Business Analysis with SAS

Apurva Godghase

Table of Contents:

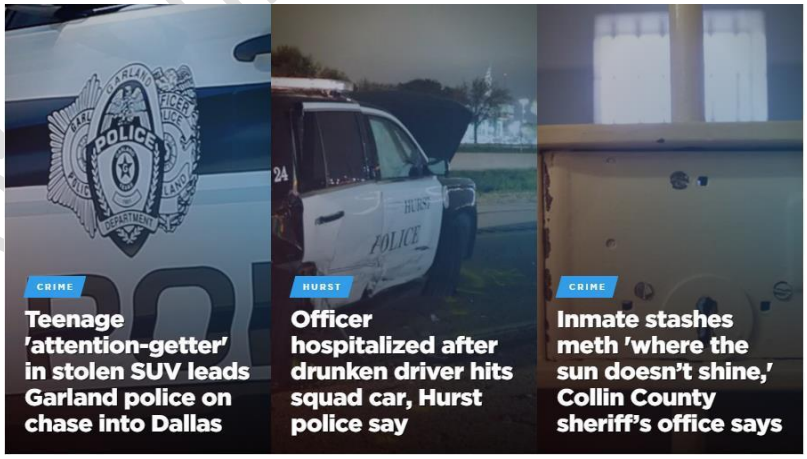
1. Executive summary

In our day to day life, it becomes necessary to understand the role of law and order in the society, specifically, to generate class discussions about both the complexities of this role and the criminal activity in general. They provide examples of real problems and situations faced by police officers at work.

The type of crime plays major role in deciding arrest during these interactions. Place and time of occurrence are contributing factors as well. We are trying to predict the generic behavior or traits of citizens that are involved in the interaction.

Whenever an officer responds to resistance or aggression by using certain levels of force, all the associated details are investigated and recorded by the Dallas Police Department (DPD) in a Response to Resistance report. This data set comes from those reports, which are entered DPD's records management system. We acknowledge there is always the possibility of mechanical or human error. The City of Dallas does not guarantee the accuracy of the information. The City of Dallas will not be responsible for any error or omission, or for the use of, or the results obtained from the use of this information.

2. Project motivation:



Since the inception, law enforcement has battled with action and reaction whether they were aware of it or not. Moreover, law enforcement is primarily a response to a stimulus, which again leads to an interesting question:

Is it possible to gauge the probability of citizen getting victimized/ assaulted by understanding action and reaction patterns to make it work for law enforcement instead of against? Is it possible for Police to focus during some scenario to lower the crime rate?

These thoughts kindled the idea of researching more on the data provided by Dallas Open Data website(https://www.dallasopendata.com/) to understand the 'action reaction' pattern between Police and citizen.

3. Data Description:

The 2015 police data records responses from police officers about interactions happened between officer and the citizens. Given any set of data, there are endless combinations of attributes that one can use to represent the data. For a preliminary examination of the data attributes, we divided the police officer record, citizen record, interaction record (force type, arrest, injury) time, geolocation as main types. This division enabled us categorized the data for analysis and narrow down the results

Time related attributes – Occurred time, Occurrence date, Officer Hire date

Officer related attributes – Badge number**,** Officer Sex, Officer Race, Officer Injury, Officer Condition, Officer in hospital, Service Type

Citizen related attributes – Citizen Number, Citizen Race, Citizen Sex, Citizen Injury, Citizen Condition Type, Citizen Arrest, Citizen Assessment type, Citizen charge type

Location related attributes- RA, Beat, Sector, Division, District, Geolocation

Interaction Related attributes- ID, File number, Force type, Force Effective, UOF Number

4. Data preparation activities

To build models appropriate for analyzing the data, it is a pre-requisite to transform the data into a required format and thereby clean that data to avoid any discrepancies (unaccepted values), missing values or unknown values that may be present in the existing dataset.

a. Date Format changes

We first changed the date format of the attributes which were heterogenous in nature to convert it into a homogenous date format.

For an instance, in the following attributes:

OCCURRED_DT – was the field which contained the date along with the time stamp
OCCURRED_TM – contained the time

There was a redundancy in terms of time captured which was making it heterogenous.Hence, there was a need of conversion to a common format with the removal of timestamp variable for the effective use of SAS enterprise miner for analysis purposes.

b. Removing blank columns

We removed blank columns named longitude and latitude as it did not have any data.

c. Timestamp

The effective timestamp available in the dataset was in the 12hour format and we realized that there is a need of conversion to a 24hour clock format.

d. Consistency

For Force type & Force Effective- There were instances where in this field contained multiple values, so to maintain the integrity with other columns, these were reduced to single values.

e. Most Probable Value

For CitChargetype, there were certain fields where in there were missing value problems and we realized that an effective solution will be the replacement of the unknown values to the most probable values for data analysis purpose.

f. Binning

There were certain instances in the dataset where broader categories were segregated into sub categories which would lead to ambiguous clusters. To get the meaningful clusters as a part of the output we kept them under one broad category.

To manipulate the time in the when most of the incidents took place we decided to create categories in time across 24 hours in day. We used binning to create categories in time.

Morning	06 AM – 12 PM
Afternoon	12 PM – 6 PM
Evening	06 PM – 12 AM
Late Night	12 AM – 6 AM

g. Dummy Coding

Dummy coding was applied to convert the nominal data to interval data to the fields having some code values to ensure easy and meaningful analysis.

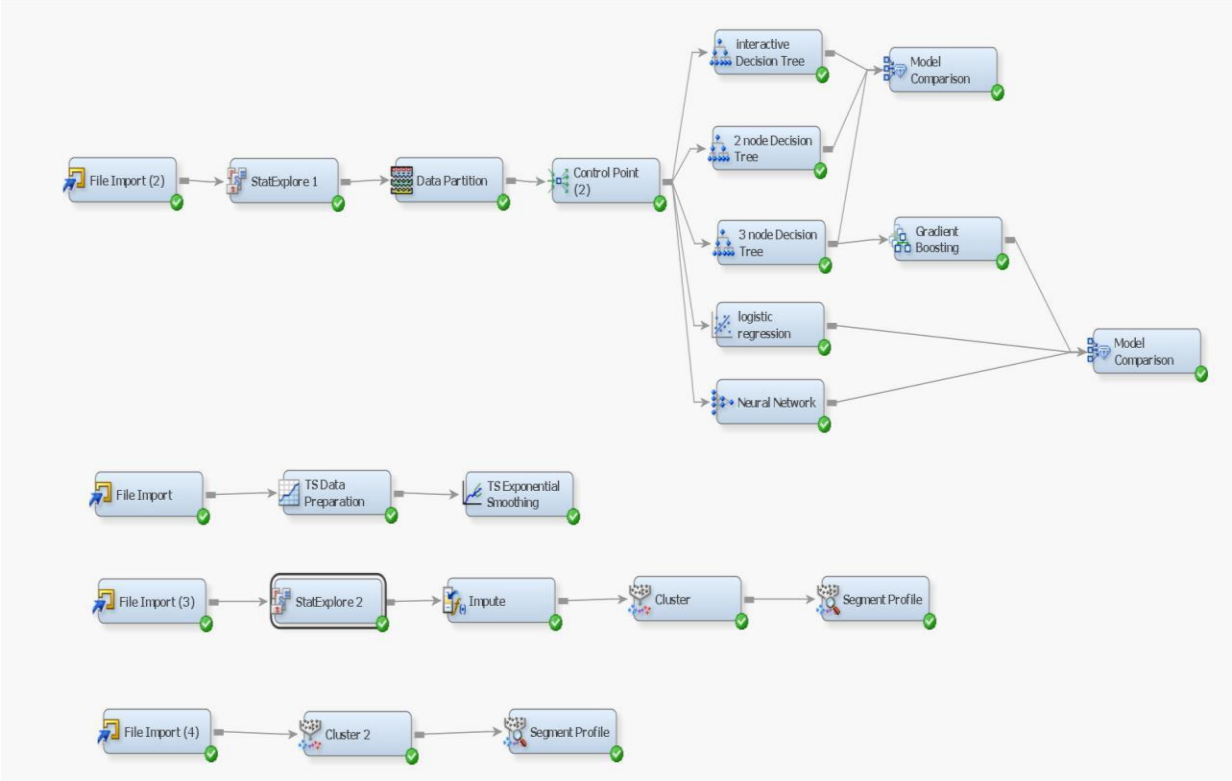
h. Data Attributes Explanation:

Variable	Description	Status
ID	Id as per the data set--	Reject
FILE NUM	File of the incident captured	Reject
OCCURRED_DT	Date when the incident happened	Input
OCCURRED_TM	Time when the incident happened	Input
CURRENT_BADGE_NO	Current Badge Number	Reject
OffSex	Officer's sex	Input
OffRace	Officer's race	Input
HIRE_DT	Date when officer was hired	Input
OFF_INJURED	Was the officer injured or not	Input
OffCondType	Condition if the officer was injured	Input
OFF_HOSPITAL	If the officer was admitted to hospital or not	Input
SERVICE_TYPE	Type of service provided by the police department	Input
UOFNum	UOF Number	Reject
ForceType	How did the police officer applied the force (held, grabbed	Input
ForceEffective	Conditional variable stating if the applied force was effective	Input
UOF_REASON	Reason of UOF	Input
Cycles_Num	Number of interactions the police has with the	Input
CitNum	Unique identifier for the citizen	Reject
CitRace	Race of the citizen	Input

CitSex	Sex of the citizen	Input
CIT_INJURED	Whether the citizen was injured or not	Input
CitCondType	What kind of injuries citizen had	Input
CIT_ARRESTED	Whether the citizen was arrested or not	Input
CIT_INFL_ASSMT	Initial assessment of the citizen at the time of the first	Input
CitChargeType	What kind of charges were imposed on the citizen	Input
ARC_Street	Street address of the incidence	Input
RA	Area code of the incidence	Input
BEAT	Beat Code of the incidence	Input
SECTOR	Sector number of the incidence	Input
DIVISION	Division of the incidence	Input
DIST_NAME	District Name of the incidence	Input
GeoLocation	Geo Location of the incidence	Input
Latitude	Missing	Reject
Latitude	Missing	Reject

The attributes marked reject have not been used in creation of models in SAS Enterprise Miner

5. BI Model Overview



6. Analysis Problem

Based on this data we attempt to answer following set of questions.

6.1 Descriptive:

Which is the region and time where most of the interactions happen? Is any specific race of citizen is observed most in all interactions?

Is any specific gender observed in interactions?

6.2 Predictive:

What is the trend of citizen getting arrested in year 2015?

What are the number of incidences occurring in a stipulated time period in the following year?

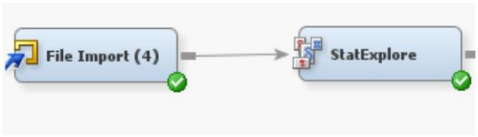
What are the most risk prone areas of the city?

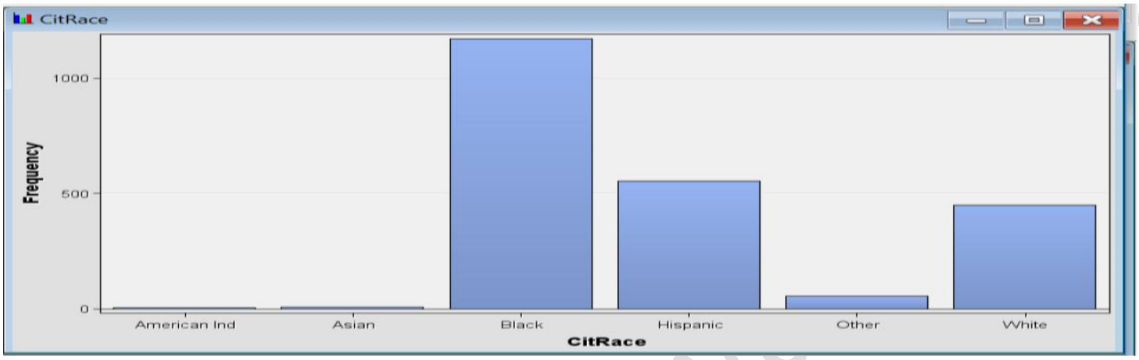
Predicting the outcome in terms of injury of the interaction between both citizen and the involved police office.

Predicting the correlation between citizen arrests and various seasons and specific time of year.

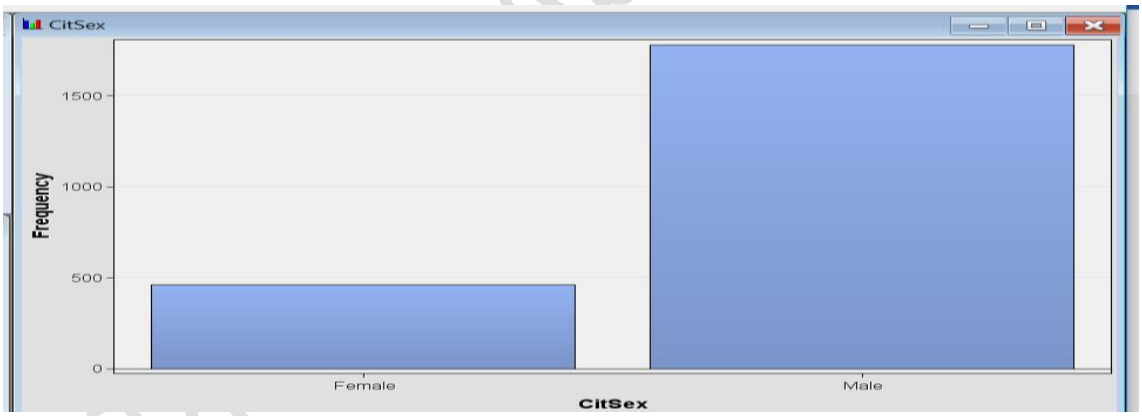
7. Findings

7.1 Analyzing the output of StatExplorer node, it is evident that black was highest involved race in the interactions.





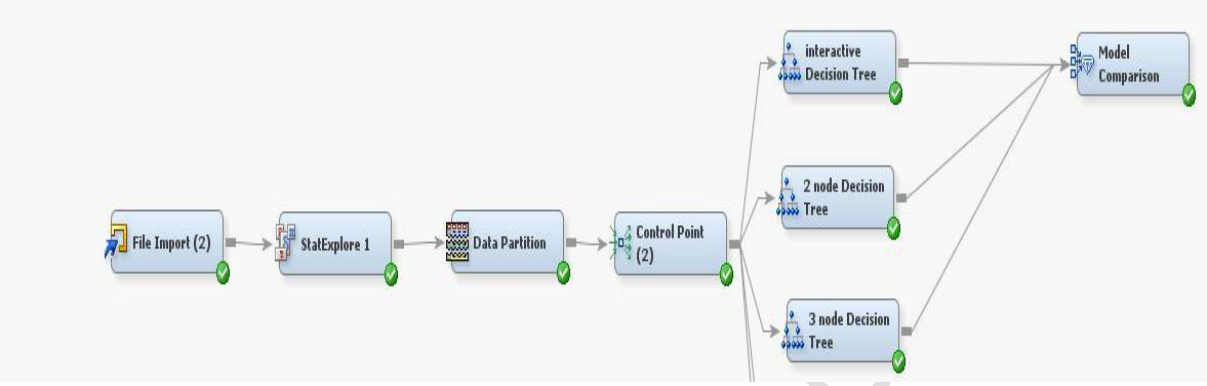
7.2 Also in all the interactions that police had Male citizens were heavily involved than women



7.3 Classification Model:

Classification technique was used to achieve third data mining objective of aiding accurate predictions and analysis. Here it discovers how the set of input and target parameters reach the conclusion

7.3.1 Decision Tree



a. **Variable Selection:** For the analysis, target is a binary variable. YES, corresponds to citizen was arrested and NO corresponds to citizen was not arrested. All the input variables were used for the classification technique are shown below with the roles and levels.

Variable	Role	Level
CIT_ARRESTED	INPUT	NOMINAL
CIT_INFL_ASSMT	INPUT	NOMINAL
CitSex	INPUT	NOMINAL
OFF_HOSPITAL	INPUT	NOMINAL
CitRace	INPUT	NOMINAL
OFF_INJURED	INPUT	NOMINAL
ForceEffective	INPUT	NOMINAL
OffSex	REJECTED	NOMINAL
ID	REJECTED	NOMINAL
OffCondType	REJECTED	NOMINAL
OffRace	REJECTED	NOMINAL
BEAT	REJECTED	INTERVAL
CitChargeType	REJECTED	NOMINAL
HIRE_DT	REJECTED	INTERVAL
ARC_Street	REJECTED	NOMINAL
OCCURRED_DT	REJECTED	INTERVAL
Years_Service	REJECTED	INTERVAL
UOFNum	REJECTED	NOMINAL
OCCURRED_TM	REJECTED	INTERVAL

UOF_Code	REJECTED	INTERVAL
SERVICE_TYPE	REJECTED	NOMINAL
UOF_REASON	REJECTED	NOMINAL
OR_CODE	REJECTED	INTERVAL
SECTOR	REJECTED	INTERVAL
C_Infl_Ass_Code	REJECTED	INTERVAL
Cycles_Num	REJECTED	NOMINAL
Osex_Code	REJECTED	INTERVAL
C_Race_Code	REJECTED	INTERVAL
C_ARR_CODE	REJECTED	INTERVAL
CURRENT_BADGE_NO	REJECTED	INTERVAL
CitNum	REJECTED	INTERVAL
Cit_Chrg_Type_Code	REJECTED	INTERVAL
CitCondType	REJECTED	NOMINAL
ForceType	REJECTED	NOMINAL
C_Sex_Code	REJECTED	INTERVAL
GeoLocation	REJECTED	NOMINAL
Coverted_OCCURRED_TM	REJECTED	NOMINAL
DIST_NAME	REJECTED	NOMINAL
DIVISION	REJECTED	NOMINAL
RA	REJECTED	INTERVAL
FILENUM	REJECTED	NOMINAL
Force_Eff_code	REJECTED	INTERVAL
CIT_INJURED	TARGET	BINARY

b. StatExplorer: Next we examined the distribution of different variables used in the dataset. From results, we can see that there are no missing values in selected input.

c. Data partition: For classification**,** we've partitioned the data sets into training, and validation data sets. Training is done for preliminary model fitting. So, 60% data was used for training. The validation data set is used to tune the model weights and for model assessment. We have used 40% data for validation purpose.

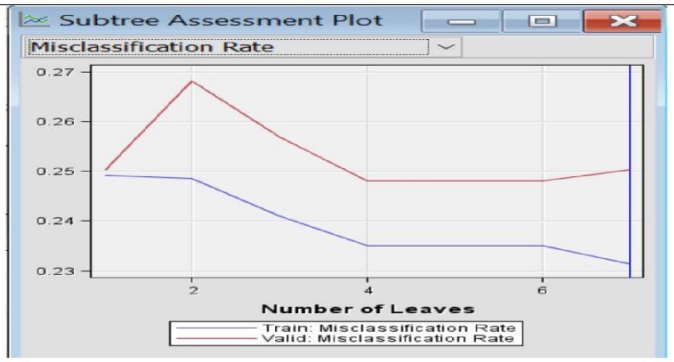
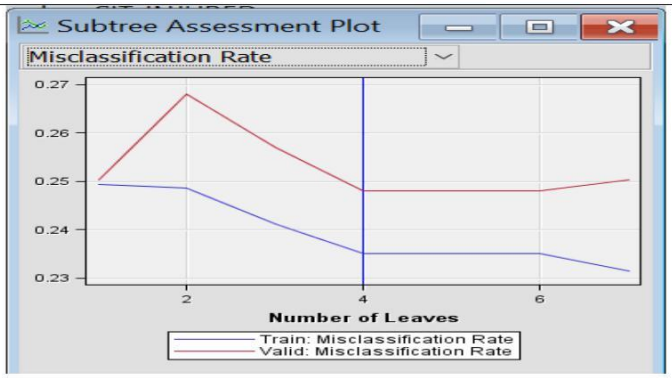
d. Control Point: For model comparison, we need to give same input to 5 different nodes . To reduce the complexity arising due to large number of connections that are made in process flow diagrams, we inserted a control point.

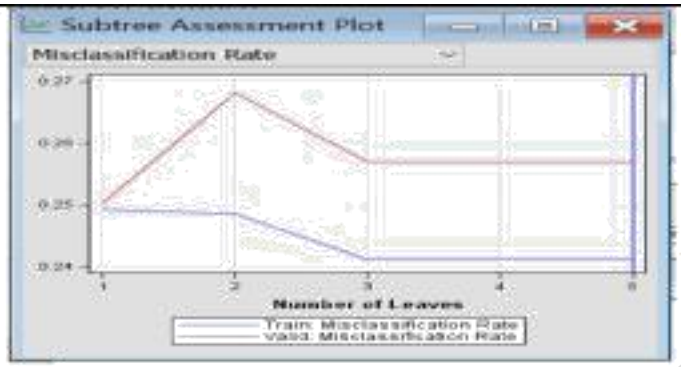
e. Decision Tree: Decision trees are very useful multiple variable analysis. To find the most optimum type of decision tree, we are implementing three different types of decision tree.

Models	Description
2 Node Decision Tree	A decision tree with maximum 2 leaf nodes. Misclassification subtree assessment is used here.
3 Node Decision Tree	A decision tree with maximum 3 leaf nodes. Misclassification subtree assessment is used here.
Interactive Decision Tree	A tree to prevent the excessive time and memory consumption that can occur with large data sets. After splitting the nodes up to desired level, the tree is frozen.

f. Classification Model Comparison (Results):

Subtree Assessment Plot

Models	Misclassification Rate
2 Node Decision Tree	
3 Node Decision Tree	

Interactive Decision Tree	

Misclassification Rate Table Comparison

Models	Misclassification Rate
3 node Decision Tree	0.2480
2 node Decision Tree	0.2502
Interactive Decision Tree	0.2569

Confusion Matrix: A confusion matrix is a tabular representation of the performance of the specified classifier. It is used to evaluate type I and type II errors.

- **For 2 Node Decision Tree:**

Confusion Matrix Terms: Calculated on Validation data

		Predicted	
		"+ve"	"-ve"
Actual	"+ve"	a	b
	"-ve"	c	d

		Predicted	
		yes	no
Actual	yes	43	292
	no	22	652

Accuracy: Proportion of correct predictions

Accuracy = (a+d) / (a+b+c+d)

= (43+642) / (43+292+22+652)

= 685/1359

= 67%

- **For 3 Node Decision Tree:**

Confusion Matrix Terms: Calculated on Validation data

		Predicted	
		yes	no
Actual	yes	18	207
	no	16	658

Accuracy: Proportion of correct predictions

Accuracy = (a+d) / (a+b+c+d)

= (18+658) / (18+207+658+16)

= 676/899

= 75.19%

- **For Interactive Decision Tree:**

Confusion Matrix Terms: Calculated on Validation data

		Predicted	
		yes	no
Actual	Yes	16	207
	No	18	658

Accuracy: Proportion of correct predictions

Accuracy = (a+d) / (a+b+c+d)

= (18+658) / (18+207+658+16)

= 676/899

= 75.19%

Models	Accuracy from Confusion Matrix
2 Node Decision Tree	67%
3 Node Decision Tree	75.19%
Interactive Decision Tree	75.19%

On assessing the results of Misclassification rate, Confusion Matrix and Subtree Assessment Plot, we conclude that 3 node decision tree is best model.

g. Fit Statics for 3 node Decision Tree (Best Model Selected)
From fit statistics for 3 node decision tree, we observed that the value of Misclassification rate and Average squared error are almost similar for training and validation.

Misclassification rate: 0.2351 for Train and 0.248 for validation
Average Squared error: 0.176 for train and 0.184 for validation

This means the results are consistent and valid. So, we proceed forward with 3 node decision tree.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
CIT_INJURED	CIT_INJURED	NOBS	Sum of Frequencies	1344	899	.
CIT_INJURED	CIT_INJURED	MISC	Misclassification Rate	0.235119	0.248053	.
CIT_INJURED	CIT_INJURED	MAX	Maximum Absolute Error	0.775061	0.775061	.
CIT_INJURED	CIT_INJURED	SSE	Sum of Squared Errors	479.4127	332.3822	.
CIT_INJURED	CIT_INJURED	ASE	Average Squared Error	0.178353	0.184862	.
CIT_INJURED	CIT_INJURED	RASE	Root Average Squared ...	0.422319	0.429956	.
CIT_INJURED	CIT_INJURED	DIV	Divisor for ASE	2688	1798	.
CIT_INJURED	CIT_INJURED	DFT	Total Degrees of Freed...	1344	.	.

7.3.2 Gradient Boosting

The Gradient Boosting node searches for the optimal partition of the data for the single target variable. Here we can see that the fit characteristics are improved on boosting.

Fit Statistics

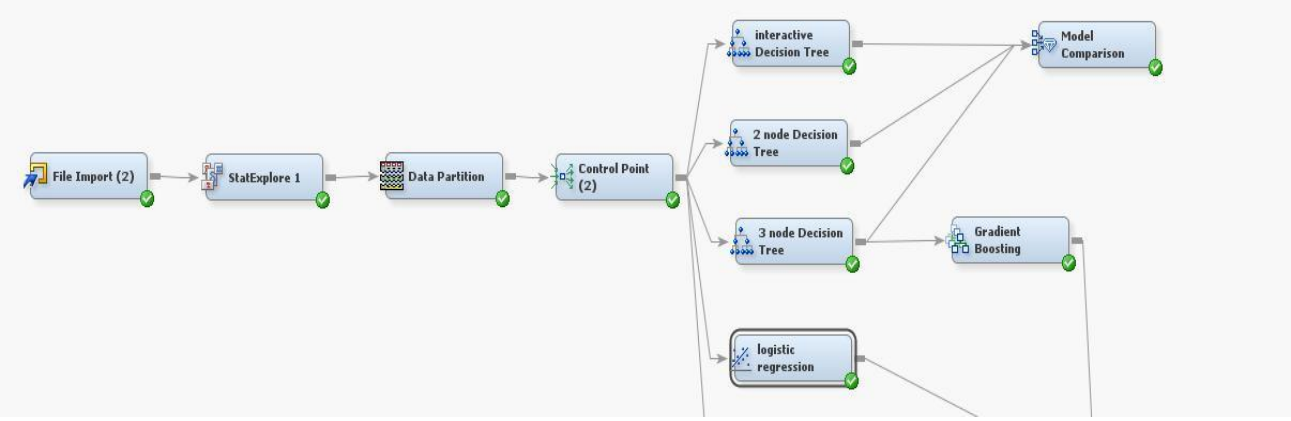
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
CIT_INJURED	CIT_INJURED	NOBS	Sum of Frequ...	1344	899	.
CIT_INJURED	CIT_INJURED	SUMW	Sum of Case ...	2688	1798	.
CIT_INJURED	CIT_INJURED	MISC	Misclassificati...	0.249256	0.250278	.
CIT_INJURED	CIT_INJURED	MAX	Maximum Abs...	0.754914	0.754914	.
CIT_INJURED	CIT_INJURED	SSE	Sum of Squar...	502.7297	337.124	.
CIT_INJURED	CIT_INJURED	ASE	Average Squa...	0.187027	0.187499	.
CIT_INJURED	CIT_INJURED	RASE	Root Average ...	0.432467	0.433012	.
CIT_INJURED	CIT_INJURED	DIV	Divisor for ASE	2688	1798	.
CIT_INJURED	CIT_INJURED	DFT	Total Degrees...	1344	.	.

Misclassification rate: 0.2492 for Train and 0.2502 for validation
Average Squared error: 0.187 for train and 0.187 for validation
This shows that the gradient boosting has improved the output of decision tree.

7.3.3 Logistic Regression

Regression is used to predict a variable from other variables. Since the output variable here is binary, logistic type of regression is implemented. The input is same as that of the decision tree. We are using Stepwise model for regression.

Property	Value
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Uses Defaults	Yes
Options	
Output Options	
Confidence Limits	No
Save Covariance	No
Covariance	No
Correlation	No
Statistics	No
Suppress Output	No
Details	No
Design Matrix	No
Score	
Excluded Variables	Reject



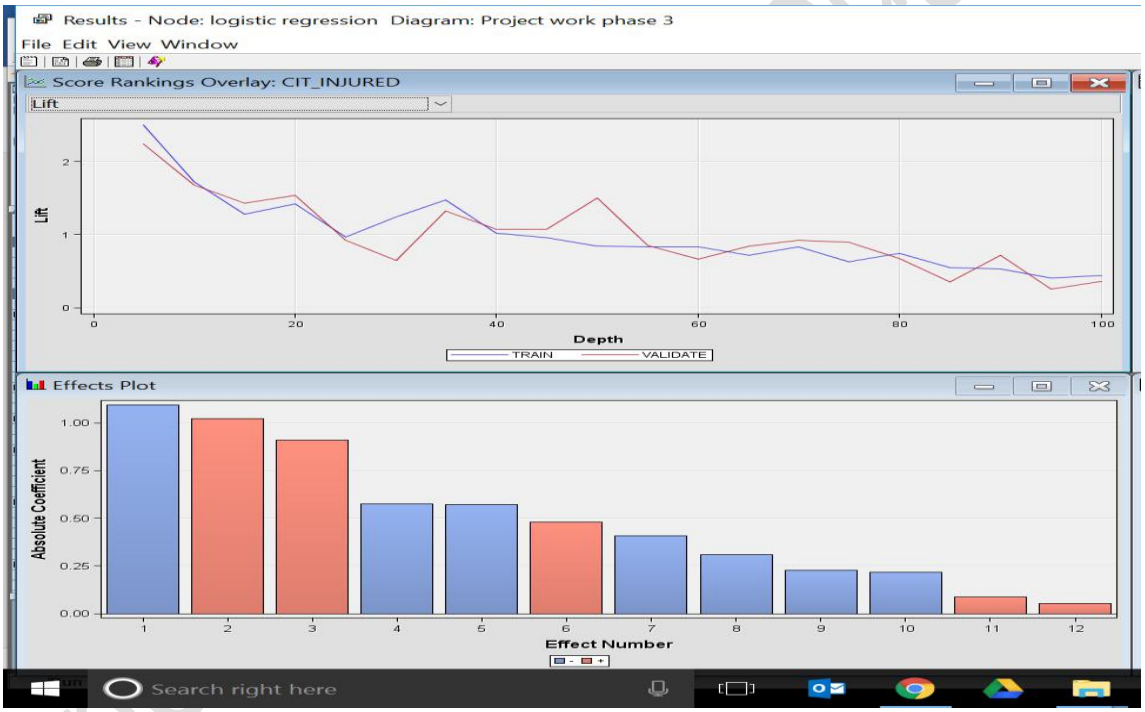
a. **Result:**

On running results, the selected model is the model trained in the last step (Step 4).

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	OFF_INJURED	1	1	44.5393		<.0001
2	CIT_INFL_ASSMT	7	2	38.2746		<.0001
3	ForceEffective	2	3	11.4995		0.0032
4	CitSex	1	4	6.3071		0.0120

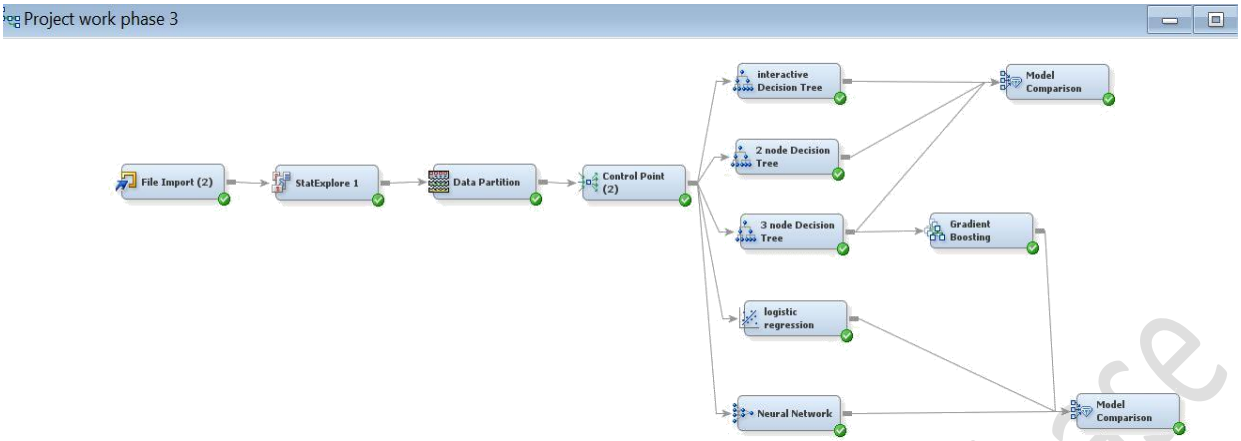
Score Ranking Overlay Statistics:



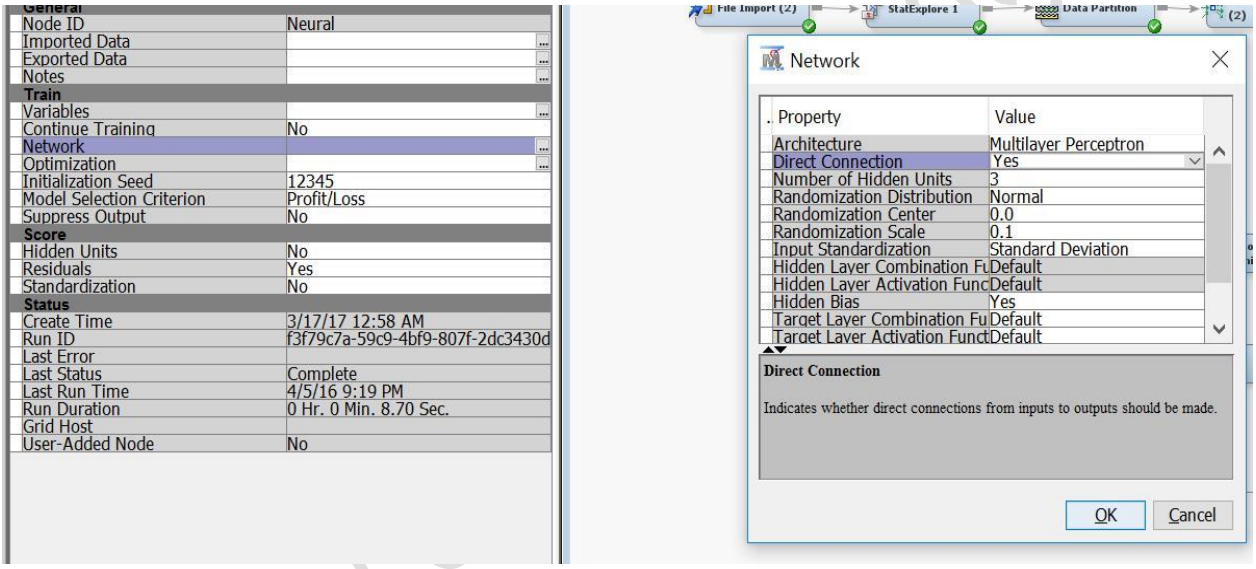
The score ranking overlay characteristics are good for the Logistic regression.

7.3.4 NEURAL NETWORK

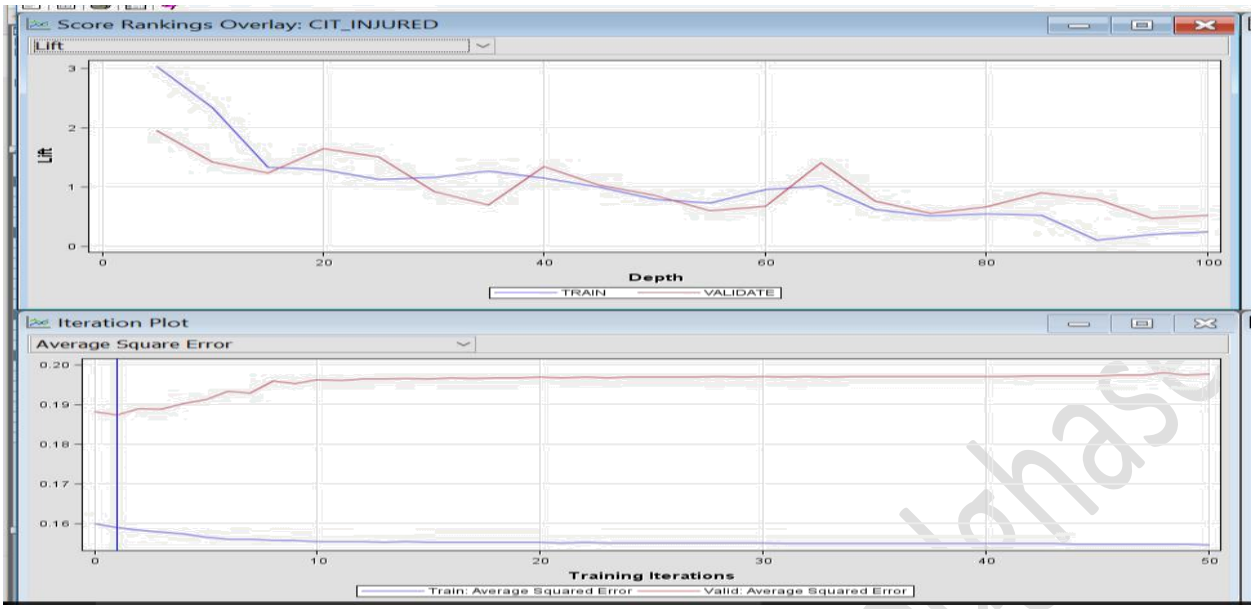
Neural networks are the type of models which accommodate nonlinear relationships between predictors and target variable than logistic regression. Since we have a good knowledge of the structure of the model that we're trying to define, we used the neural network node here.



Here, direct connection is made from input to the output node and we are specifying number of hidden units as 3 as seen in the screenshot.



Results of the Neural Network:



7.3.5 Model Comparison:

Models	Avg Square Error	Misclassification Rate
Neural Network	0.1739	0.2680
Logical Regression	0.1531	0.2436
Decision Tree with Gradient boosting	0.1870	0.2502

On comparing the models based on the score ranking overlay, Avg Square Error and misclassification rate comparison, we conclude that the Logistic Regression model is the best fit.

7.3.6 Time Series Analysis:

As Time Series Analysis is used to analyze past data and predict the trend for future. So, we exploited this node to analyze the data for the entire year on a week by week basis and plot it graphically to predict the data for the coming weeks. Since time series analysis is performed on interval data, we used inputs as mentioned.

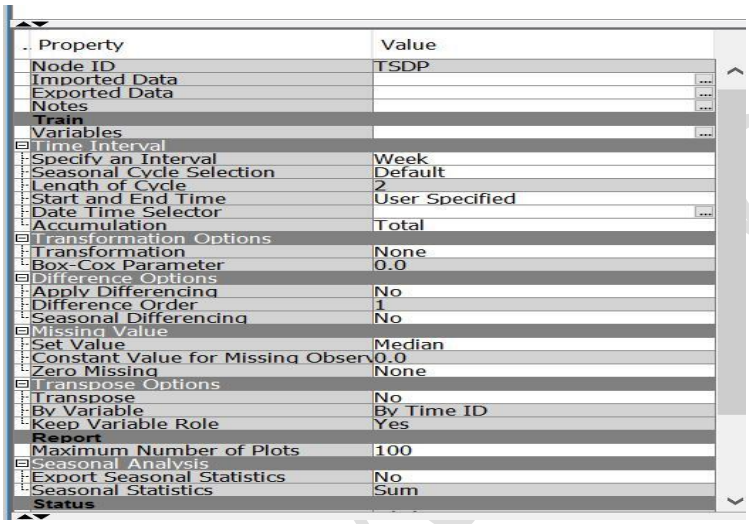
The graphs and report are as follows:

a. **Variable Selection:** In our case the Target Variable is 'whether the citizen got arrested or not'. Hence, our prediction for the first 2 weeks of the next year tells about the fact that whether the *citizen will get arrested or not*.

Variable	Role	Level
CURRENT_BADGE_NO	INPUT	INTERVAL
Coverted_OCCURRED_TM	INPUT	NOMINAL
OR_CODE	INPUT	INTERVAL
RA	INPUT	INTERVAL
Osex_Code	INPUT	INTERVAL
Force_Eff_code	INPUT	INTERVAL
C_Sex_Code	INPUT	INTERVAL
HIRE_DT	INPUT	INTERVAL
C_Infl_Ass_Code	INPUT	INTERVAL
C_Race_Code	INPUT	INTERVAL
CitNum	INPUT	INTERVAL
BEAT	INPUT	INTERVAL
UOF_Code	INPUT	INTERVAL
SECTOR	INPUT	INTERVAL
Cit_Chrg_Type_Code	INPUT	INTERVAL
Years_Service	INPUT	INTERVAL
OffCondType	REJECTED	NOMINAL
OffRace	REJECTED	NOMINAL
ID	REJECTED	NOMINAL
OCCURRED_TM	INPUT	INTERVAL
UOFNum	REJECTED	NOMINAL
OFF_INJURED	REJECTED	NOMINAL
SERVICE_TYPE	REJECTED	NOMINAL
OFF_HOSPITAL	REJECTED	NOMINAL
OffSex	REJECTED	NOMINAL
UOF_REASON	REJECTED	NOMINAL
CIT_ARRESTED	REJECTED	NOMINAL
CIT_INFL_ASSMT	REJECTED	NOMINAL
CitSex	REJECTED	NOMINAL
Cycles_Num	REJECTED	NOMINAL
CIT_INJURED	REJECTED	NOMINAL
CitChargeType	REJECTED	NOMINAL
ARC_Street	REJECTED	NOMINAL

CitRace	REJECTED	NOMINAL
CitCondType	REJECTED	NOMINAL
ForceEffective	REJECTED	NOMINAL
FILENUM	REJECTED	NOMINAL
GeoLocation	REJECTED	NOMINAL
ForceType	REJECTED	NOMINAL
DIST_NAME	REJECTED	NOMINAL
DIVISION	REJECTED	NOMINAL
C_ARR_CODE	TARGET	INTERVAL
OCCURRED_DT	TIMEID	INTERVAL

b. TS Data Preparation: Here we're preprocessing the data to make it suitable for the time series analysis. So, we are setting user specified start and end time and Time Interval as 'week'. We're replacing missing values by Median value.



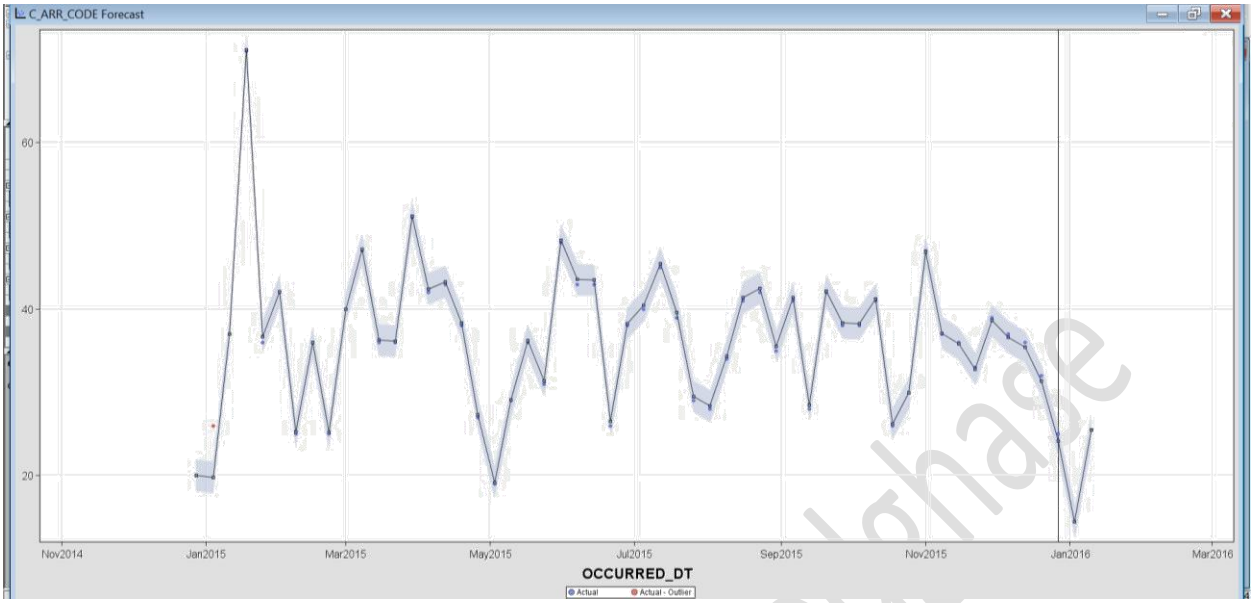
Property	Value
Node ID	TSDP
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Time Interval	
Specify an Interval	Week
Seasonal Cycle Selection	Default
Length of Cycle	2
Start and End Time	User Specified
Date Time Selector	
Accumulation	Total
Transformation Options	
Transformation	None
Box-Cox Parameter	0.0
Difference Options	
Apply Differencing	No
Difference Order	1
Seasonal Differencing	No
Missing Value	
Set Value	Median
Constant Value for Missing Obser	0.0
Zero Missing	None
Transpose Options	
Transpose	No
By Variable	By Time ID
Keep Variable Role	Yes
Report	
Maximum Number of Plots	100
Seasonal Analysis	
Export Seasonal Statistics	No
Seasonal Statistics	Sum
Status	

c. TS Exponential Smoothing: We are performing TS Exponential Smoothing to give relatively more weight to recent observations in forecasting than the older observations.

d. Result:

Highest number of arrest occur during the spring season.

Lately, we again see a rise in the rate of arrest during the holiday season and a dip in winters.



7.3.7 Clustering 1

We have used Automatic clustering with 10 clusters and with maximum 5 number of clusters which will be acceptable for final solution.

General	
Node ID	Clus
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Cluster Variable Role	Segment
Internal Standardization	Standardization
Number of Clusters	
Specification Method	Automatic
Maximum Number of Cluster	10
Selection Criterion	
Clustering Method	Ward
Preliminary Maximum	50
Minimum	2
Final Maximum	5
CCC Cutoff	3
Encoding of Class Variables	
Ordinal Encoding	Rank
Nominal Encoding	GLM
Initial Cluster Seeds	
Seed Initialization Method	Default
Minimum Radius	0.0
Drift During Training	No
Training Options	
Use Defaults	Yes
Settings	...
Missing Values	
Interval Variables	Default
Nominal Variables	Default
Ordinal Variables	Default
Scoring Imputation Method	None
Score	
Cluster Variable Role	Segment
Hide Original Variables	Yes
Cluster Label Editor	...
Report	
Cluster Graphs	Yes
Tree Profile	Yes
Distance Plot and Table	Yes

a. StatExplorer:

On analyzing the output of StatExplorer, we see that there are missing values so we need to remove those values.

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	NOMINAL	2
REJECTED	INTERVAL	18
REJECTED	NOMINAL	23

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

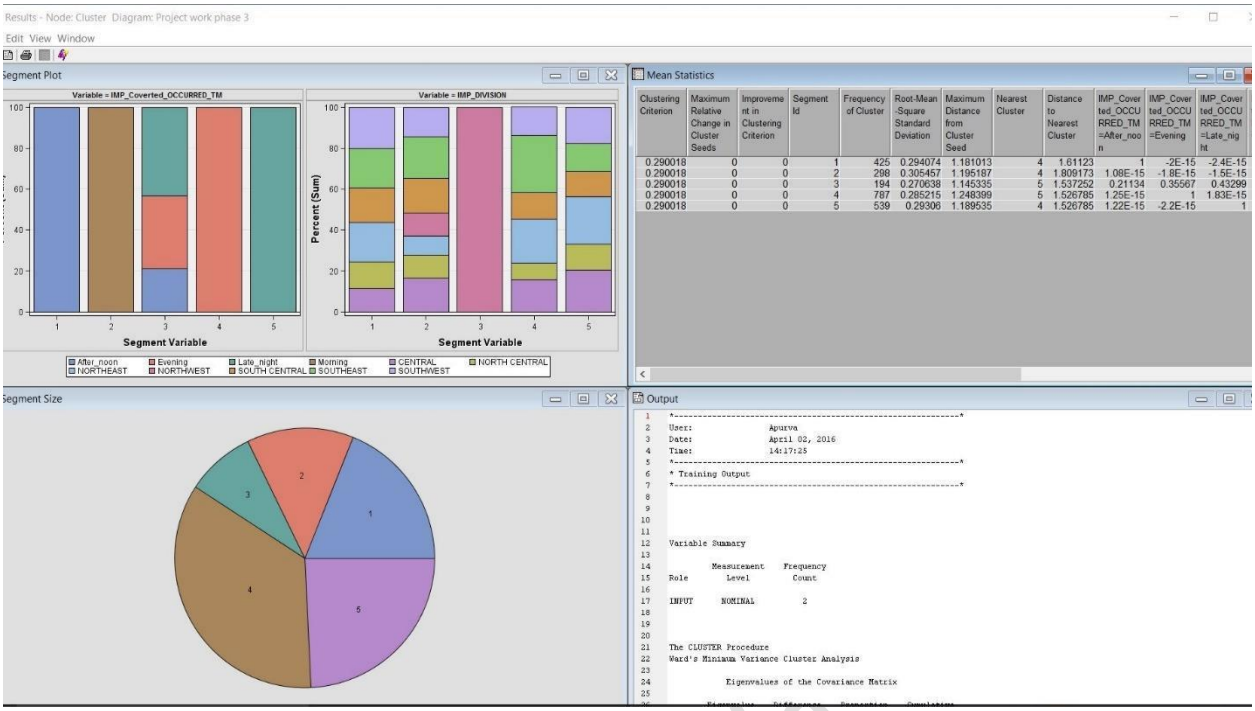
Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Coverted_OCCURRED_TM	INPUT	5	7	Evening	37.90	Late_night	27.73
TRAIN	DIVISION	INPUT	8	1	SOUTHEAST	19.44	NORTHEAST	18.06

b. Impute Node

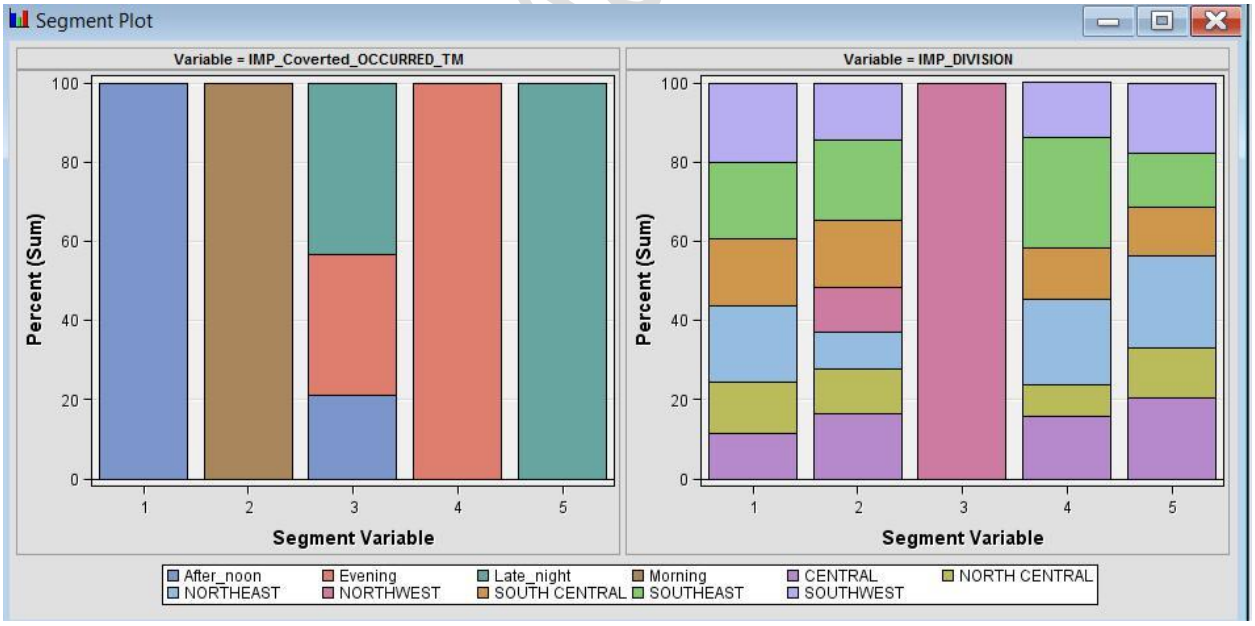
Since there were few missing values for occurrence time and division. Before performing clustering, we used impute node for treating missing values. The default input method that we used was Tree surrogate since time was numerical variable.

Property	Value
General	
Node ID	Impt
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Tree Surrogate
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Mean
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	.
Method Options	
Random Seed	12345
Tuning Parameters	
Tree Imputation	
Score	
Hide Original Variables	Yes
Indicator Variables	
Type	None
Source	Imputed Variables
Role	Rejected
Report	
Validation and Test Data	No
Distribution of missing	No
Status	
Create Time	4/2/16 2:16 PM
Run ID	47f0e918-ad74-404b-bbec-b0a
Last Error	
Last Status	Complete
Last Run Time	4/2/16 2:17 PM
Run Duration	0 Hr. 0 Min. 3.60 Sec.
Grid Host	
User-Added Node	No

c. Results:



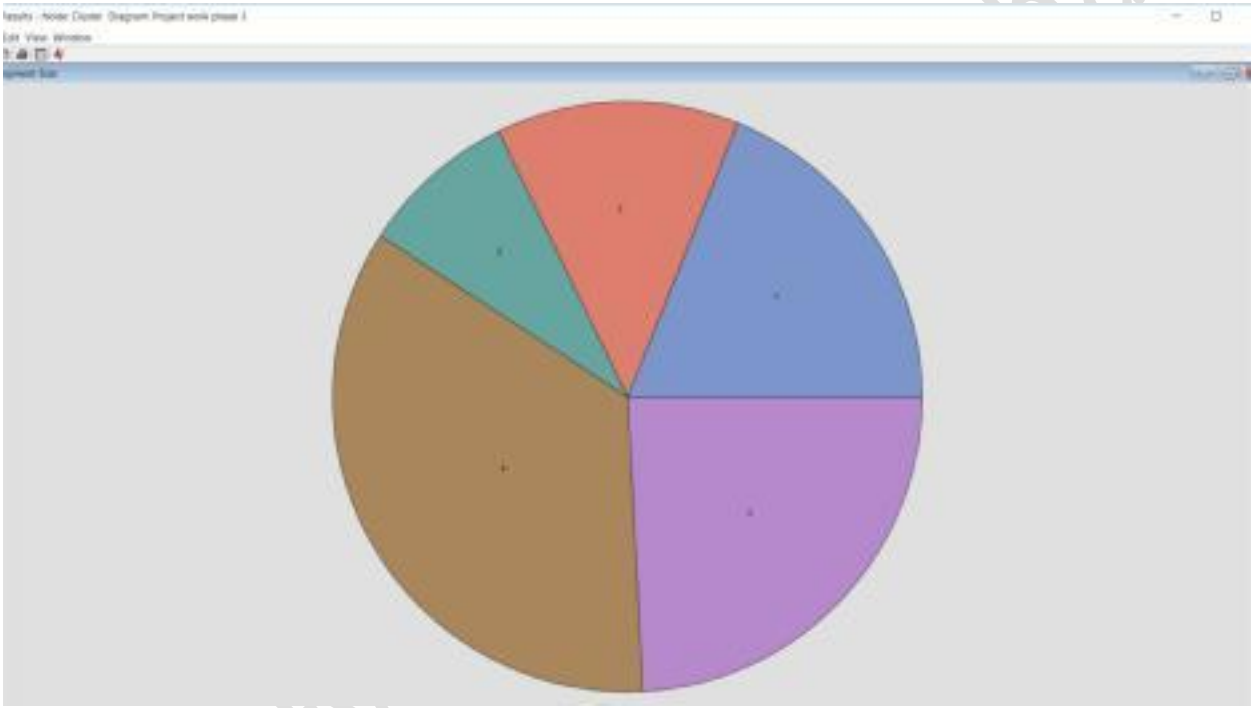
d. Segment Plot:



Clusters 1,2,4 and 5 are pure cluster with the occurrence time variable. (we had different time within our dataset we binned the times into 4 categoies of morning, afternoon, evening and late night). Cluster 1 depicts the after-noon time which is from 12 PM to 6 PM. Cluster 2 depicts the morning time which is form 6 AM to 12 PM. Cluster 4 depicts the evening time which is form 6PM to 12 AM. Cluster 5 depicts the late-night time which is from 12 AM to 6 AM. Obviously, cluster 3 is combination of different times.

On the other side, we have cluster 3 as a pure cluster which depicts NORTH WEST division. While other clusters are combination of different divisions such as NORTH EAST, SOUTH CENTRAL, SOUTH EAST, SOUTH WEST, and NORTH CENTRAL.

e. Segment size:



The clusters 1 and 4 and 5 have highly significant observation and clusters 2 and 3 have almost significant observations.

Statistics:

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	IMP_Covered_OCCURRED_TM=After_noon	IMP_Covered_OCCURRED_TM=Evening	IMP_Covered_OCCURRED_TM=Late_night	IMP_Covered_OCCURRED_TM=Morning	IMP_DIVISION=CENTRAL	IMP_DIVISION=NORTH	IMP_DIVISION=NORTHEAST	IMP_DIVISION=NORTHWEST	IMP_DIVISION=SOUTH CENTRAL	IMP_DIVISION=SOUTHEAST	IMP_DIVISION=SOUTHWEST
0.290018	0	0	1	425	0.294074	1.181013	4	1.61123	1	-2E-15	-2.4E-15	-9.2E-16	0.115294	0.129412	0.192941	2.78E-16	0.167059	0.195294	0.2
0.290018	0	0	2	298	0.305457	1.195187	4	1.809173	1.08E-15	-1.8E-15	-1.5E-15	1	0.16443	0.114094	0.090604	0.114094	0.171141	0.201342	0.144295
0.290018	0	0	3	194	0.270638	1.145335	5	1.537252	0.21134	0.35567	0.43299	1.11E-16	3.33E-16	4.16E-16	2.78E-17	1	2.78E-16	2.78E-17	5.27E-16
0.290018	0	0	4	787	0.285215	1.248399	5	1.526785	1.25E-15	1	1.83E-15	-1.3E-15	0.15629	0.080051	0.21601	2.78E-16	0.130877	0.279543	0.13723
0.290018	0	0	5	539	0.29306	1.189535	4	1.526785	1.22E-15	-2.2E-15	1	-1.1E-15	0.205937	0.124304	0.233766	2.78E-16	0.122449	0.137291	0.176252

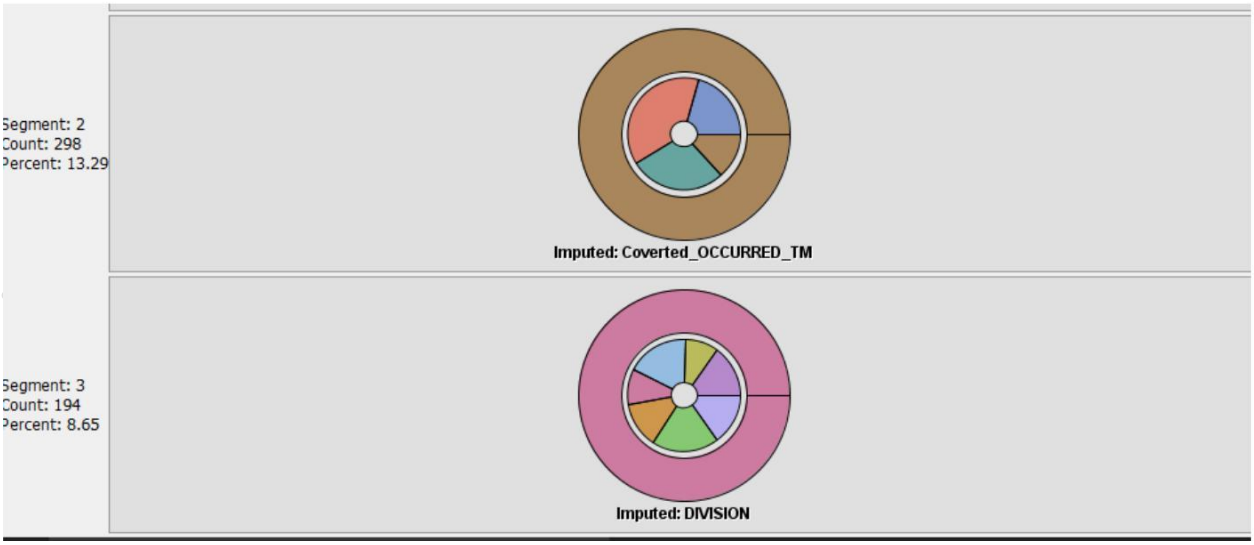
From the results it can be concluded that:

The probability of interactions between police and citizens during evening is higher in SOUTHEAST.

The probability of interaction between police and citizens in the NORTHEAST have a higher rate in late night.

f. Segment Profile:

We examined the segmented/clustered data and identified the factors that diiferenitate data segments from the populations.



7.3.8 Clustering 2

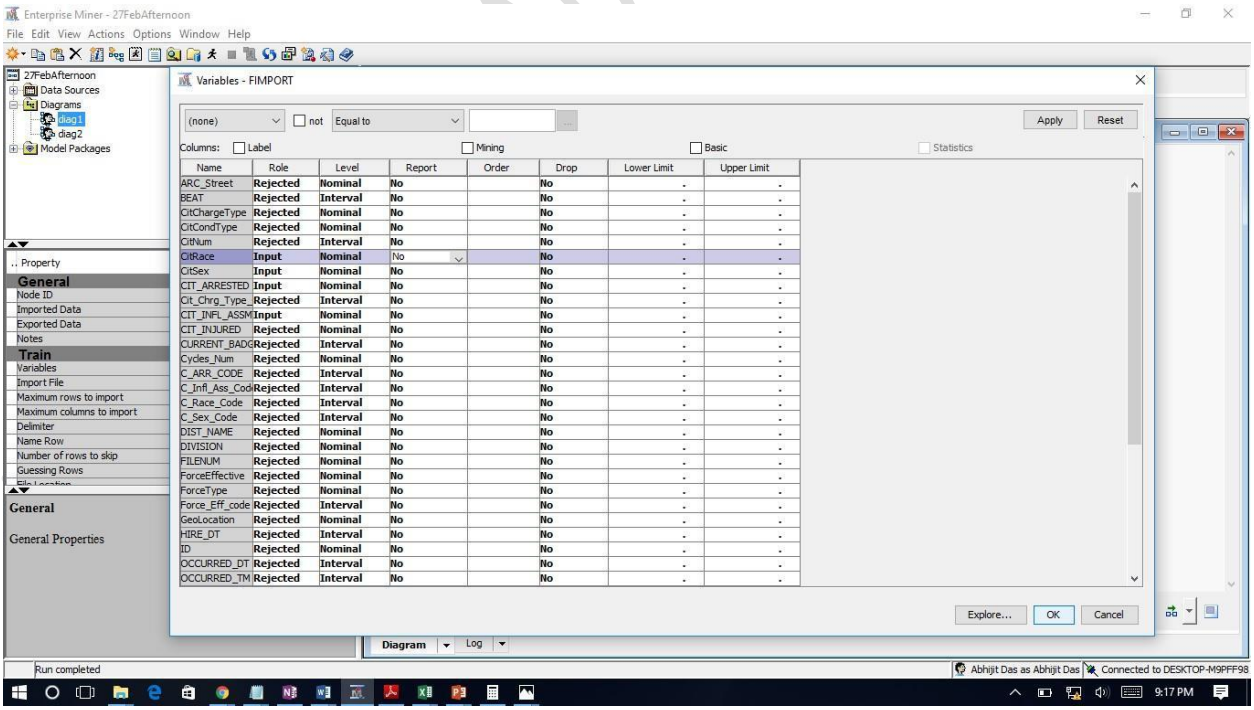
Clustering is an unsupervised learning method, which groups a set of object based on their similarity.

We have tasked clustering with grouping Citizen Sex, Citizen state assessment at the time of encounter, Citizen arrest and Citizen race to find out common traits within these attributes that lead to a citizen being arrested. Using these we can understand the general trend and effectively predict the risk behaviors.

As this is an unsupervised analysis technique we have set all the parameters to the input.

a. File import node:

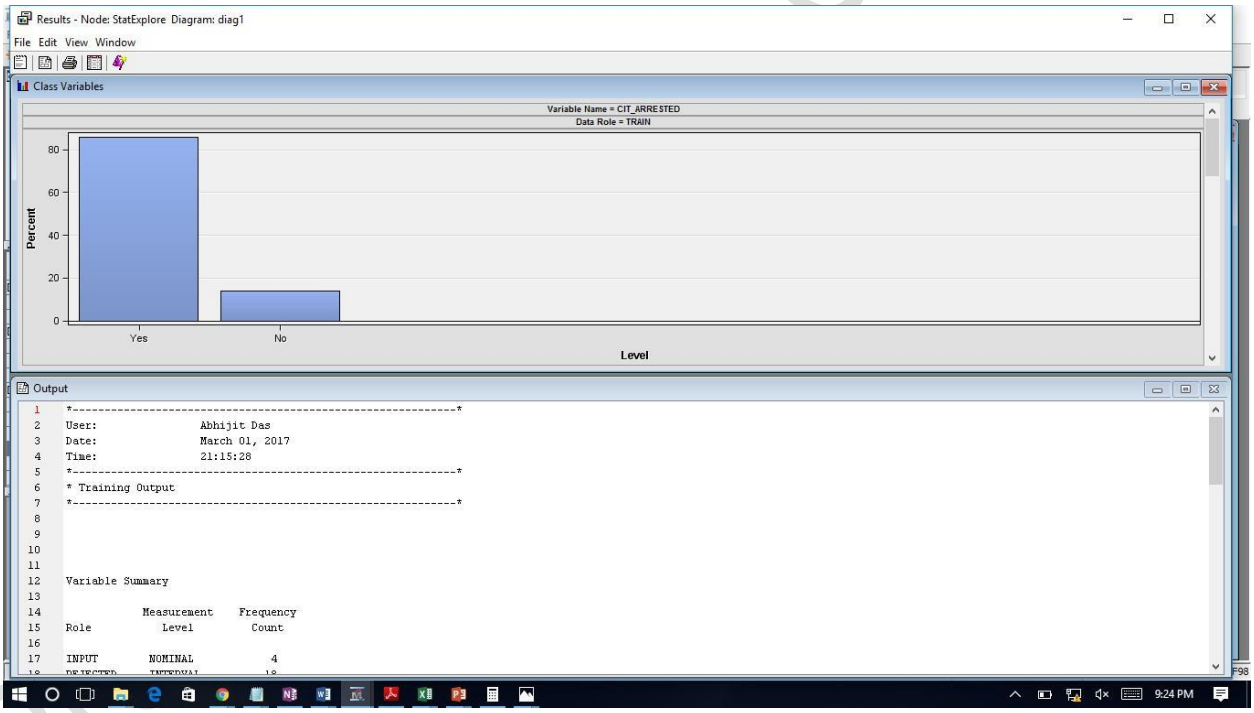
Input Variables: CitSex, CitRace, CIT_ARRESTED, CIT_INFL_ASSMT



The general description of the data set:

```
25    The CONTENTS Procedure
26
27    Data Set Name        EMWS1.FIMPORT_DATA          Observations          2243
28    Member Type          DATA                        Variables             42
29    Engine               V9                          Indexes               0
30    Created              03/01/2017 17:36:38         Observation Length    688
31    Last Modified        03/01/2017 17:36:38         Deleted Observations  0
32    Protection                                       Compressed            NO
33    Data Set Type                                    Sorted                NO
34    Label
35    Data Representation  WINDOWS_64
36    Encoding             wlatin1  Western (Windows)
```

b. StatExplore node: Reported the summary and association summary



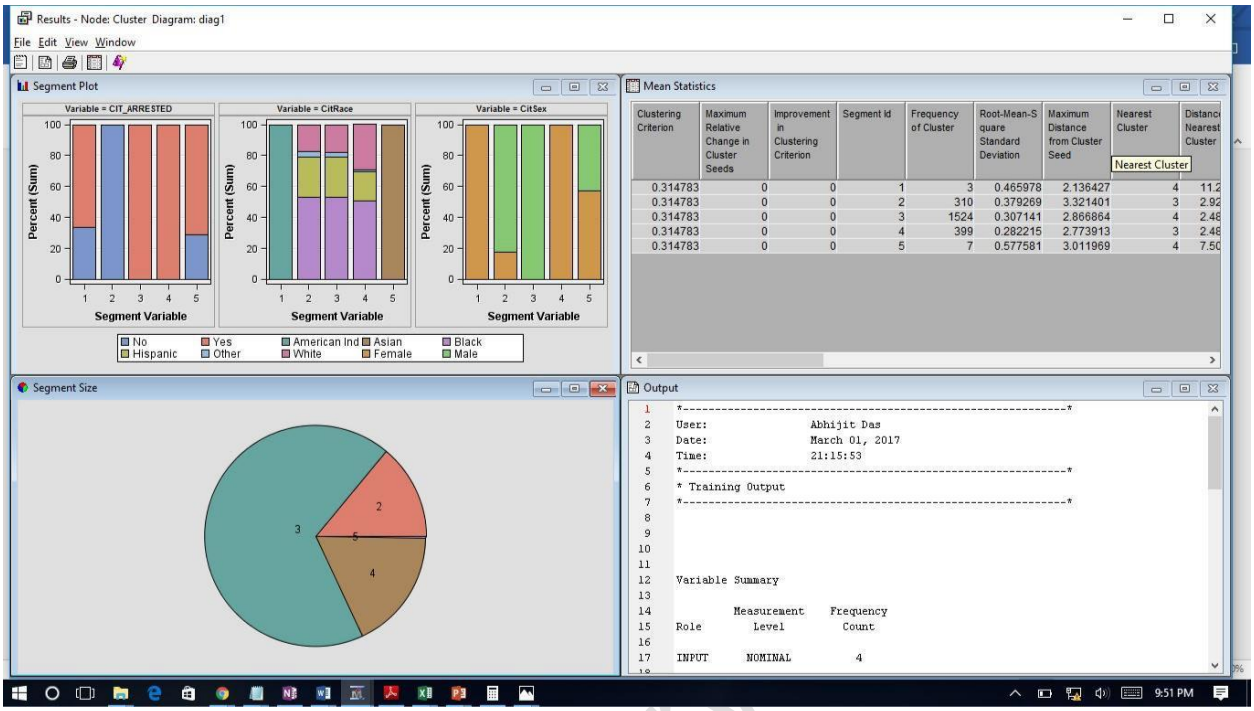
Accordingly, almost 80% of the interaction reported culminated in arrest of the citizen.

c. Control point: Incorporated the control point for addition of analysis points in future without changing the existing the current set up.

d. Cluster node: We have used user defined clustering with 5 clusters. Encoding method of Nominal variables: GLM (Generalized Linear Modelling)

Train	
Variables	
Cluster Variable Role	Segment
Internal Standardization	Standardization
Number of Clusters	
Specification Method	User Specify
Maximum Number of Clusters	5
Selection Criterion	
Clustering Method	Ward
Preliminary Maximum	50
Minimum	2
Final Maximum	20
CCC Cutoff	3
Encoding of Class Variables	
Ordinal Encoding	Rank
Nominal Encoding	GLM
Initial Cluster Seeds	
Seed Initialization Method	Default
Minimum Radius	0.0
Drift During Training	No
Training Options	
Use Defaults	Yes
Settings	
Missing Values	
Interval Variables	Default
Nominal Variables	Default
Ordinal Variables	Default
Scoring Imputation Method	None
Score	
Cluster Variable Role	Segment
Hide Original Variables	Yes

Running the cluster node produced the following results:



e. Segment plot:

Cluster 2,3,4 are pure cluster with the arrested variable. While the cluster 3, 4 depict the characteristics which associate with citizen getting arrested, the cluster 2 represents the characteristics associated with the case the citizen was not arrested.

Cluster 1 and 4 are entirely comprised of female and cluster 2 is a pure cluster of male. We can decompose the characteristics on a gender basis using this information.

The clustering of the Race is a heterogenous in the desired clusters of 2,3,4.

f. Segment Size:

The cluster 2,3,4 have significant observation, while the segment 1 and 5 do not have significant number of values.

Statistics:

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	CIT_ARRESTED=No	CIT_ARRESTED=Yes	CIT_INFL_ASSMT=Alchohol
0.314783	0	0	1	3	0.465978	2.136427	4	11.23372	0.333333	0.666667	0.333333
0.314783	Maximum Relative Change in Cluster Seeds			310	0.379269	3.321401	3	2.927823	1	6.11E-15	0.096774
0.314783	0	0	3	1524	0.307141	2.866864	4	2.484345	-2.4E-15	1	0.175853
0.314783	0	0	4	399	0.282215	2.773913	3	2.484345	-1.3E-15	1	0.205514
0.314783	0	0	5	7	0.577581	3.011969	4	7.504933	0.285714	0.714286	-2.8E-17

CIT_INFL_ASSMT=Alchohol and unknown drugs	CIT_INFL_ASSMT=FD	CIT_INFL_ASSMT=Marijuana	CIT_INFL_ASSMT=Mentally unstable	CIT_INFL_ASSMT=None detected	CIT_INFL_ASSMT=Unknown	CIT_INFL_ASSMT=Unknown Drugs	CitRace=American Ind	CitRace=Asian	CitRace=Black	CitRace=Hispanic	CitRace=Other	CitRace=White	CitSex=Female	CitSex=Male
0.333333	1.39E-17	0	0	0.333333	0	-1.4E-17	1	-8.7E-19	0	-2.8E-17	0	0	1	1.11E-16
0.029032	0.106452	0.025806	0.070968	0.277419	0.306452	0.087097	9.97E-18	-1.7E-17	0.532258	0.258065	0.035484	0.174194	0.177419	0.822581
0.107612	0.09252	0.020997	0.137139	0.114173	0.223753	0.127953	-3.1E-17	-3.9E-17	0.529528	0.261811	0.026903	0.181759	-3.3E-15	1
0.095238	0.0401	-1.1E-16	0.275689	0.115288	0.170426	0.097744	1.21E-17	-2.3E-17	0.503759	0.190476	0.010025	0.295739	1	-5.2E-15
-1.4E-17	0.142857	0.285714	0.428571	0.142857	0	-1.4E-17	0	1	0	2.78E-17	6.94E-18	0	0.571429	0.428571

From the results, it can be observed:

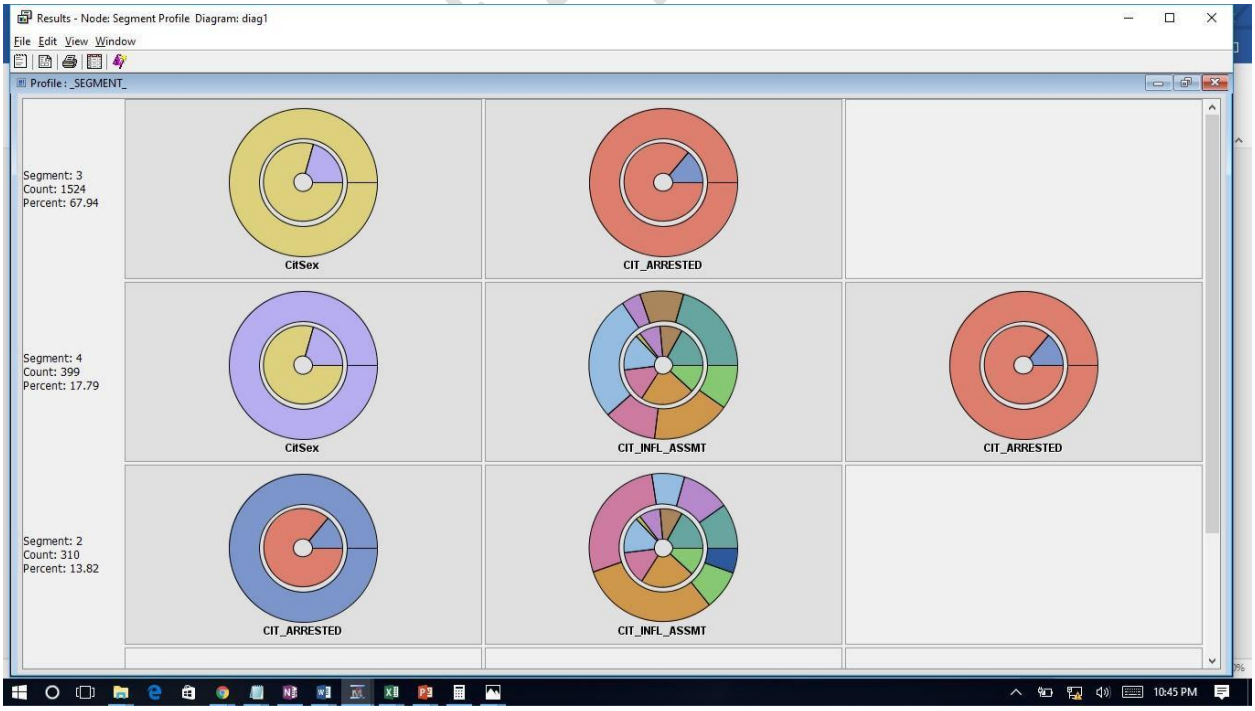
If the accuse is male, the probability of getting arrested during an encounter with police is higher.

Intoxication during the police interaction significantly increases the chance of the citizen getting arrested.

Further, an ethnic group(Black) has higher chances of getting arrested.

g. Segment Profile:

We examined the segmented/clustered data and identified the factors that differentiate data segments from the populations.



8. Summary of Findings

We can conclude the following interesting insights:

1. Almost 80% of the interactions between police and citizen will result in arrest of the citizen.
2. The Southeast area of the city and Northeast area of the city notch the highest number of interactions and hence arrests.
3. The Southeast is most susceptible area, 83% of the cases will result in arrests, followed by the Northeast region where, 79% of cases will result in arrest
4. Men have higher probability of being involved in interactions with cops- almost 77% of incidents recorded have a citizen whose gender is male.
5. For every injured citizen, there will be a 25% chance of the involved officer getting injured
6. From the model, it is predicted that for the first 10 days for the following year there will be a minimum of 35 incidents and a maximum of 43 incidents

9. Managerial Implications

As the Dallas police chief, more efforts should be focused on the Northeast and Southwest areas of the city. Increased patrolling during evening will make these areas safer.

A part of budget maybe allocated for special training to avoid possible injuries to officers during the interaction.

10. References

https://www.dallasopendata.com/Public-Safety/Police-2015-Response-to-Resistance/594v-2cnd
https://support.sas.com/documentation/onlinedoc/miner/
https://support.sas.com/documentation/onlinedoc/miner/casestudy_59123.pdf

Special Thanks to SAS Software YouTube Channel