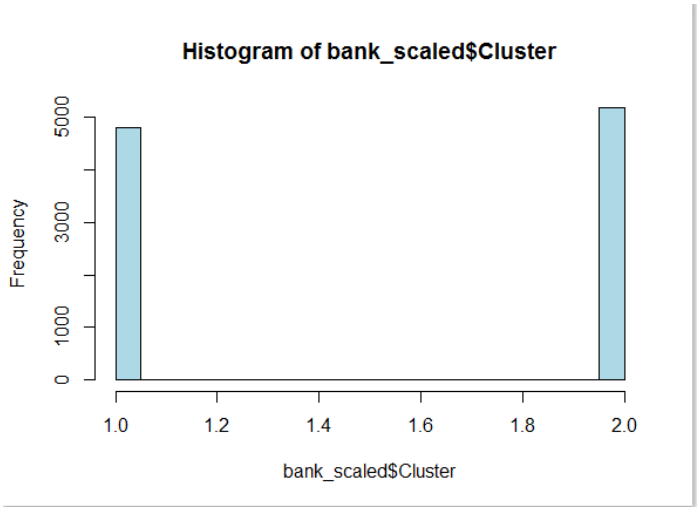
```
Y
  0   1
595 105
```

5. Tree selection



Optimal number of K=9

6. Data distribution



Histogram for clustering of arrest made in the dataset