Philadelphia Crime Analysis

Project Report

Group 14

Abhidha Arya

Neha Chaudhary

Table of Contents

Executive Summary

Crime rate in the US has varied over time and there is a lot of data pertaining to criminogenic factors such as time, place, and socio-demographics. Crime Analysis is the examination of relationships between such factors. To identify crime and be better prepared to handle the criminal activities, it is important to understand patterns in crime. Our project tries to predict crime type whether Violent or Non-Violent is more likely to occur given a month and place in Philadelphia. Victims are also of interest as it is believed that certain ethnicities and gender are more likely to be the target of a specific type of crime. We are interested in finding the factors that influence crimes and find out variables which are related to Violent and Non-Violent crime. We are forming our model based on approximately 6 years of data comprising of crime reports of Philadelphia City. The purpose of this project is to mine the Philadelphia crime data using the SAS Enterprise Miner and understand the crime pattern. We will use various prediction techniques such as decision tree, logistic regression, and neural networks to find out which are the relevant variables that affects the Philadelphia crime data. Based on results received after predictive modeling, we will target specific zones and focus on factors/variables which are important and that contributed most towards building the model. Given the result of model, Philadelphia Police Department can use these facts and analysis to decide upon the patrolling techniques.

Project motivation and background

According to Philadelphia Police Department US Census Bureau Philadelphia has a violent crime rate that is 227% higher than the Pennsylvania average and 176% higher than the national average. For Non-Violent crime, Philadelphia is 74% higher than the Pennsylvania average and 27% higher than the national average. In Philadelphia, you have 1 in 24 chance of becoming a victim of any crime and a 1 in 98 chance of becoming a victim of any violent crime. We are trying to find out factors that affect the crime which in turn will be useful in mitigating crime.

Data Description

The dataset used in our project for mining, is a second-hand data, which was obtained from an online source Kaggle.com. This data stores details of various crimes occurred in Philadelphia during the years 2011 – 2016. We are taking into consideration attributes such as Zip code, Gender, Offense Type, Zone/Beat and Ethnicity. The entire data is in the form of one .csv file. The file is attached below:

Crime_Final.csv

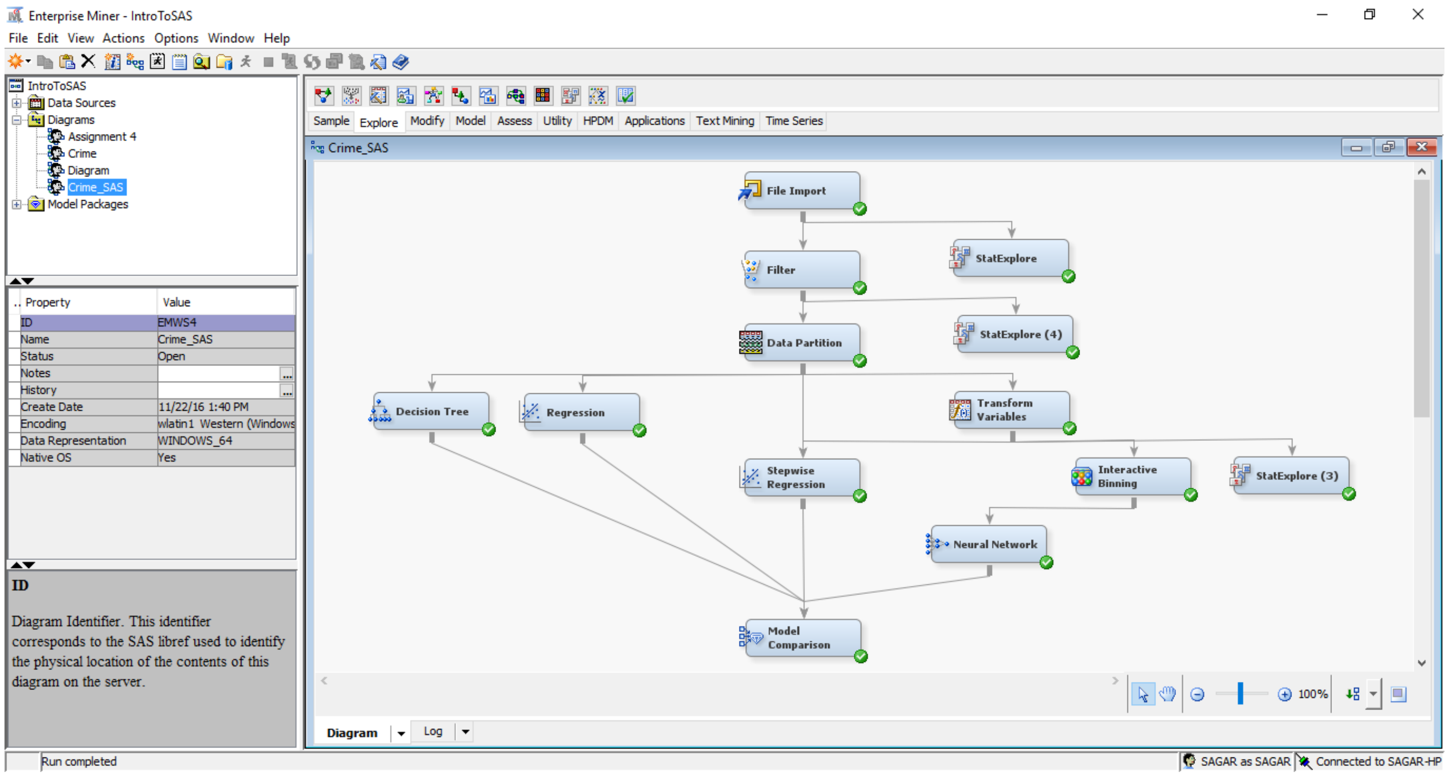
Variables Description

Given below the complete list of input variables along with description which we have taken in our analysis:

Variables	**Description**
Zip_Code	Zip code of the crime location.
Month	Month in which the crime has occurred.
Longitude	Longitude of the location of crime.
Location_Block	Address where the crime has occurred.
Latitude	Latitude of the location of crime.
Gender	Gender of the victim.
Dc_Dist	District code where the crime has occurred.
Zone_Beat	Zone in which crime has occurred.
Census_Tract_2000	Neighbourhood code.
Date_Reported	Date on which crime was reported.
General_Offense_Number	Unique Offense Number.
Hispanic_Non_Hispanic	Specifies whether the victim was hispanic or non-hispanic(1 or 0).
Hundred_Block_Location	Name of the location block within 100 yards.
Occurred Date or Date Range Start	Range of Start Date of Crime.
Occurred Date Range End	Range of End Date of Crime
Offense Code	Unique Offense Code.
Offense Code Extension	Extension to specify the exact offense in case of multiple categories of offense in the same offense code.
Offense_Type	Type of the offense based on the offense code.
Premise	Premise in which crime was committed.
RMS_CDW_ID	Unique identifier for each crime record.
Summarized_Offense_Description	Description of the offense based on the offense code.
Summarized_Offense_Code	Summarized offense codes based on similar crime types.
Violent	Depicts whether the crime is Violent or Non-Violent (1 or 0).

BI TECHNIQUES

Process flow Diagram



Group-14 Project Report

Step 1 - File Import

File Import node is used to import the data so that SAS Enterprise Miner can interpret the dataset. The data is imported in the form of .csv files

We have assigned our target variable and rejected the irrelevant variables for our analysis.

Variables - FIMPORT

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Census_Tract_2000	Input	Nominal	No		No	.	.
Date_Reported	Time ID	Interval	No		No	.	.
Dc_Dist	Input	Nominal	No		No	.	.
General_Offense_Number	Rejected	Interval	No		No	.	.
Hispanic_Non_Hispanic	Input	Binary	No		No	.	.
Hundred_Block_Location	Rejected	Nominal	No		No	.	.
Latitude	Input	Interval	No		No	.	.
Location	Rejected	Nominal	No		No	.	.
Longitude	Input	Interval	No		No	.	.
Month	Input	Interval	No		No	.	.
Occurred_Date_Range_End	Time ID	Interval	No		No	.	.
Occurred_Date_or_Date_Range_Star	Time ID	Interval	No		No	.	.
Offense_Code	Rejected	Nominal	No		No	.	.
Offense_Code_Extension	Rejected	Interval	No		No	.	.
Offense_Type	Rejected	Nominal	No		No	.	.
Premise	Input	Nominal	No		No	.	.
RMS_CDW_ID	Rejected	Interval	No		No	.	.
Summarized_Offense_Description	Rejected	Nominal	No		No	.	.
Summary_Offense_Code	Rejected	Nominal	No		No	.	.
Violent	Target	Binary	No		No	.	.
Year	Rejected	Interval	No		No	.	.
Zipcode	Input	Interval	No		No	.	.
Zone_Beat	Input	Nominal	No		No	.	.

Target Variable is- **Violent**

Violent is a **binary variable i.e. either 0(Non-Violent) or 1(Violent)**

Rejected Variables:

We have explicitly marked some variables as rejected due to below reasons-

1. Direct relation with the output: General_Offense_Number, Offense_Type, Summary_Offense_Code, Offense_Code_Extension, Summarized_Offense_Description, Offense_Code have been rejected as there is direct relation between these variables and the output variable.
2. Insignificant Variables: Location, Hundred_Block_Location have been rejected as Zipcode is already filtering data according to location. Also RMS_CDW_ID and Year have been rejected as these are insignificant.

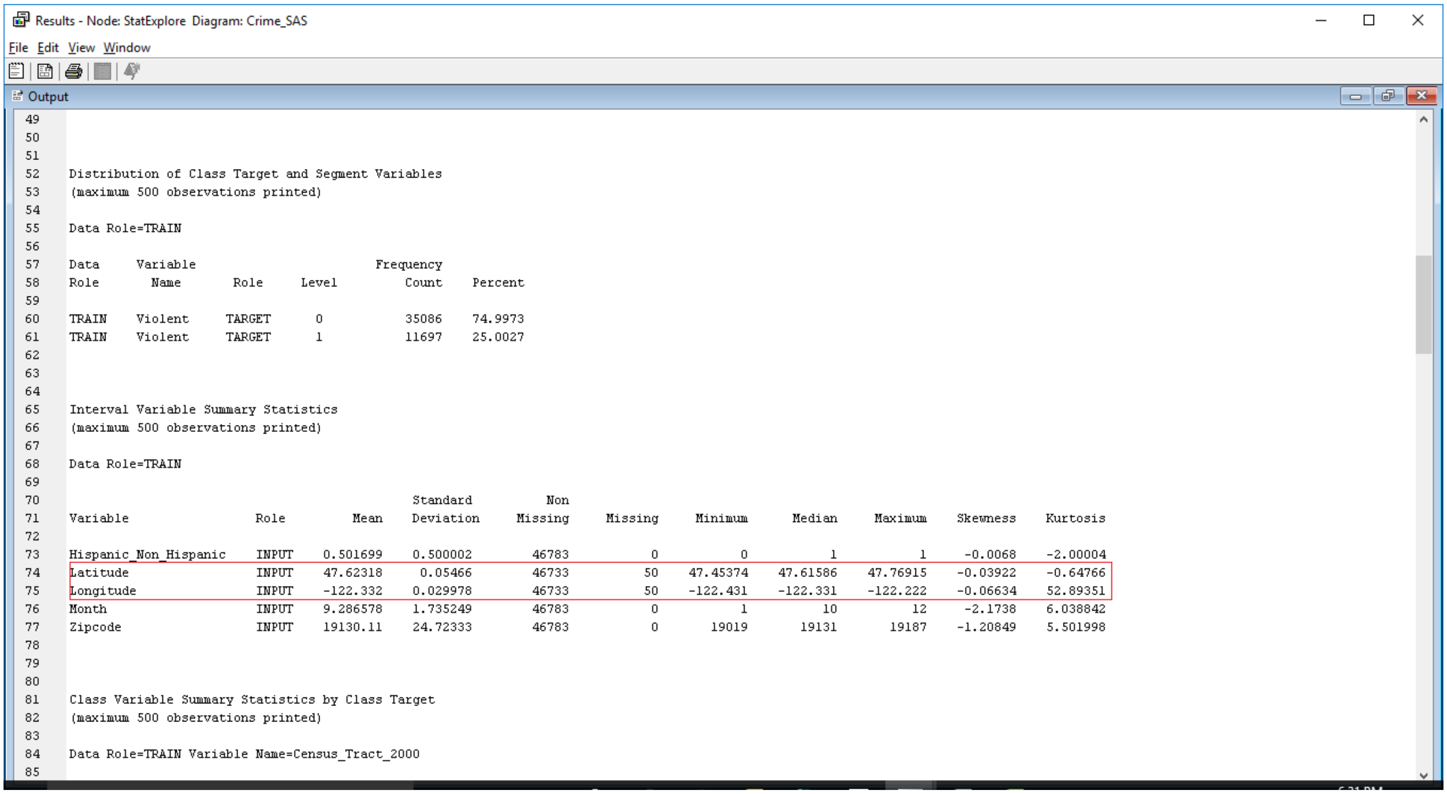
Violent or Non-Violent distribution as shown by below histogram:-

Out of 46783 observations, only 11697 observations **(25%)** are under violent category whereas 35086 observations **(75%)** are in Non-Violent Category.

Step 2 - Data Pre-Processing

We have used Stat Explore node to analyze the variable statistics. The below graph shows the relevance of input variables to the target variable "Violent".

We observed that there are missing values in the data set as shown by the stat explore node.


Results - Node: StatExplore Diagram: Crime_SAS

File Edit View Window

Output

Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	Violent	TARGET	0	35086	74.9973
TRAIN	Violent	TARGET	1	11697	25.0027

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Hispanic_Non_Hispanic	INPUT	0.501699	0.500002	46783	0	0	1	1	-0.0068	-2.00004
Latitude	INPUT	47.62318	0.05466	46733	50	47.45374	47.61586	47.76915	-0.03922	-0.64766
Longitude	INPUT	-122.332	0.029978	46733	50	-122.431	-122.331	-122.222	-0.06634	52.89351
Month	INPUT	9.286578	1.735249	46783	0	1	10	12	-2.1738	6.038842
Zipcode	INPUT	19130.11	24.72333	46783	0	19019	19131	19187	-1.20849	5.501998

Class Variable Summary Statistics by Class Target
(maximum 500 observations printed)

Data Role=TRAIN Variable Name=Census_Tract_2000

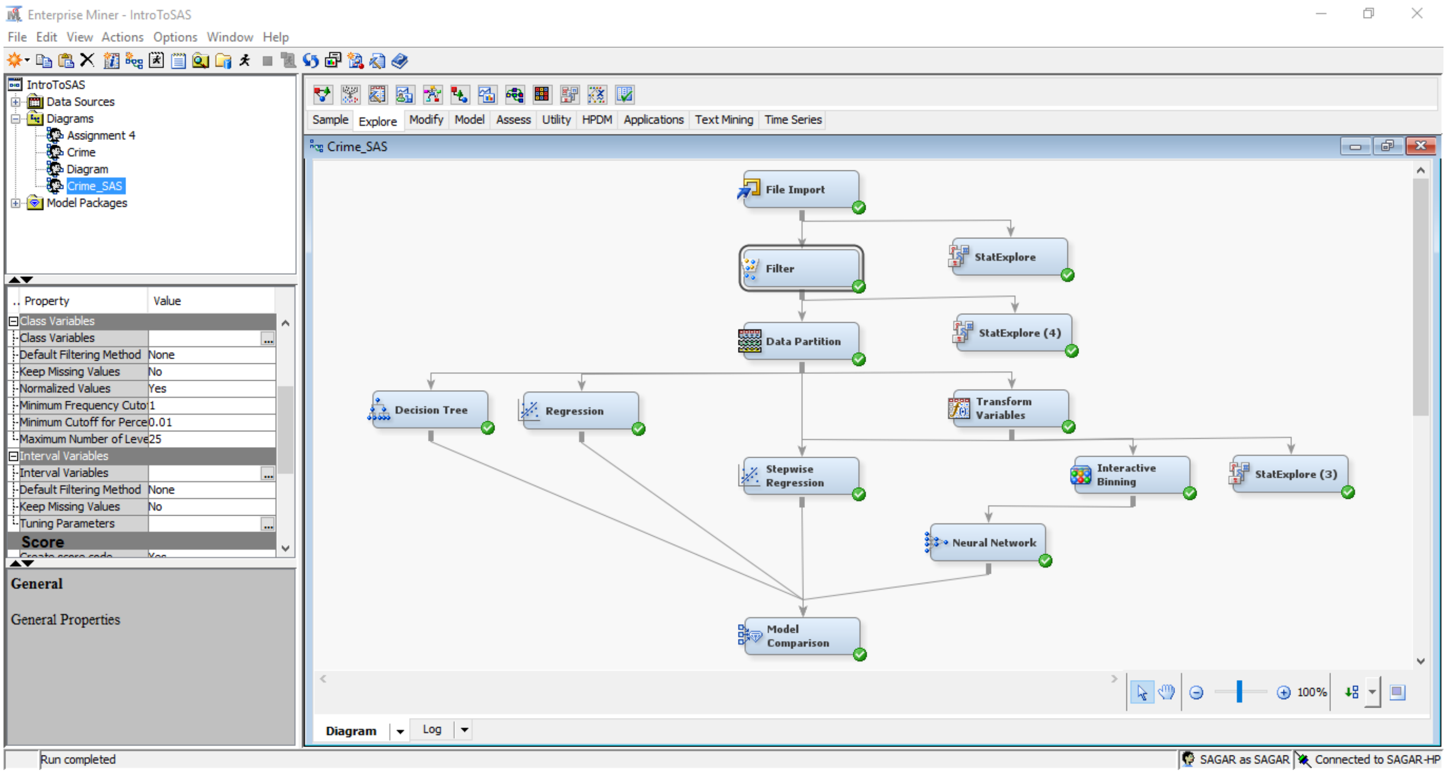
Variables which were having missing values are:

1. Latitude
2. Longitude

The above missing values were removed using the **Filter** node.

Step 2.1 Filter Node

Filter node excluded the missing values from the dataset. The default filtering method was kept as none and the property of filter node **Keep Missing Values** was kept as **No**.



Now again we have added a stat explorer node to check whether the missing values have been removed or not.

Results - Node: StatExplore (4) Diagram: Crime_SAS

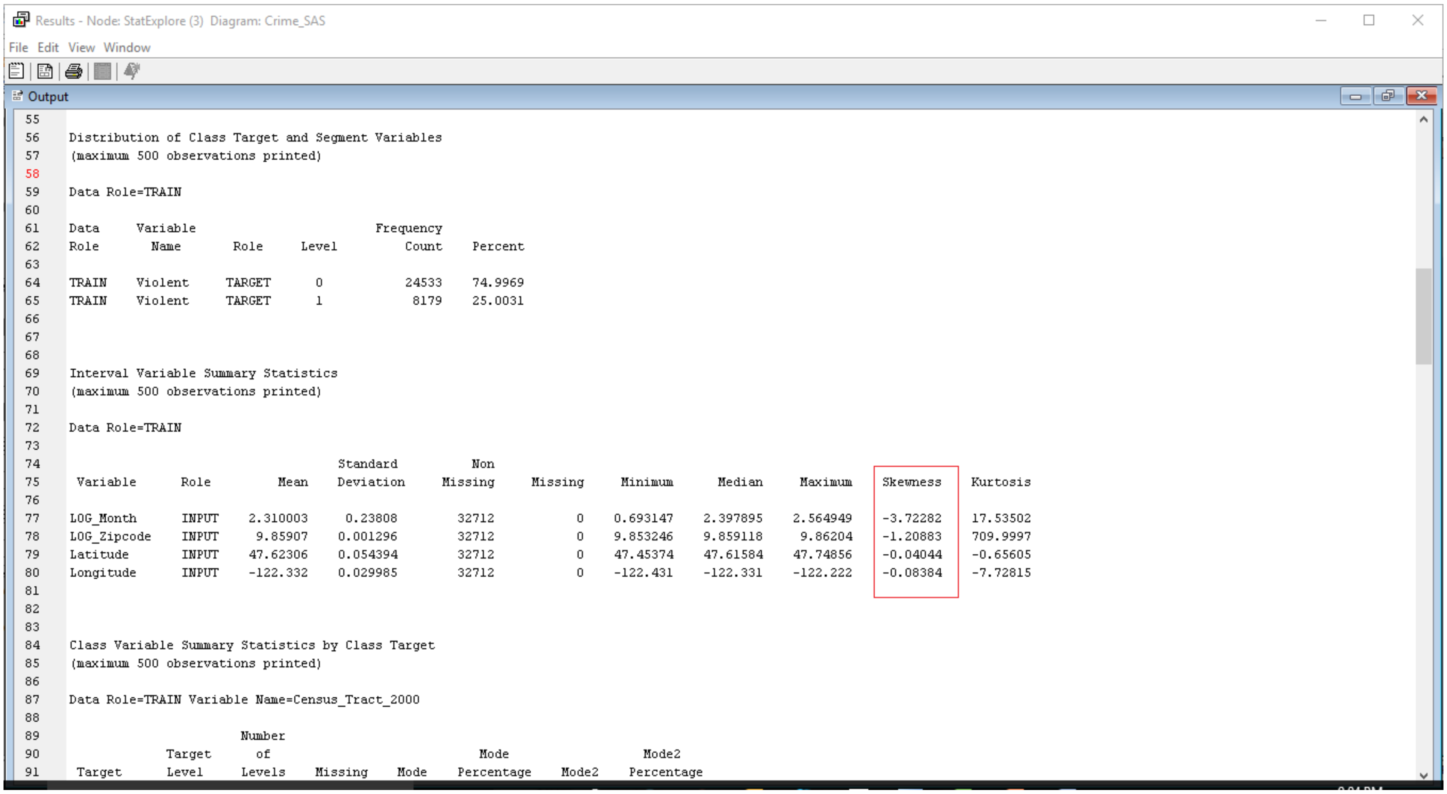
File Edit View Window

Output

```
Data Role=TRAIN

Data     Variable                         Frequency
Role       Name       Role     Level         Count     Percent

TRAIN    Violent     TARGET      0           35048     74.9963
TRAIN    Violent     TARGET      1           11685     25.0037



Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                                Standard          Non
Variable                  Role         Mean    Deviation      Missing     Missing     Minimum      Median     Maximum    Skewness    Kurtosis

Hispanic_Non_Hispanic    INPUT     0.501765     0.500002        46733           0           0           1           1    -0.00706    -2.00004
Latitude                 INPUT     47.62318      0.05466        46733           0    47.45374    47.61586    47.76915    -0.03922    -0.64766
Longitude                INPUT     -122.332     0.029978        46733           0    -122.431    -122.331    -122.222    -0.06634    52.89351
Month                    INPUT     9.284938     1.735222        46733           0           1          10          12    -2.17403    6.039133
Zipcode                  INPUT     19130.11     24.72228        46733           0       19019       19131       19187    -1.20871    5.502456



Class Variable Summary Statistics by Class Target
(maximum 500 observations printed)

Data Role=TRAIN Variable Name=Census_Tract_2000

                        Number
             Target      of                            Mode                  Mode2
 Target      Level     Levels    Missing    Mode    Percentage    Mode2    Percentage

Violent        0         164        0       8100       6.07        9300       2.90
```

Step 2.2 - Transform Variable

We are using transform variable as part of data preprocessing to reduce the skewness of the input variables. Transformation method has been set as Log for transforming the variables of the dataset.

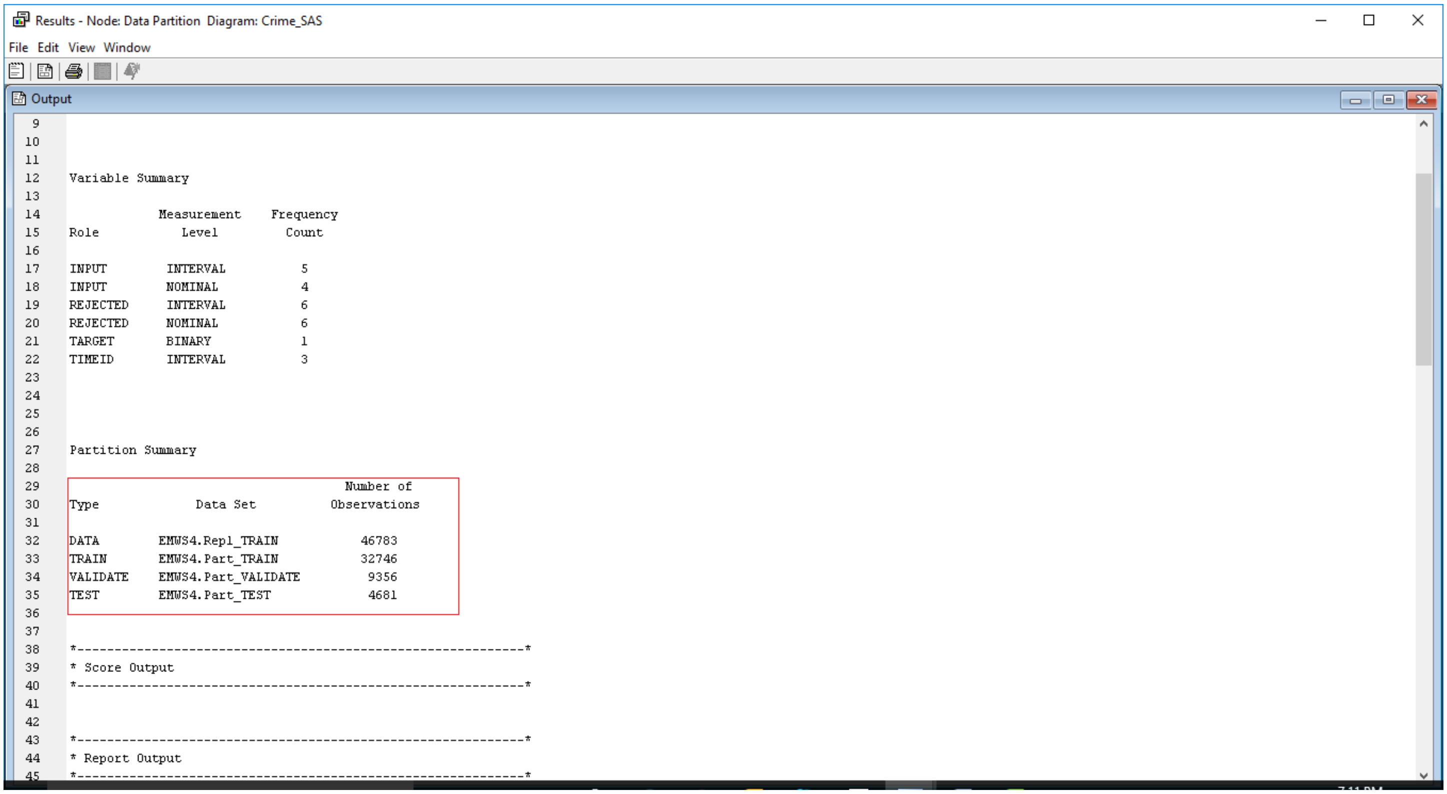



```
Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

Data Role=TRAIN

Data     Variable                          Frequency
Role       Name       Role     Level           Count     Percent

TRAIN    Violent     TARGET      0             24533     74.9969
TRAIN    Violent     TARGET      1              8179     25.0031



Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                         Standard         Non
 Variable      Role         Mean     Deviation     Missing     Missing     Minimum      Median     Maximum    Skewness    Kurtosis

LOG_Month      INPUT    2.310003       0.23808       32712           0    0.693147    2.397895    2.564949    -3.72282    17.53502
LOG_Zipcode    INPUT     9.85907      0.001296       32712           0    9.853246    9.859118     9.86204    -1.20883    709.9997
Latitude       INPUT    47.62306      0.054394       32712           0    47.45374    47.61584    47.74856    -0.04044    -0.65605
Longitude      INPUT    -122.332      0.029985       32712           0    -122.431    -122.331    -122.222    -0.08384    -7.72815



Class Variable Summary Statistics by Class Target
(maximum 500 observations printed)

Data Role=TRAIN Variable Name=Census_Tract_2000

                       Number
             Target      of                              Mode                 Mode2
 Target      Level     Levels    Missing    Mode    Percentage    Mode2    Percentage
```

As we can see from above screenshot the skewness of the variables have decreased.

Step 3 - Data Partition

Once we filtered all the missing values and cleaned the data, we are applying data partition node to partition data into training, validation and testing in order to build prediction model. Data allocation is as below.



Partition Summary

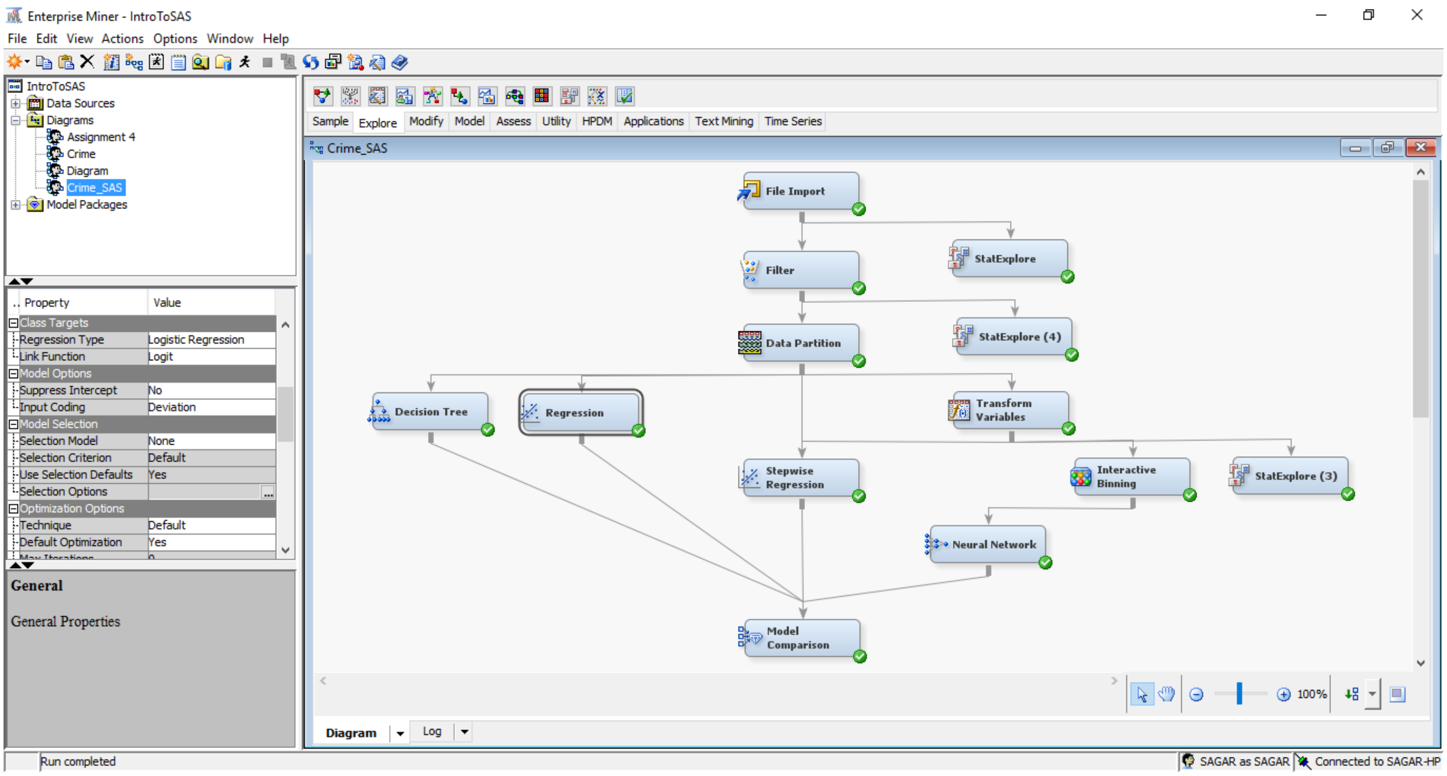
Type	**Percentage of Data**	**No of Observations**
Data	100%	46783
Train	70%	32746
Validate	20%	9356
Test	10%	4681

Step 4 - Predictive Modelling

Step 4.1- Regression

We are performing regression analysis on the dataset to find significant variables which drive the target variable Violent.

As our target variable is binary - Violent or Non Violent, we are using Logit regression to predict, given new set of data whether the crime will be Violent or Non-Violent.



Results of Regression

Fit Statistics table for Regression-

Results - Node: Regression Diagram: Crime_SAS

File Edit View Window

Fit Statistics

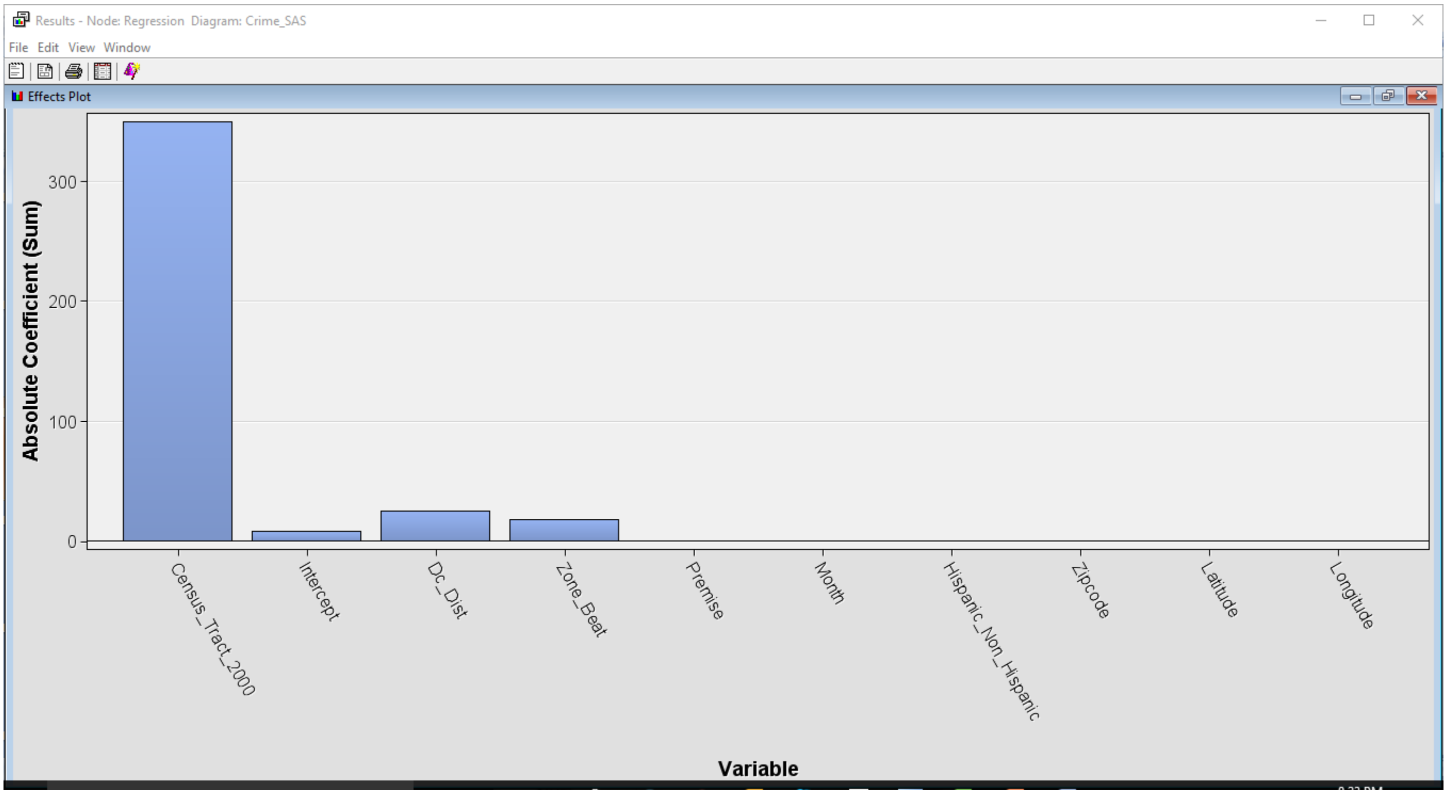
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Violent		_AIC_	Akaike's Information Criterion	34960.04	.	.
Violent		_ASE_	Average Squared Error	0.174445	0.177559	0.178474
Violent		_AVERR_	Average Error Function	0.526905	0.535646	0.537927
Violent		_DFE_	Degrees of Freedom for Error	32520	.	.
Violent		_DFM_	Model Degrees of Freedom	226	.	.
Violent		_DFT_	Total Degrees of Freedom	32746	.	.
Violent		_DIV_	Divisor for ASE	65492	18712	9362
Violent		_ERR_	Error Function	34508.04	10023	5036.074
Violent		_FPE_	Final Prediction Error	0.176869	.	.
Violent		_MAX_	Maximum Absolute Error	0.97371	0.99978	0.970015
Violent		_MSE_	Mean Square Error	0.175657	0.177559	0.178474
Violent		_NOBS_	Sum of Frequencies	32746	9356	4681
Violent		_NW_	Number of Estimate Weights	226	.	.
Violent		_RASE_	Root Average Sum of Squares	0.417666	0.421377	0.422462
Violent		_RFPE_	Root Final Prediction Error	0.420558	.	.
Violent		_RMSE_	Root Mean Squared Error	0.419115	0.421377	0.422462
Violent		_SBC_	Schwarz's Bayesian Criterion	36857.66	.	.
Violent		_SSE_	Sum of Squared Errors	11424.73	3322.483	1670.878
Violent		_SUMW_	Sum of Case Weights Times Freq	65492	18712	9362
Violent		_MISC_	Misclassification Rate	0.246229	0.249359	0.249733

The fit statistics tell us about the accuracy of the model. As we can see the Misclassification Rate (0.249359), Mean Squared Error (0.177559) and Root Mean Square Error (0.421377) are low.

Low value of Mean square error specifies that the predicted values are close to the actual values.

Misclassification Rate is the fraction of cases assigned to the wrong class, so lower the value of misclassification rate, and better the model.

Effect Plot Window



From the effects plot window we can see greater the absolute value of variable, the more important that variable is to the regression model. In the data set most important variables and thus significant predictor variables are Census_Tract_2000, Dc_Dist, Zone/Beat.

Results - Node: Regression Diagram: Crime_SAS

File Edit View Window

Output

```
                             Odds Ratio Estimates

                                                                 Point
Effect                                                        Estimate

Census_Tract_2000     100 vs NULL                                2.421
Census_Tract_2000     1000 vs NULL                               5.022
Census_Tract_2000     10001 vs NULL                              0.355
Census_Tract_2000     10002 vs NULL                              0.476
Census_Tract_2000     101 vs NULL                                4.734
Census_Tract_2000     10100 vs NULL                              0.330
Census_Tract_2000     10101 vs NULL                              0.289
Census_Tract_2000     10200 vs NULL                              0.103
Census_Tract_2000     10300 vs NULL                              0.294
Census_Tract_2000     10301 vs NULL                              0.082
Census_Tract_2000     10401 vs NULL                              0.160
Census_Tract_2000     10402 vs NULL                              0.349
Census_Tract_2000     10500 vs NULL                            999.000
Census_Tract_2000     10501 vs NULL                            999.000
Census_Tract_2000     10600 vs NULL                            769.674
Census_Tract_2000     10601 vs NULL                            999.000
Census_Tract_2000     10701 vs NULL                            999.000
Census_Tract_2000     10702 vs NULL                            999.000
Census_Tract_2000     10800 vs NULL                            999.000
Census_Tract_2000     10900 vs NULL                              1.380
Census_Tract_2000     1100 vs NULL                              22.862
Census_Tract_2000     11001 vs NULL                              0.096
Census_Tract_2000     11002 vs NULL                              0.096
Census_Tract_2000     11101 vs NULL                              0.096
Census_Tract_2000     11102 vs NULL                              0.061
Census_Tract_2000     11200 vs NULL                            999.000
Census_Tract_2000     11300 vs NULL                            853.160
Census_Tract_2000     11301 vs NULL                            999.000
Census_Tract_2000     11401 vs NULL                            999.000
Census_Tract_2000     11402 vs NULL                            999.000
Census_Tract_2000     11500 vs NULL                            999.000
Census_Tract_2000     11600 vs NULL                            999.000
```

Results - Node: Regression Diagram: Crime_SAS

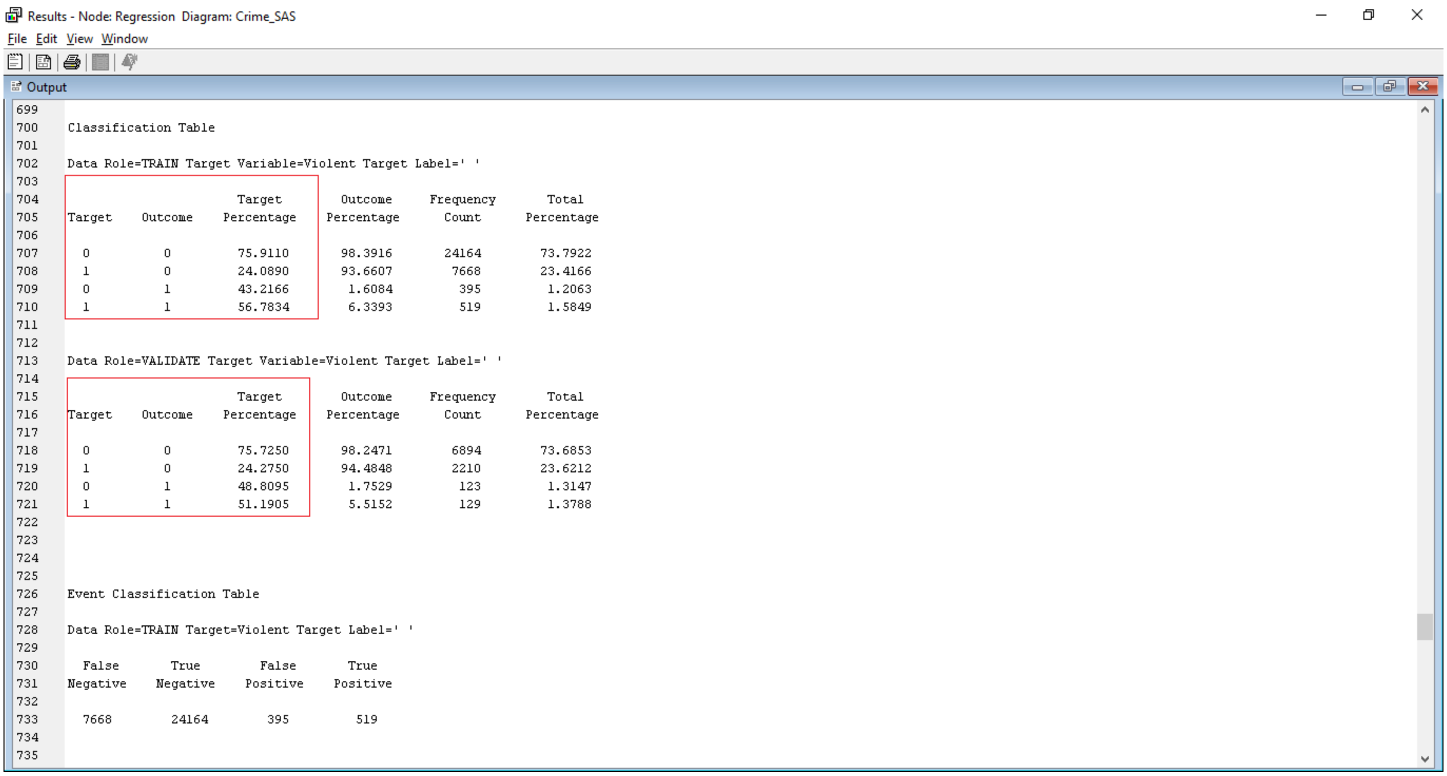
File Edit View Window

Output

```
Dc_Dist               O vs W                                   999.000
Dc_Dist               Q vs W                                   711.592
Dc_Dist               R vs W                                   999.000
Dc_Dist               S vs W                                   999.000
Dc_Dist               U vs W                                     .
Hispanic_Non_Hispanic                                            0.963
Month                                                            0.888
Premise               Driveway vs Vehicle                        0.963
Premise               Garage/Carport vs Vehicle                  0.963
Premise               Multi-Unit Dwelling vs Vehicle             0.999
Premise               Parking Lot vs Vehicle                     1.020
Premise               Sidewalk vs Vehicle                        0.967
Premise               Single Family Dwelling vs Vehicle          0.821
Premise               Store vs Vehicle                           0.924
Premise               Streets/Parkways vs Vehicle                1.014
REP_Latitude                                                     .
REP_Longitude                                                    .
Zipcode                                                          1.000
Zone_Beat             99 vs W3                                  15.897
Zone_Beat             B1 vs W3                                   2.011
Zone_Beat             B2 vs W3                                   1.702
Zone_Beat             B3 vs W3                                   .
Zone_Beat             C1 vs W3                                   2.192
Zone_Beat             C2 vs W3                                   2.746
Zone_Beat             C3 vs W3                                   .
Zone_Beat             D1 vs W3                                   0.911
Zone_Beat             D2 vs W3                                   0.789
Zone_Beat             D3 vs W3                                   .
Zone_Beat             E1 vs W3                                   1.788
Zone_Beat             E2 vs W3                                   3.998
Zone_Beat             E3 vs W3                                   .
Zone_Beat             F1 vs W3                                   1.663
Zone_Beat             F2 vs W3                                   0.314
Zone_Beat             F3 vs W3                                   .
Zone_Beat             G1 vs W3                                   2.505
Zone_Beat             G2 vs W3                                   1.299
Zone_Beat             G3 vs W3                                   .
```

From above output window, we can see that Census_Tract_2000, Dc_Dist and Zone/Beat are important variables.

Classification Table



Results - Node: Regression Diagram: Crime_SAS

File Edit View Window

Output

Classification Table

Data Role=TRAIN Target Variable=Violent Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	75.9110	98.3916	24164	73.7922
1	0	24.0890	93.6607	7668	23.4166
0	1	43.2166	1.6084	395	1.2063
1	1	56.7834	6.3393	519	1.5849

Data Role=VALIDATE Target Variable=Violent Target Label=' '

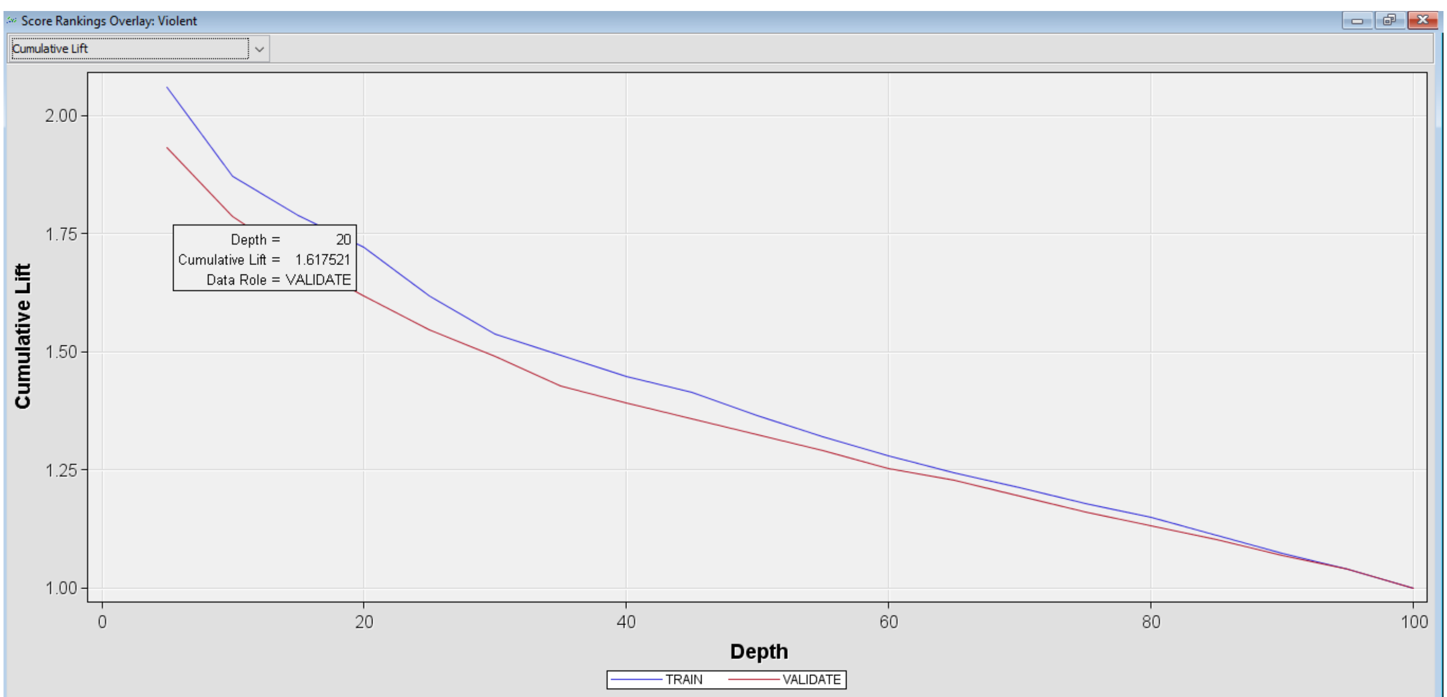
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	75.7250	98.2471	6894	73.6853
1	0	24.2750	94.4848	2210	23.6212
0	1	48.8095	1.7529	123	1.3147
1	1	51.1905	5.5152	129	1.3788

Event Classification Table

Data Role=TRAIN Target=Violent Target Label=' '

False Negative	True Negative	False Positive	True Positive
7668	24164	395	519

Based on the result from classification table, regression model is 51.1905% accurate for true positive (target variable= 1, predicted outcome = 1) and 75.7250% accurate for true negative (target variable =0, predicted outcome =0).

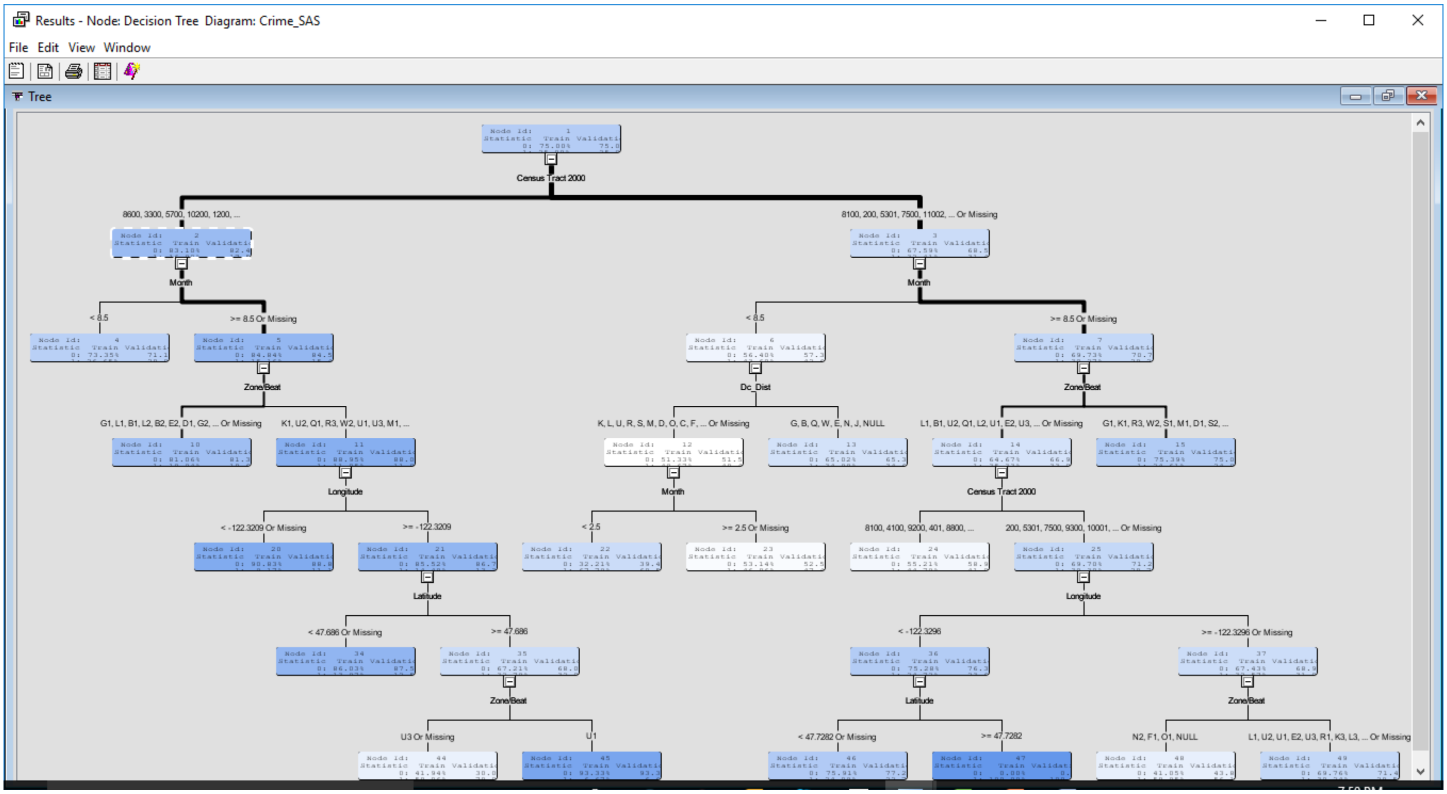


Lift is the improvement obtained by modelling that is the ratio between the result obtained with and without the predictive modelling. Cumulative lift value at the top twenty percentile in the validation data for Regression node is **1.617521.**

Step 4.2- Decision Tree

After regression, we ran Decision Tree node on our data set as decision tree do not require any assumptions of linearity in the data and very ease to interpret.

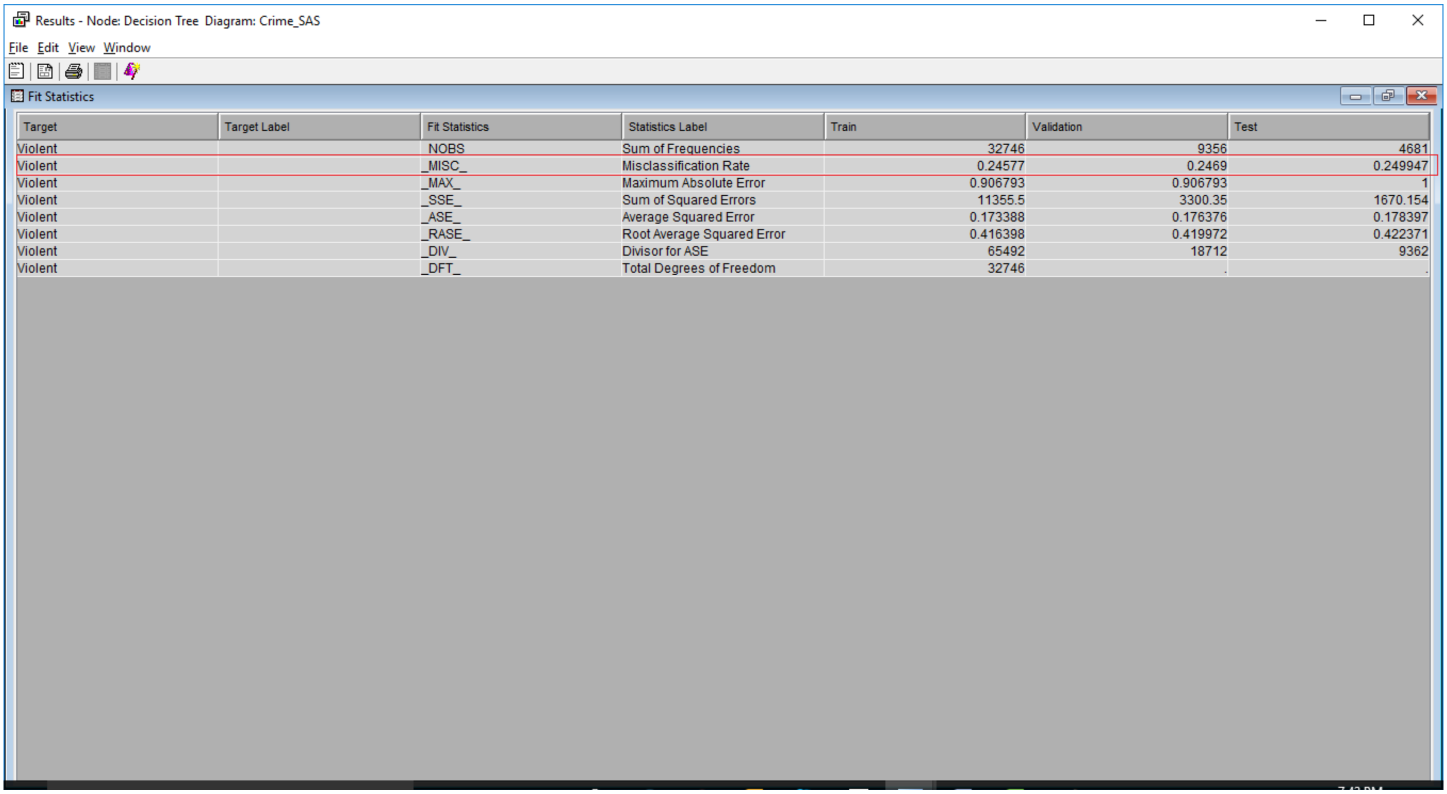
Results of Decision tree is as below-



From above result, we can see that root node selected by decision tree is **Census_Tract_2000** which is the Neighbourhood code in which crime has occurred.

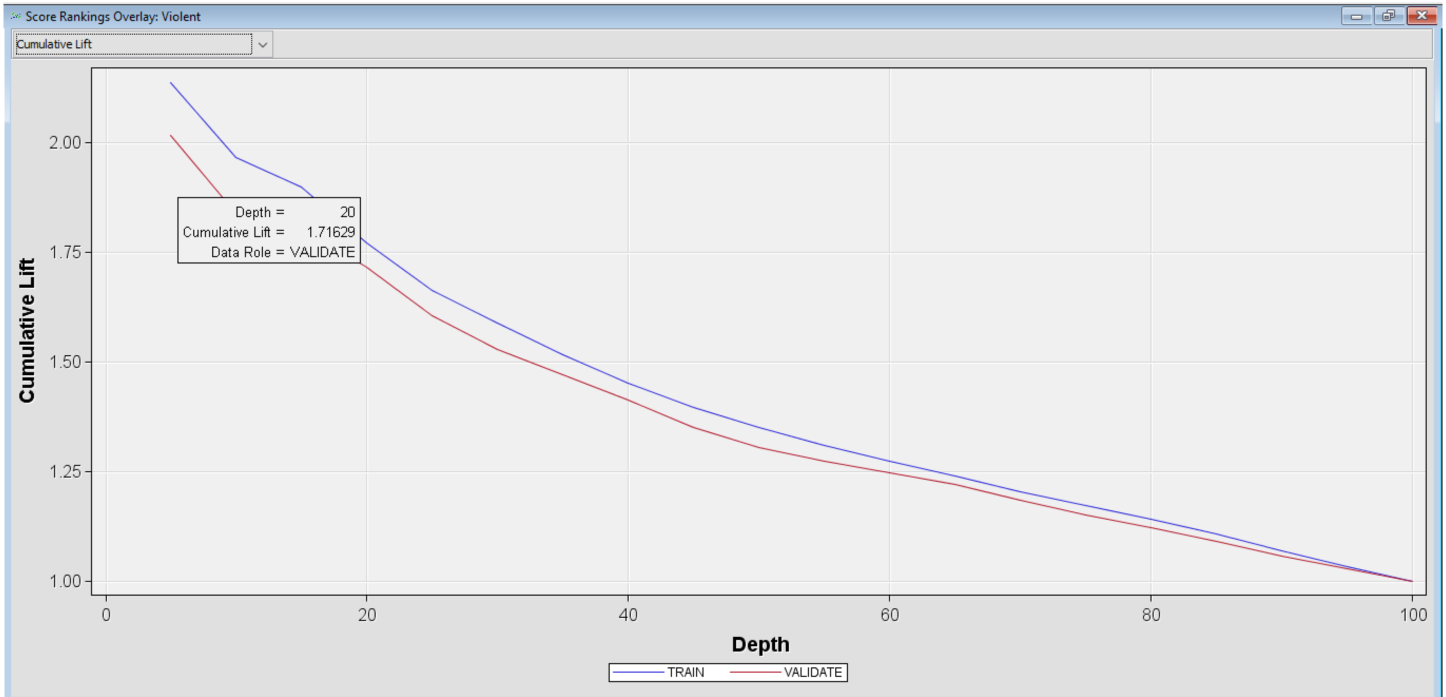
Other variables considered are Month, Zone/beat, Latitude, Longitude, Dc_Dist.

Fit Statistics for Decision Tree Node



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Violent		_NOBS_	Sum of Frequencies	32746	9356	4681
Violent		_MISC_	Misclassification Rate	0.24577	0.2469	0.249947
Violent		_MAX_	Maximum Absolute Error	0.906793	0.906793	1
Violent		_SSE_	Sum of Squared Errors	11355.5	3300.35	1670.154
Violent		_ASE_	Average Squared Error	0.173388	0.176376	0.178397
Violent		_RASE_	Root Average Squared Error	0.416398	0.419972	0.422371
Violent		_DIV_	Divisor for ASE	65492	18712	9362
Violent		_DFT_	Total Degrees of Freedom	32746	.	.

The Misclassification Rate in case of Decision Tree is coming out to be **0.2469** which is comparatively low compared to logistic regression model.



Cumulative lift value at the top twenty percentile in the validation data for Decision Tree Node is **1.71629.**

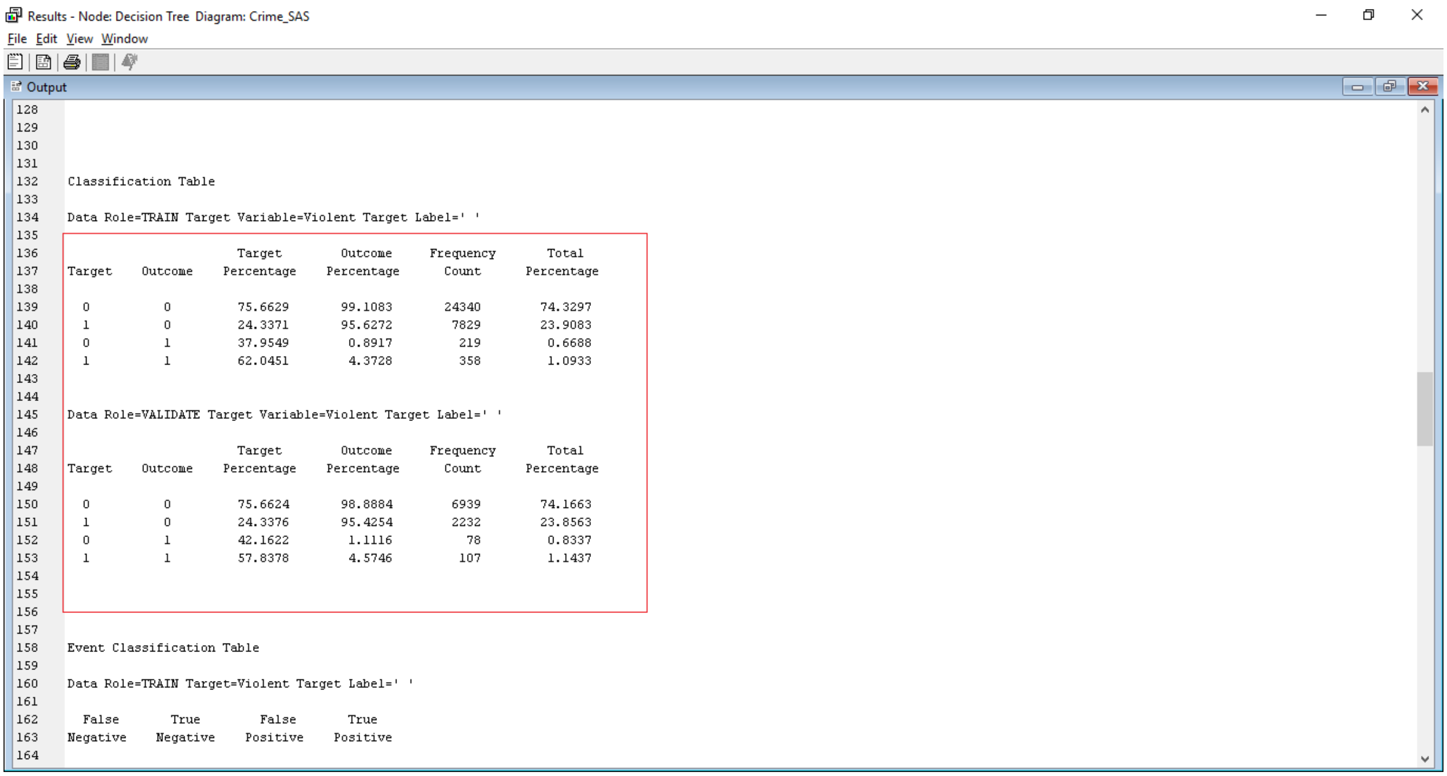
Variable Importance in Decision Trees

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Census_Tract_2000	Census Tract 2000	2	1.0000	1.0000	1.0000
Zone_Beat	Zone/Beat	3	0.6288	0.5508	0.8760
Month		6	0.5814	0.6964	1.1978
REP_Longitude	Replacement: Longitude	7	0.3889	0.4217	1.0845
REP_Latitude	Replacement: Latitude	4	0.3444	0.4160	1.2077
Dc_Dist		1	0.1192	0.1576	1.3221
Zipcode		0	0.0000	0.0000	.
Hispanic_Non_Hispanic	Hispanic/Non-Hispanic	0	0.0000	0.0000	.
Premise		0	0.0000	0.0000	.

From above Variable importance window, we can see that the important variables are Census_Tract_2000, Zone/Beat, Month, Longitude, Latitude, Dc_Dist.

Classification Table for Decision Tree


Results - Node: Decision Tree Diagram: Crime_SAS
File Edit View Window
Output

Classification Table

Data Role=TRAIN Target Variable=Violent Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	75.6629	99.1083	24340	74.3297
1	0	24.3371	95.6272	7829	23.9083
0	1	37.9549	0.8917	219	0.6688
1	1	62.0451	4.3728	358	1.0933

Data Role=VALIDATE Target Variable=Violent Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	75.6624	98.8884	6939	74.1663
1	0	24.3376	95.4254	2232	23.8563
0	1	42.1622	1.1116	78	0.8337
1	1	57.8378	4.5746	107	1.1437

Event Classification Table

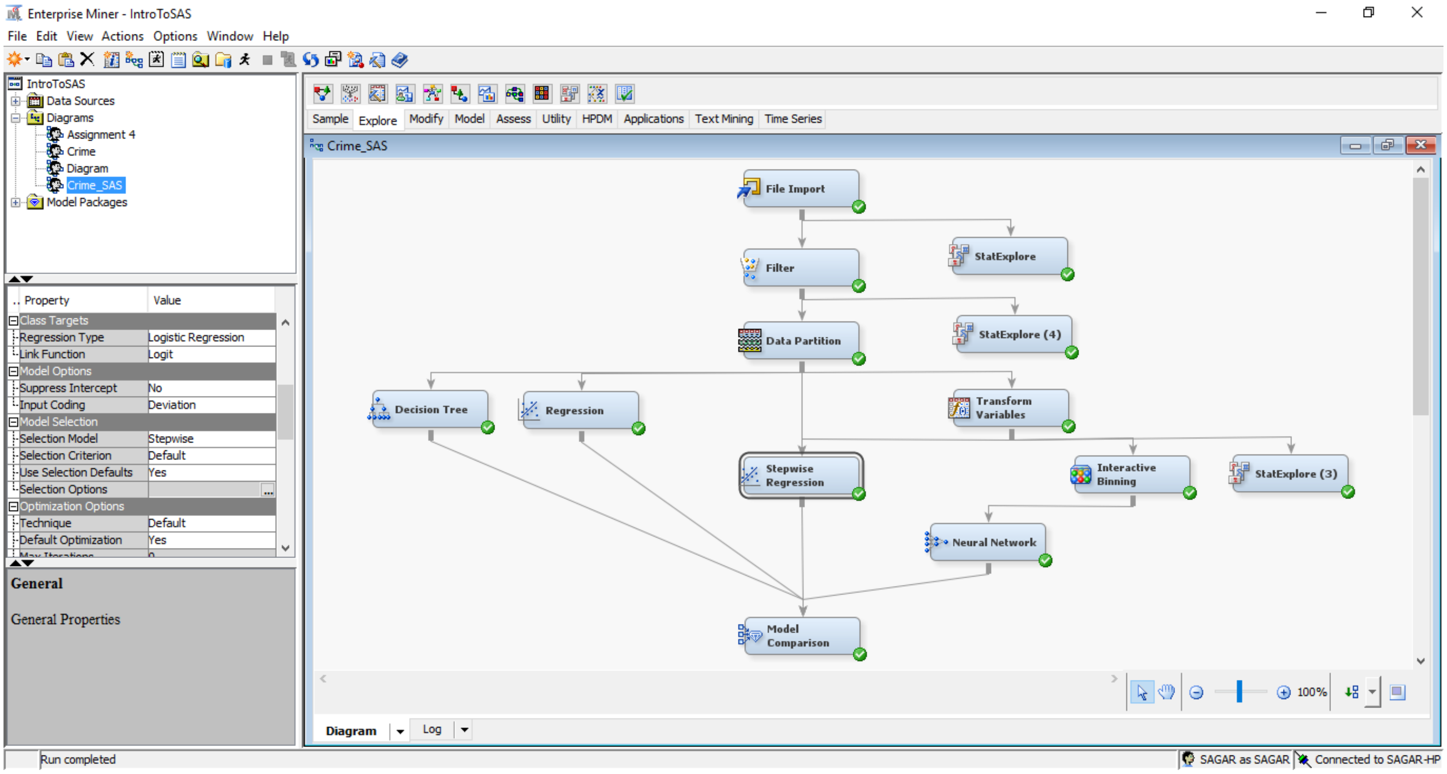
Data Role=TRAIN Target=Violent Target Label=' '

False Negative	True Negative	False Positive	True Positive

Based on the result from classification table, regression model is 57.8378% accurate for true positive (target variable= 1, predicted outcome = 1) and 75.6624% accurate for true negative (target variable =0, predicted outcome =0).

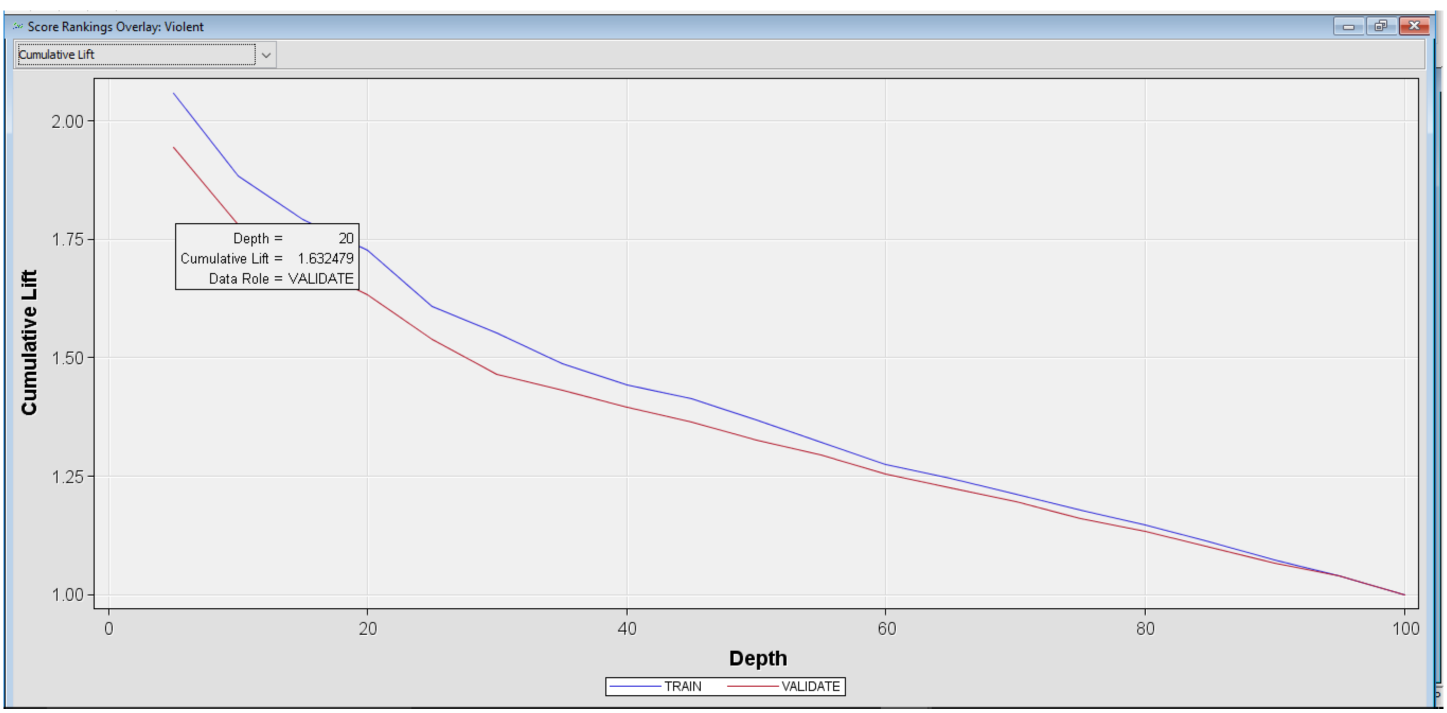
Step 4.3 - Stepwise Regression – Transform Variable

We have used Stepwise Regression node after transforming the variables and compared the results with other predictive models.



Cumulative Lift

The below graph shows cumulative lift in case of Stepwise Regression-



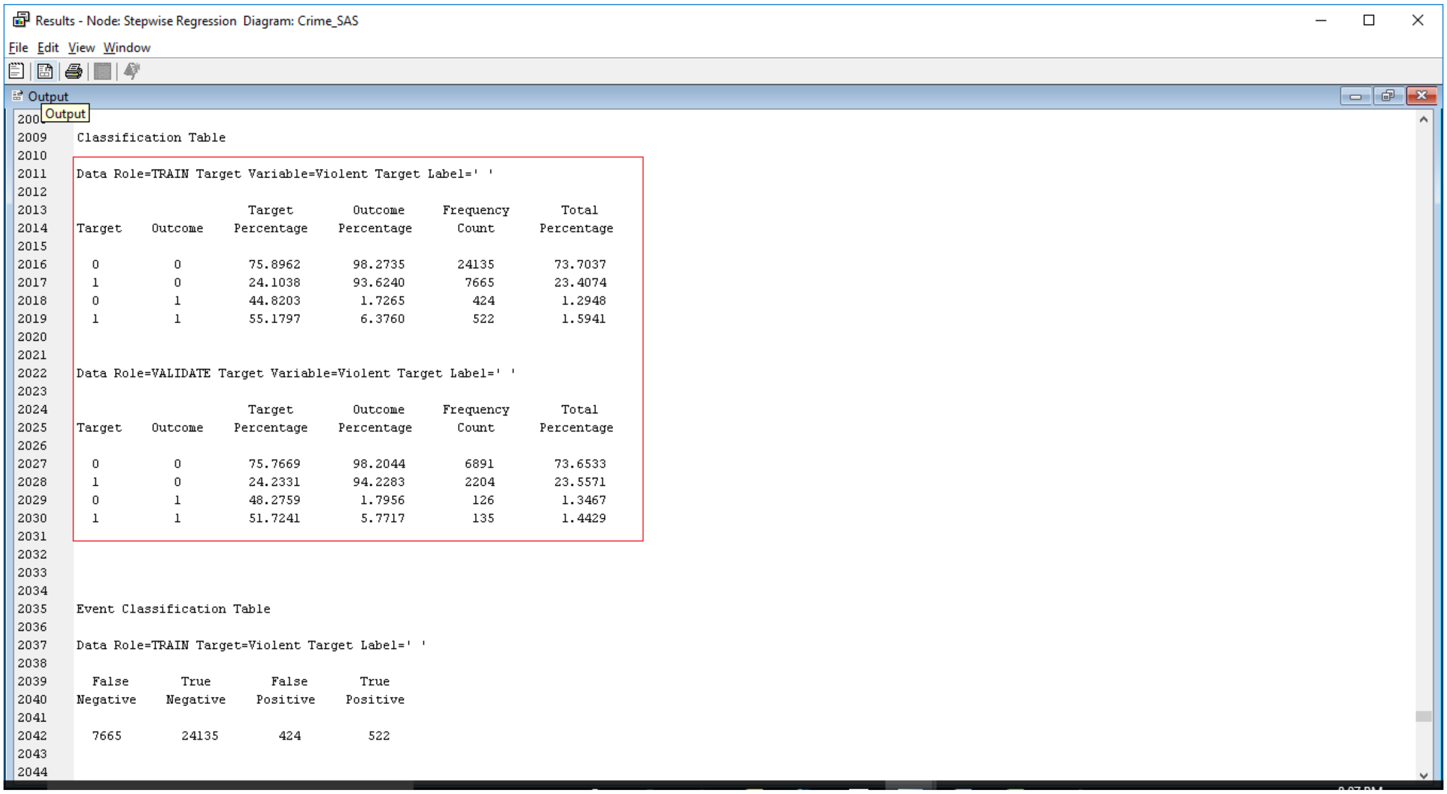
We can see that cumulative lift at the top twenty percentile for stepwise regression is **1.632479**.

Fit Statistics for Stepwise Regression

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Violent		_AIC_	Akaike's Information Criterion	34950.97	.	.
Violent		_ASE_	Average Squared Error	0.174495	0.177452	0.178409
Violent		_AVERR_	Average Error Function	0.527071	0.535408	0.537707
Violent		_DFE_	Degrees of Freedom for Error	32530	.	.
Violent		_DFM_	Model Degrees of Freedom	216	.	.
Violent		_DFT_	Total Degrees of Freedom	32746	.	.
Violent		_DIV_	Divisor for ASE	65492	18712	9362
Violent		_ERR_	Error Function	34518.97	10018.55	5034.015
Violent		_FPE_	Final Prediction Error	0.176812	.	.
Violent		_MAX_	Maximum Absolute Error	0.972811	0.999787	0.969486
Violent		_MSE_	Mean Square Error	0.175653	0.177452	0.178409
Violent		_NOBS_	Sum of Frequencies	32746	9356	4681
Violent		_NW_	Number of Estimate Weights	216	.	.
Violent		_RASE_	Root Average Sum of Squares	0.417725	0.421251	0.422385
Violent		_RFPE_	Root Final Prediction Error	0.42049	.	.
Violent		_RMSE_	Root Mean Squared Error	0.41911	0.421251	0.422385
Violent		_SBC_	Schwarz's Bayesian Criterion	36764.62	.	.
Violent		_SSE_	Sum of Squared Errors	11428	3320.49	1670.268
Violent		_SUMW_	Sum of Case Weights Times Freq	65492	18712	9362
Violent		_MISC_	Misclassification Rate	0.247023	0.249038	0.249733

The fit statistics tell us about the accuracy of the model. As we can see the Misclassification Rate (0.249038), Mean Squared Error (0.177452) and Root Mean Square Error (0.421251) are high.
High value of Mean square error specifies that the predicted values are not close to the actual values.
Misclassification Rate is the fraction of cases assigned to the wrong class, so lower the value of misclassification rate, and better the model.

Classification Table for Stepwise Regression



Results - Node: Stepwise Regression Diagram: Crime_SAS

File Edit View Window

Output

Classification Table

Data Role=TRAIN Target Variable=Violent Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	75.8962	98.2735	24135	73.7037
1	0	24.1038	93.6240	7665	23.4074
0	1	44.8203	1.7265	424	1.2948
1	1	55.1797	6.3760	522	1.5941

Data Role=VALIDATE Target Variable=Violent Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	75.7669	98.2044	6891	73.6533
1	0	24.2331	94.2283	2204	23.5571
0	1	48.2759	1.7956	126	1.3467
1	1	51.7241	5.7717	135	1.4429

Event Classification Table

Data Role=TRAIN Target=Violent Target Label=' '

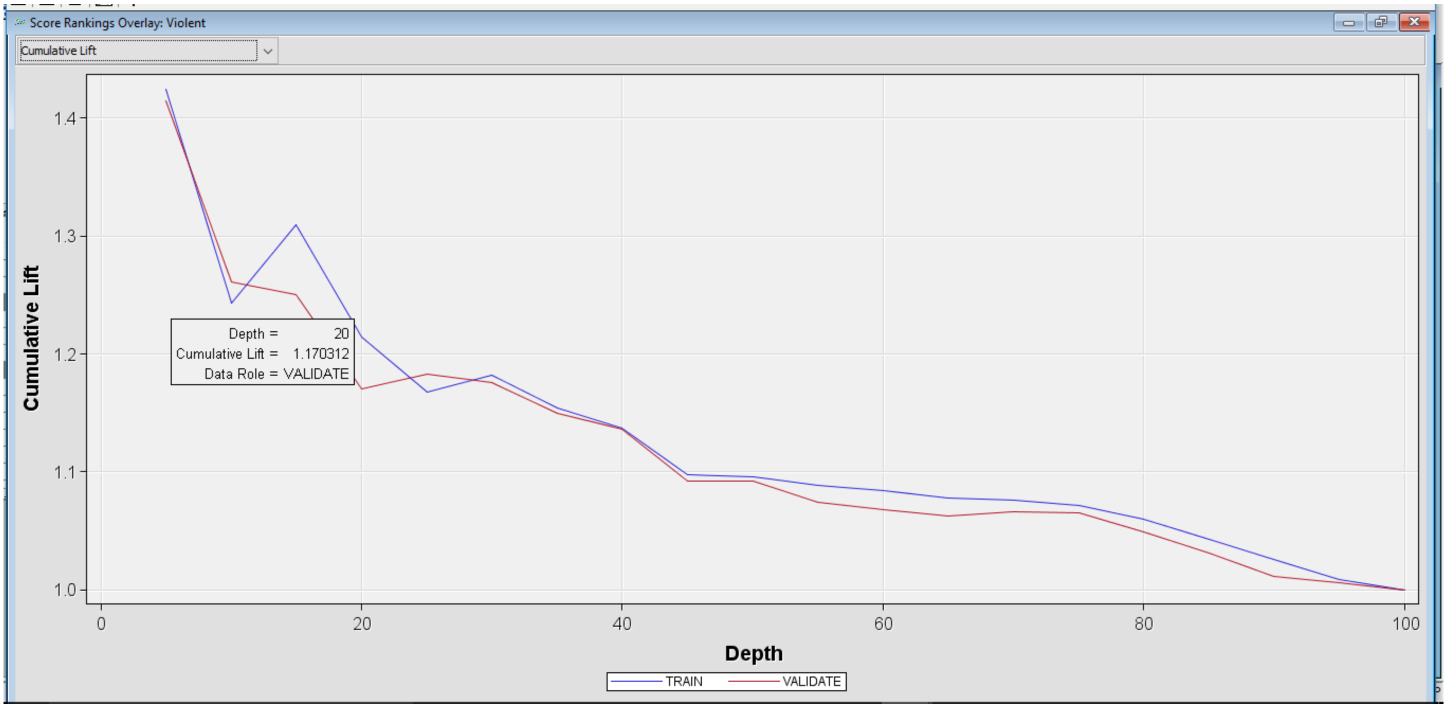
False Negative	True Negative	False Positive	True Positive
7665	24135	424	522

Although True positive rate and True negative rate is higher in stepwise regression as compared to Logistic regression model, but in comparison to Decision tree model that we presented, it is slightly lower.

Step 4.4 - Neural Network

In order to find best fit model, we are implementing Neural Network on our data set that can be used to construct, train, and validate multilayer feed forward neural networks.

Cumulative Lift



As we can see from the above screenshot the cumulative lift at the top twenty percentile for Neural Network node is **1.170312**.

Fit Statistics for Neural Network

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Violent		_DFT_	Total Degrees of Freedom	32746	.	.
Violent		_DFE_	Degrees of Freedom for Error	32397	.	.
Violent		_DFM_	Model Degrees of Freedom	349	.	.
Violent		_NW_	Number of Estimated Weights	349	.	.
Violent		_AIC_	Akaike's Information Criterion	37279.87	.	.
Violent		_SBC_	Schwarz's Bayesian Criterion	40210.26	.	.
Violent		_ASE_	Average Squared Error	0.186079	0.186488	0.185954
Violent		_MAX_	Maximum Absolute Error	0.816096	0.815502	0.815167
Violent		_DIV_	Divisor for ASE	65492	18712	9362
Violent		_NOBS_	Sum of Frequencies	32746	9356	4681
Violent		_RASE_	Root Average Squared Error	0.431368	0.431842	0.431224
Violent		_SSE_	Sum of Squared Errors	12186.67	3489.558	1740.9
Violent		_SUMW_	Sum of Case Weights Times Freq	65492	18712	9362
Violent		_FPE_	Final Prediction Error	0.190088	.	.
Violent		_MSE_	Mean Squared Error	0.188083	0.186488	0.185954
Violent		_RFPE_	Root Final Prediction Error	0.435991	.	.
Violent		_RMSE_	Root Mean Squared Error	0.433686	0.431842	0.431224
Violent		_AVERR_	Average Error Function	0.55857	0.5597	0.558324
Violent		_ERR_	Error Function	36581.87	10473.11	5227.028
Violent		_MISC_	Misclassification Rate	0.250015	0.25	0.25016
Violent		_WRONG_	Number of Wrong Classifications	8187	2339	1171

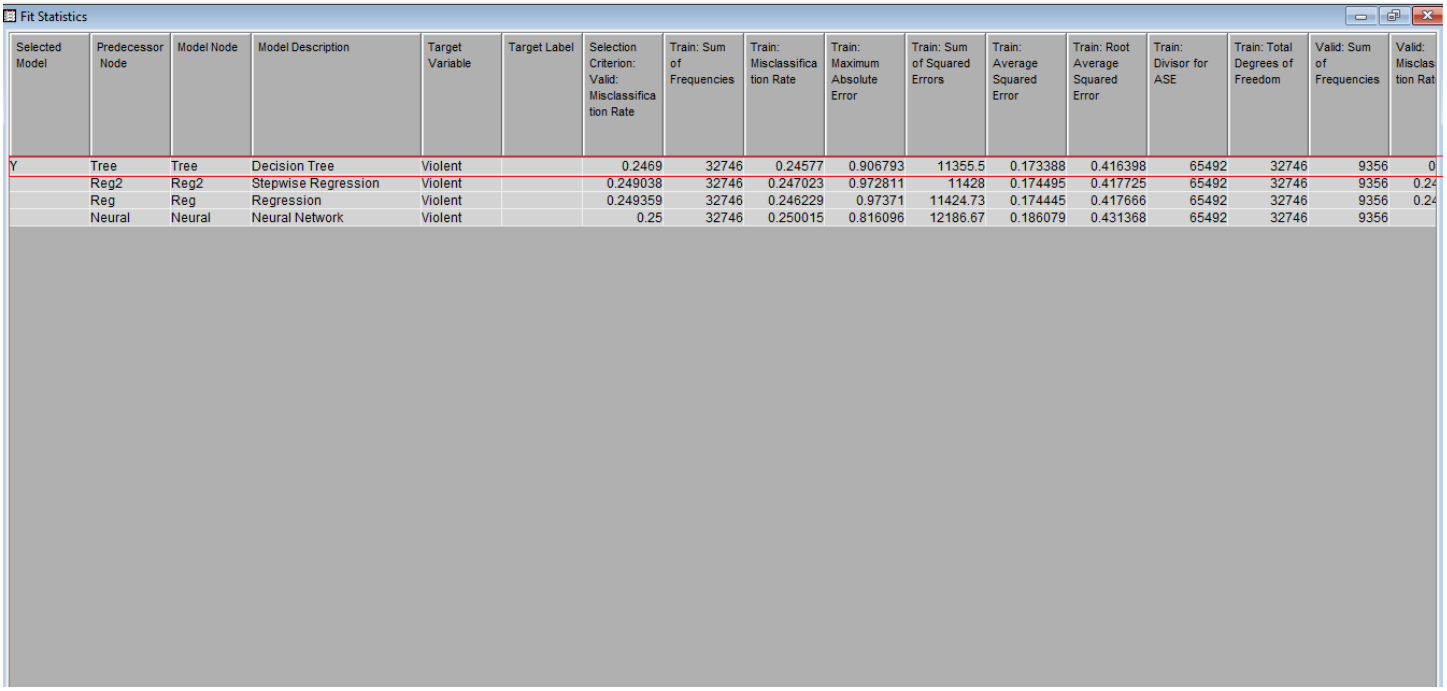
Mean Squared Error (0.186488), Root Mean Squared Error (0.431842), and Misclassification Rate (0.25) for the Neural Network is comparatively more than the Decision Tree.

Output

```
_MISC_      Misclassification Rate              0.25        0.25       0.25
_WRONG_     Number of Wrong Classifications    8187.00     2339.00    1171.00




Classification Table

Data Role=TRAIN Target Variable=Violent Target Label=' '

                      Target      Outcome     Frequency     Total
Target    Outcome   Percentage   Percentage     Count     Percentage

  0          0       74.9985        100         24559      74.9985
  1          0       25.0015        100          8187      25.0015


Data Role=VALIDATE Target Variable=Violent Target Label=' '

                      Target      Outcome     Frequency     Total
Target    Outcome   Percentage   Percentage     Count     Percentage

  0          0         75           100          7017        75
  1          0         25           100          2339        25




Event Classification Table

Data Role=TRAIN Target=Violent Target Label=' '

  False       True        False       True
Negative    Negative    Positive    Positive

  8187       24559          0           0
```

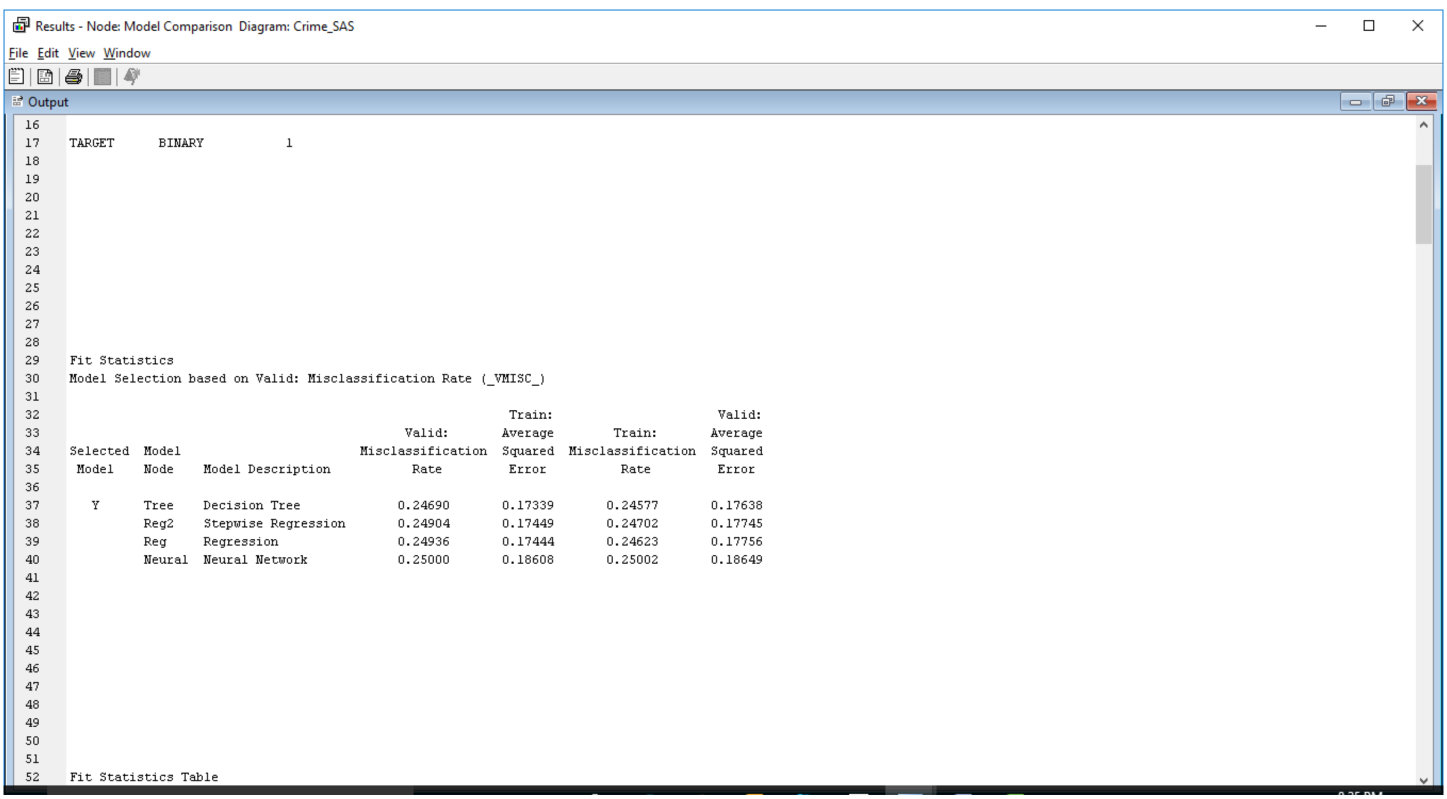
As we can see from the above classification table the true positive rate and true negative rate are comparatively lower than the decision tree which means decision tree is better model than neural network.

Step 5- Model Comparison



Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclas tion Rat
Y	Tree	Tree	Decision Tree	Violent		0.2469	32746	0.24577	0.906793	11355.5	0.173388	0.416398	65492	32746	9356	0
	Reg2	Reg2	Stepwise Regression	Violent		0.249038	32746	0.247023	0.972811	11428	0.174495	0.417725	65492	32746	9356	0.24
	Reg	Reg	Regression	Violent		0.249359	32746	0.246229	0.97371	11424.73	0.174445	0.417666	65492	32746	9356	0.24
	Neural	Neural	Neural Network	Violent		0.25	32746	0.250015	0.816096	12186.67	0.186079	0.431368	65492	32746	9356	



Results - Node: Model Comparison Diagram: Crime_SAS

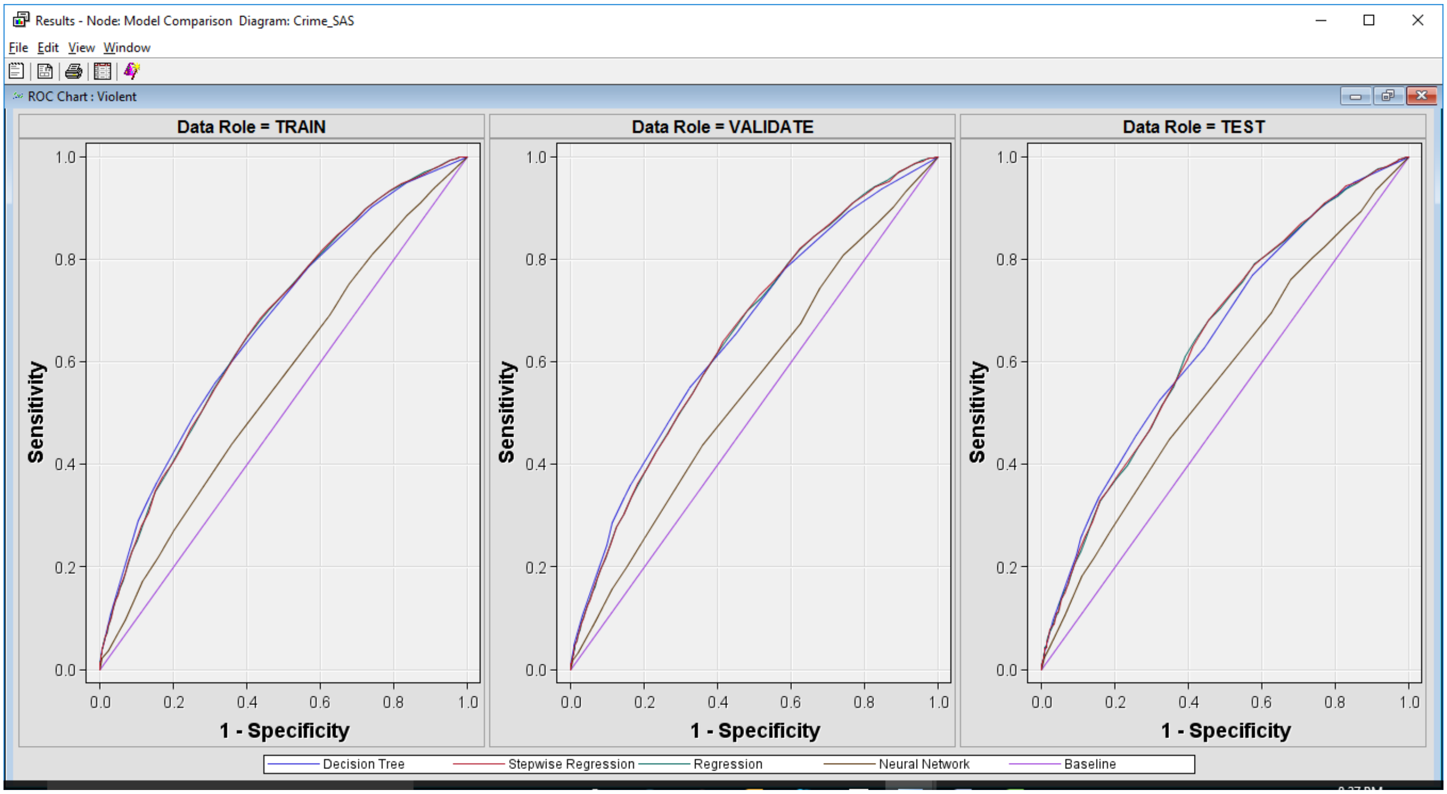
File Edit View Window

Output

```
TARGET      BINARY          1

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

                                                       Train:                     Valid:
                                          Valid:      Average        Train:       Average
Selected  Model                    Misclassification  Squared  Misclassification  Squared
 Model    Node    Model Description        Rate         Error         Rate         Error

   Y      Tree    Decision Tree          0.24690      0.17339       0.24577      0.17638
          Reg2    Stepwise Regression    0.24904      0.17449       0.24702      0.17745
          Reg     Regression             0.24936      0.17444       0.24623      0.17756
          Neural  Neural Network         0.25000      0.18608       0.25002      0.18649

Fit Statistics Table
```

Decision Tree is selected as best model by SAS Enterprise Miner for predicting Violent or Non Violent Crime. It is based on the misclassification rate, as the misclassification rate of decision tree (0.24690) is Least compared to other predictive models. The Model is not as accurate as the misclassification rate and mean squared error is a little high.

ROC Chart



After Running Model Comparison node, best fit model is Decision tree.
From the above ROC Chart, Decision Tree Curve is closer to 1.0 (top left) compared to other models, which shows better accuracy.

We can see from above stat table, misclassification rate is low 0.24577 compared to regression and neural network.

Conclusion

During the project we have explored different predictive models and we have determined some of the prominent variables from our analysis which affect the target variable.

- Census_Tract_2000
- Zone_Beat
- Month
- Longitude
- Latitude
- Dc_Dist

Census_Tract_2000 has a positive correlation with Violent. Violent crime will depend on the specific values of this variable.

Also, from the regression stats we have analyzed that the variables like Dc_Dist and Zone_Beat have a positive correlation with Violent. This is interesting as we can predict the category of crime for these particular Districts and Zone/Beat.

Recommendation

- As we know that Zone is positively correlated to the type of crime, PPD can use this information and manage its patrolling according to the zones; more resources can be allotted to the zones where Violent crime occurs.
- Also Month is positively correlated to the type of crime, PPD can use this information and be alert when violent crimes are at their highest during peak months.

  This way Philadelphia Police Department can deploy their resources in a more effective manner and devise solutions to crime problems by formulating crime prevention strategies.

References

1. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_logistic_sect030.htm
2. http://www.ats.ucla.edu/stat/sas/output/sas_logit_output.htm
3. http://www.simafore.com/blog/bid/62333/4-key-advantages-of-using-decision-trees-for-predictive-analytics
4. http://chicago.cbslocal.com/2015/10/22/violent-crime-statistics-for-every-city-in-america/