CRIME ANALYSIS IN DALLAS COUNTY

MIS 6324 Business Analytics with SAS

Professor: Zhe Zhang

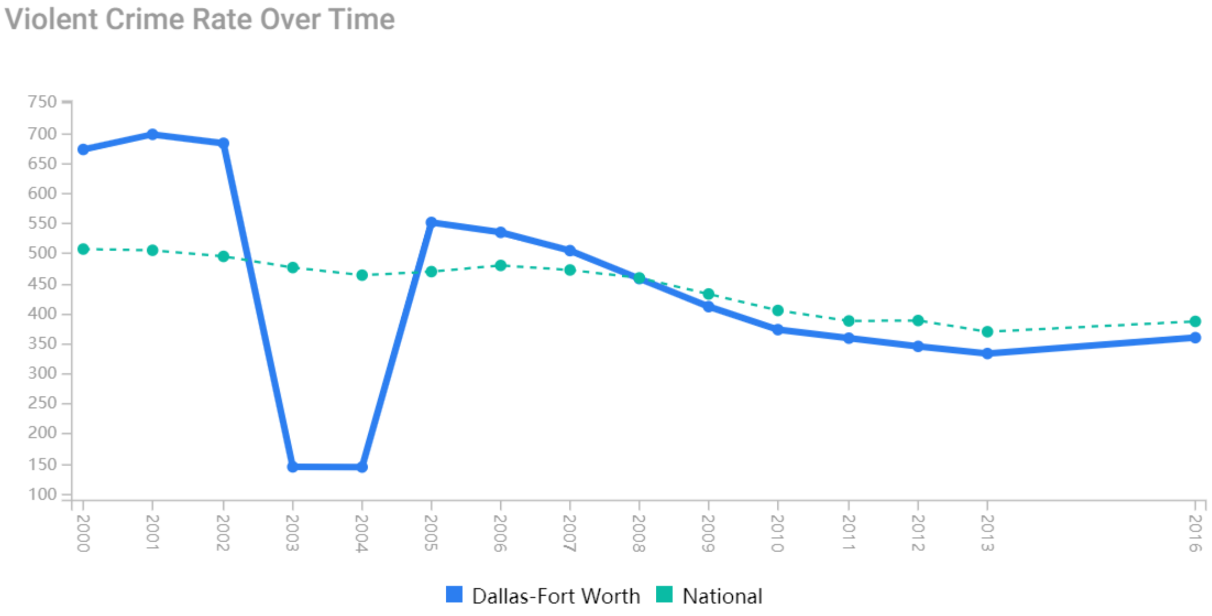
Bing Fu-bxf160930
Minjing Chen-mxc176330
Tuo Zhang-txz160830
Yuqiu Jiang-yxj172030

Contents

Background

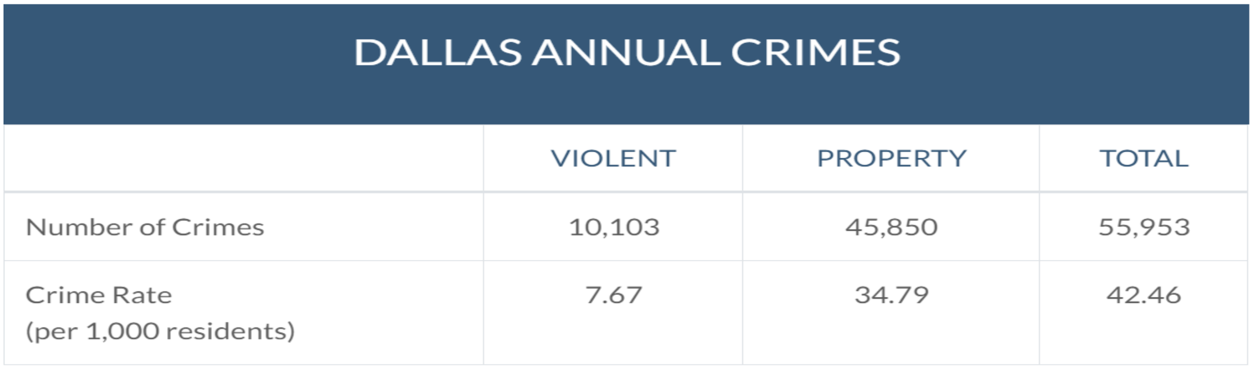
Based on Dallas Police Department information, DFW shows a crime rate which is noticeably lower than the average national rate. This means that for comparably-sized cities all across America, DFW area is actually safer than other cities. Especially from the year 2009 to the year 2016.

Figure-1



However, we found that Dallas County has a crime rate of 42 per one thousand residents (see Figure-2, this data reflects 2016 calendar year, which was released from FBI in September 2017), it has been one of the highest crime rates in America compared to all communities of all sizes, from the smallest towns to the largest cities. Within Texas, more than 91% of the communities have a lower crime rate than Dallas County.

Figure-2



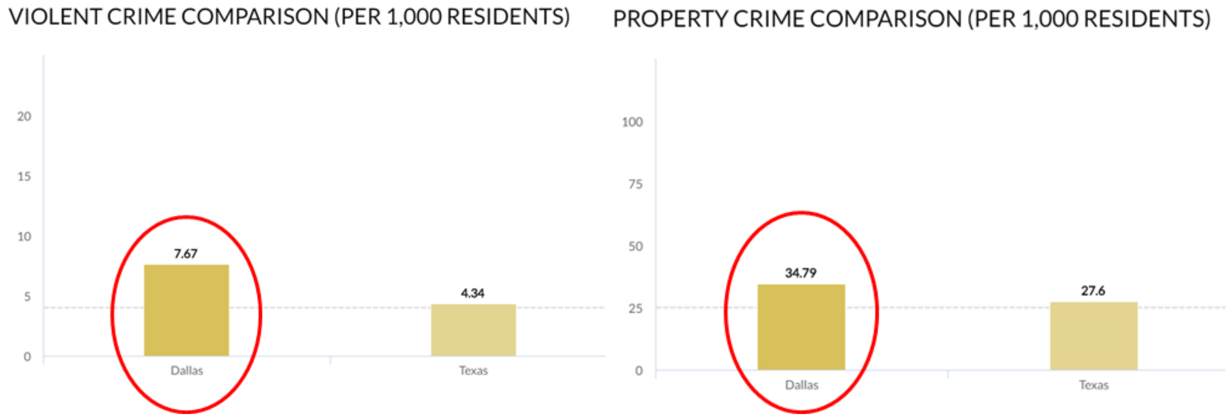
DALLAS ANNUAL CRIMES

	VIOLENT	PROPERTY	TOTAL
Number of Crimes	10,103	45,850	55,953
Crime Rate (per 1,000 residents)	7.67	34.79	42.46

Figure-3

Melissa
Decatur
Denton
McKinney
Frisco
Allen
Greenville
Lavon Lake
Plano
Royse City
Grapevine
Garland
Quinlan
Azle
Irving
Dallas
Fort Worth
Arlington
Terrell
Benbrook
DeSoto
Kaufman
Mansfield
Burleson
Kemp
Waxahachie
Cleburne
Ennis
Gun Barrel City
Google

Figure-4



In this project, we will turn to take a look at the details of the crime cases in Dallas County. We use the methodologies (descriptive analysis and predictive analysis) in SAS Enterprise Miner to dig out what is the major contribution variables to the general rate of crime in Dallas County? Which area of Dallas County tends to have more crimes than the others? Also, based on the location, day and time information, we can predict what type of crime will mostly happen to the citizen.

Data Cleaning

Data cleaning is the first step we need to do before starting the analyzing. It is a process that finding dirty or coarse data which may incorrect, incomplete, inaccurate or irrelevant, and then replacing or modifying them into right data format or information we need.

The dataset is used for this project is from Kaggle second-hand data. The data set is retrieved from https://www.kaggle.com/carrie1/dallaspolicereportedincidents. Below are the detail steps we did on data cleaning:

1. Drop unnecessary columns within the dataset. The original raw data has 103 columns, but most of the variables are redundant and not mandatory for data mining. Based on our target, we select 12 out of 103 variables, which are related to the final descriptive and predictive analysis.
2. Transfer ratio data into straight interval data. Time is discrete data ranging from 0:00:00 to 23:59:59. We classify the time data into three periods and transfer them into interval data: from 0:00:01 to 8:00:00 set them as 1, from 8:00:01 to 16:00:00 set them as 2, from 16:00:01 to 0:00:00 set them as 3.
3. Combine similar meaning descriptions to one terminology. Type of location columns has 34 unique items in total. As too much unique items will make the result vague, we find some of them can be combined into one category, like 'apartment' 'complex/build', 'apartment parking lot', 'apartment residence', etc. All of these description categories can be classified as 'apartment'. Finally, we reduce the unique items number from 34 to 20 after repeating the process mentioned before.
4. Extract analysis related data and got rid of no-related data. Latitude and longitude data are essential to our analysis, but the information is embedded in the location column, which included not only latitude and longitude information also, occurrence address and zip code. We first split information at location column and then only keep the latitude and longitude data.
5. Abandon unrelated data. The original data includes 3 classes of offenses under United States federal law: F(Felony), M(Misdemeanor) and I (Infraction). The target offense class we are focusing on is felony and misdemeanor, thus, infraction data has been got rid of.
6. Fortunately, after screening the whole dataset, we generate a new dataset with no missing data in the whole dataset.

Below are the detail variables and final data type we use in the SAS Enterprise Miner.

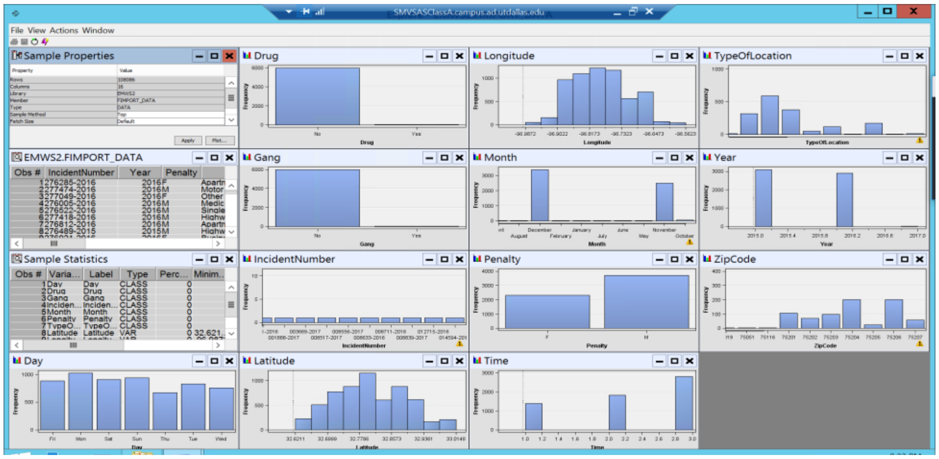
Table-1

Variables	Type of Data	Role
Incident Number	Nominal	ID
Year	Interval	Input
Month	Nominal	Input
Day	Nominal	Input
Time	Interval	Input
Type of Location	Nominal	Input
Zip code	Nominal	Input
Latitude	Interval	Input
Longitude	Interval	Input
Penalty	Binary	Target
Gang	Binary	Input
Drug	Binary	Input

Data Exploration

Below is the overall data exploration for these 12 variables.

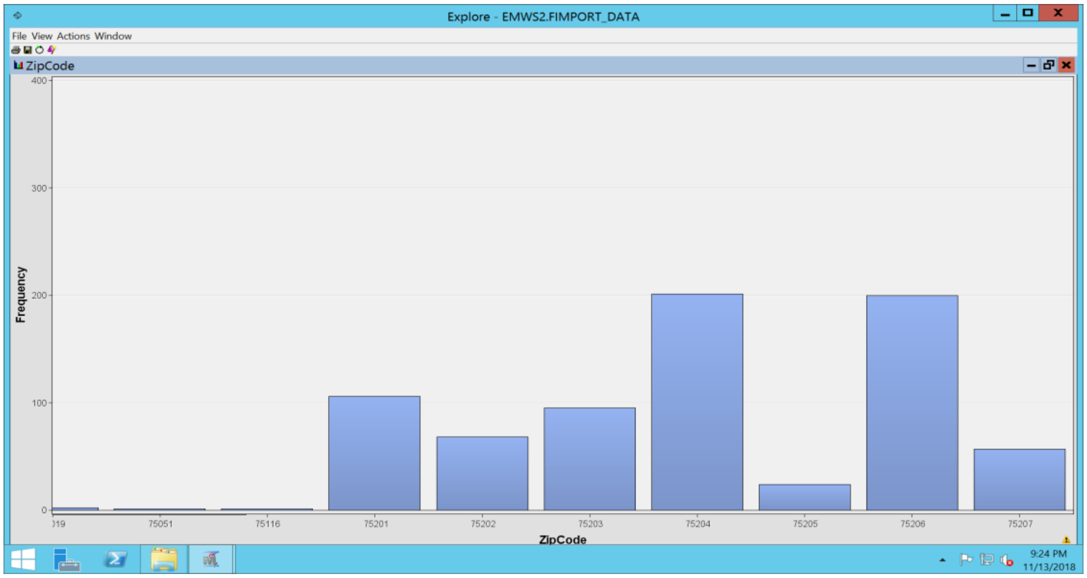
Figure-5



Here we pick up some typical variables and explore their data in detail. These exploration data can be a good reference during our analysis. It will give readers a big picture of how the data distributed, and what kind of prediction we may have after we run SAS Enterprise Miner.

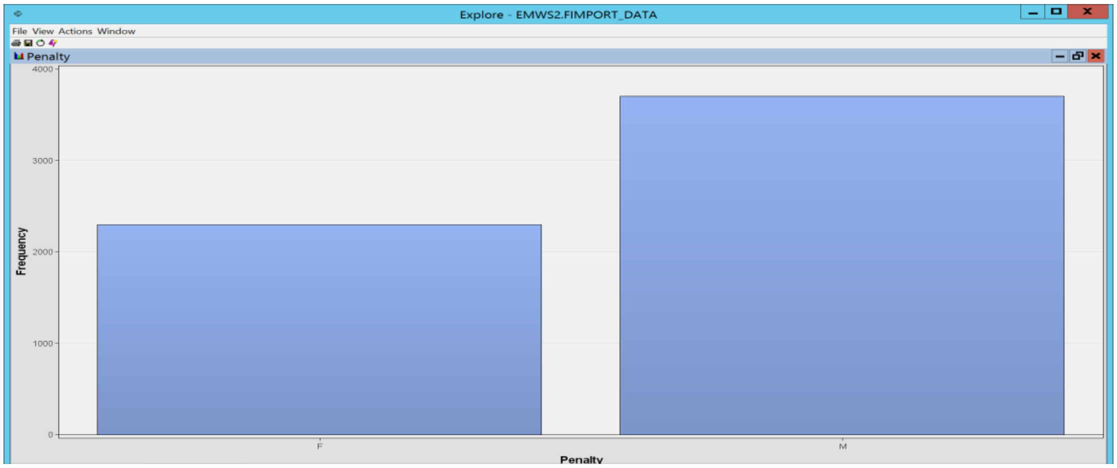
Zip code 75204 and 75206 are more dangerous compared with other Dallas County areas, which may indicate that those areas may not be an ideal location for living, study or work.

Figure-6



For the data exploration, we also can illustrate that the ratio for felony and misdemeanor is 2:3,

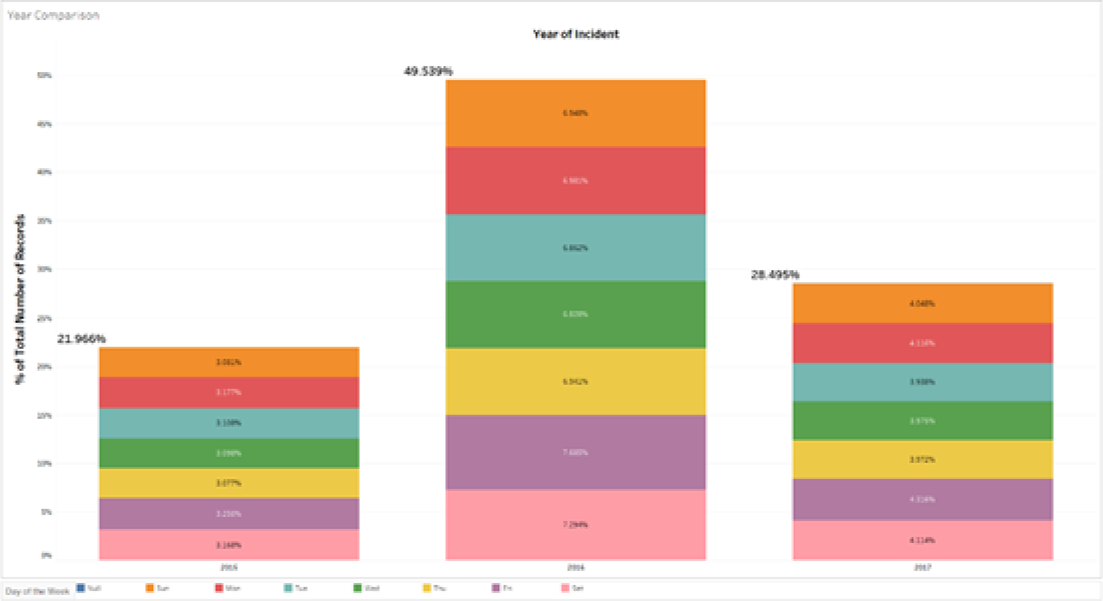
Figure-7



Descriptive Analysis

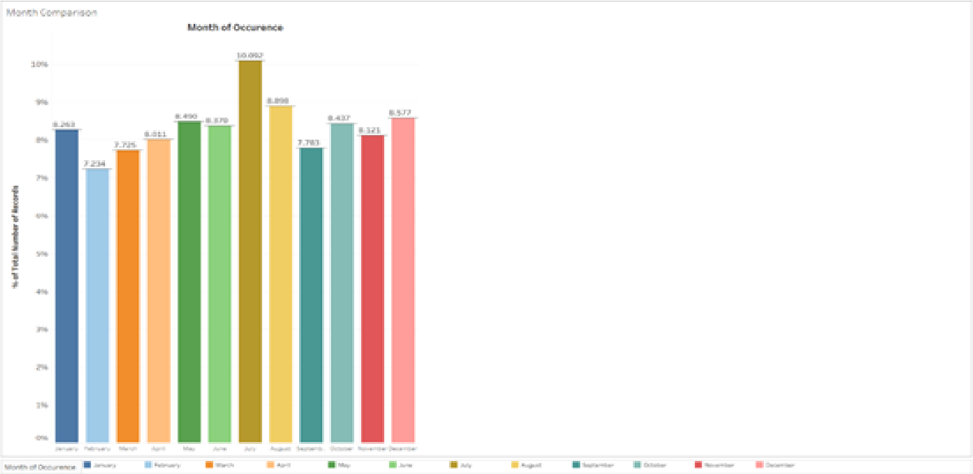
After we clean the data and explore all 12 variables, we would like to conduct a simple data visualization process to see what kind of patterns exist in our data. This visualization analysis would make us more familiar with our data and bring us more ideas about research direction. Here we use Tableau as our visualization tool.

Figure-8



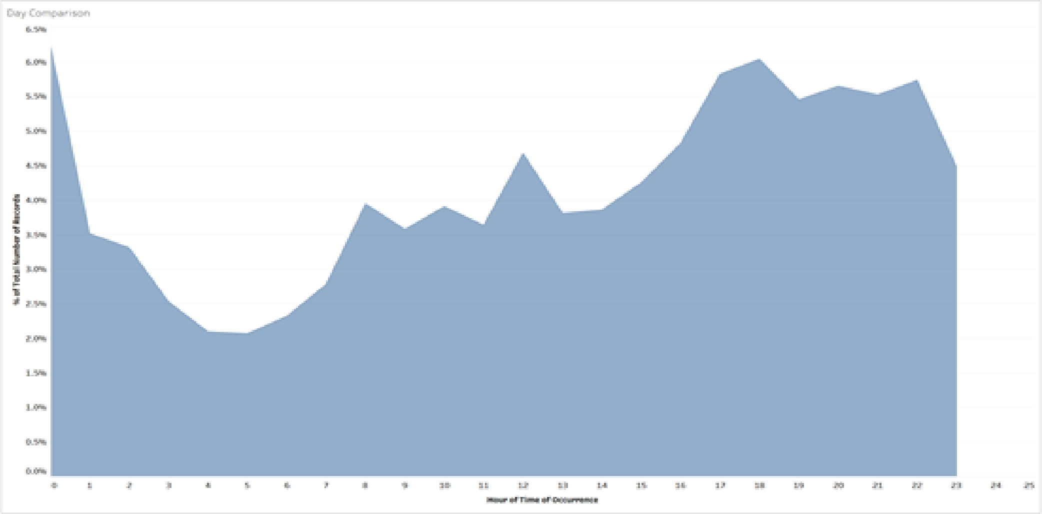
First, we generate a bar chart to show the percentage of the accident records each year. It displays each year's crime amount of the whole three years amount percentage. It is 22 % , 50% and 28% for 2015, 2016 and 2017 records respectively. The amount of accidents increases from 2015 to 2016 and decreases from 2016 to 2017. Different colors represent different days of a week. It directly shows that accidents happened on Friday are higher than any other days.

Figure-9



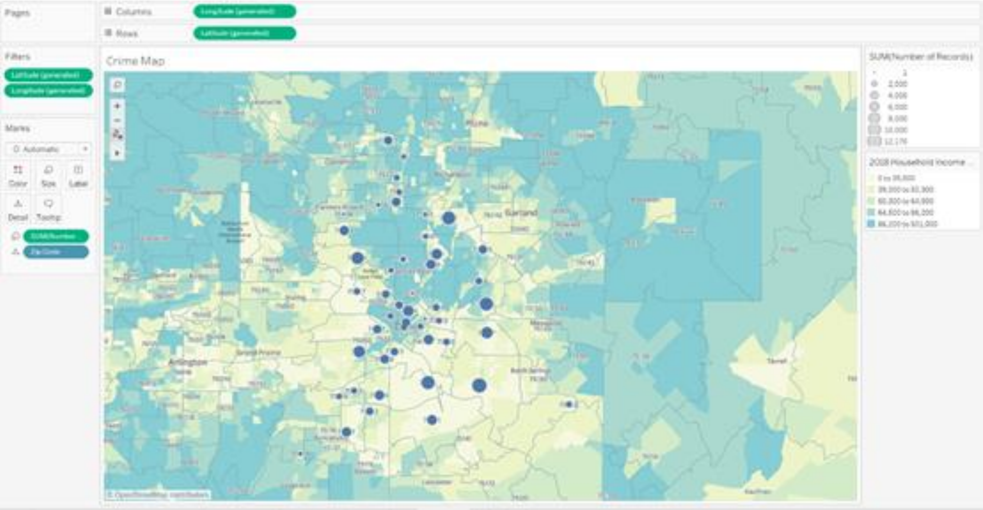
Second, we use the bar chart to analyze the distribution of the total accidents in 12 months. Each bar represents the sum of the accident records for a specific month in the dataset. Based on the bar chart, we can inform readers that July has the highest records than other months' records. In addition, it shows that accidents happened during the period from May to August and the period from December to January are higher than the rest months.

Figure-10



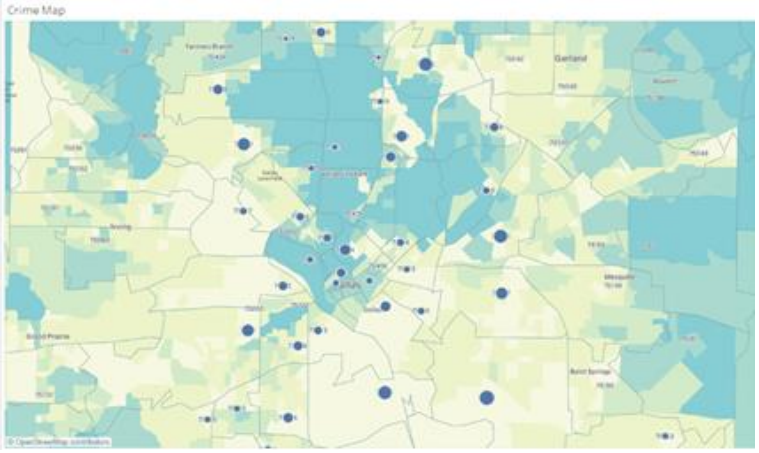
Next, we used 'area' to show the trend of accident occurrence in 24 hours. We sum up the records at a point of time in the dataset. Based on this graph, the number of accidents happened at 5 a.m. is the lowest and the number of accidents happened at 12 a.m. is the highest compared with the rest time within one day. Since 5 a.m., accidents occurrence increases gradually and reaches a periodic peak at 12 p.m. From 4 p.m. to 12 a.m., accidents occurrence keeps in high level.

Figure-11



At last, we try to take advantage of map tool in Tableau to geocode our data. The dots on the map represent the total number of accidents of different zip code. The size of the dot displays the quantity of the accidents. If the dot is larger, it means the total number of accidents in this specific area are higher than areas covered by smaller dots. The color of this map reveals the household income. If the color is darker, then the household income in that area is higher than other areas with smaller dots.

Figure-12

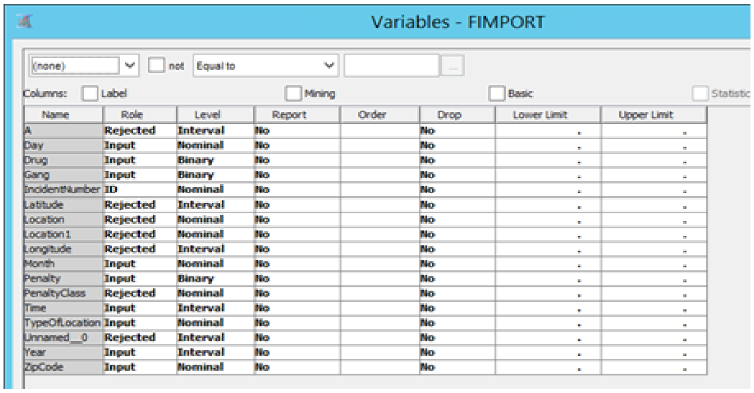

Crime Map

When we zoom in the map, the distribution pattern is clearer. We can see that if the household income is higher, the dots will be smaller, which means a fewer accidents occurrence.

Clustering Analysis

Based on what we have got from above, we would like to conduct a clustering analysis to further study on how many clusters could these accidents form and what kind of properties each cluster has. At last, we use latitude and longitude to geocode these clusters on a map and show the patterns of their distribution to readers.

Figure-13


Variables - FIMPORT

(none) □ not Equal to

Columns: □ Label □ Mining □ Basic □ Statistic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
A	Rejected	Interval	No		No	.	.
Day	Input	Nominal	No		No	.	.
Drug	Input	Binary	No		No	.	.
Gang	Input	Binary	No		No	.	.
IncidentNumber	ID	Nominal	No		No	.	.
Latitude	Rejected	Interval	No		No	.	.
Location	Rejected	Nominal	No		No	.	.
Location1	Rejected	Nominal	No		No	.	.
Longitude	Rejected	Interval	No		No	.	.
Month	Input	Nominal	No		No	.	.
Penalty	Input	Binary	No		No	.	.
PenaltyClass	Rejected	Nominal	No		No	.	.
Time	Input	Interval	No		No	.	.
TypeOfLocation	Input	Nominal	No		No	.	.
Unnamed__0	Rejected	Interval	No		No	.	.
Year	Input	Interval	No		No	.	.
ZipCode	Input	Nominal	No		No	.	.

Because records with similar properties may not be geographically close to each other, we exclude geo locations such as latitude, longitude, location. However, zip code is a nominal variable and we keep it in our analysis.

Figure-14

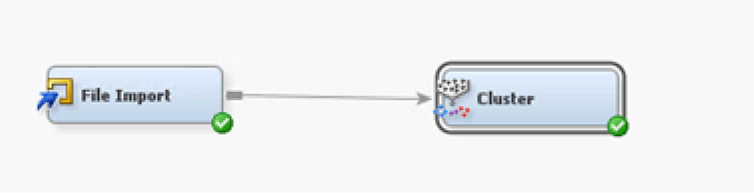
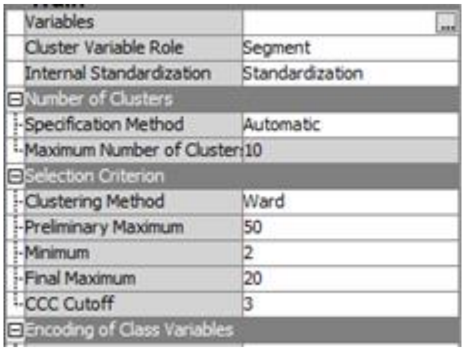


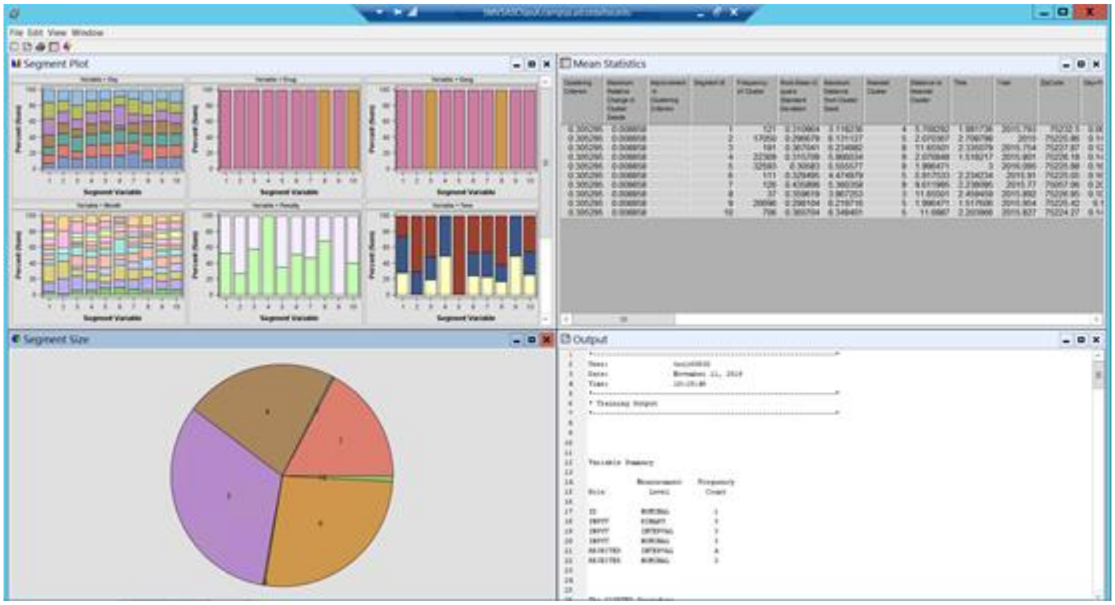
Figure-15



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	20
CCC Cutoff	3
⊟Encoding of Class Variables	

We use Ward as our Clustering Method to measure the cluster distance. The clustering analysis produces the results below:

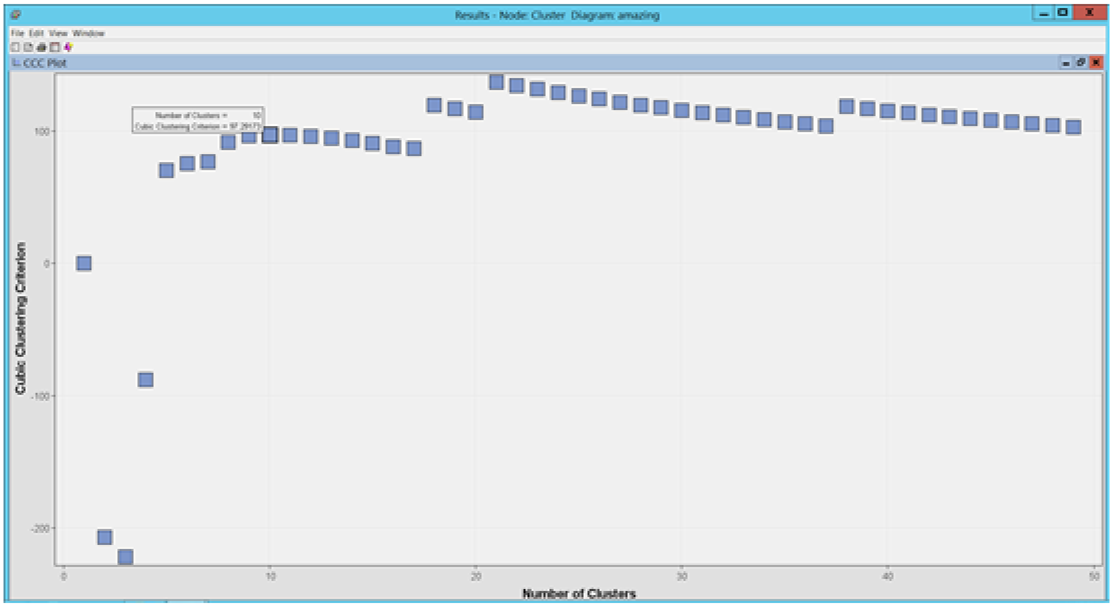
Figure-16



```
* Training Output

Variable Summary

                Measurement    Frequency
Role               Level         Count

ID                NOMINAL          1
INPUT             BINARY           1
INPUT             INTERVAL         3
INPUT             NOMINAL          1
REJECTED          INTERVAL         4
REJECTED          NOMINAL          1
```

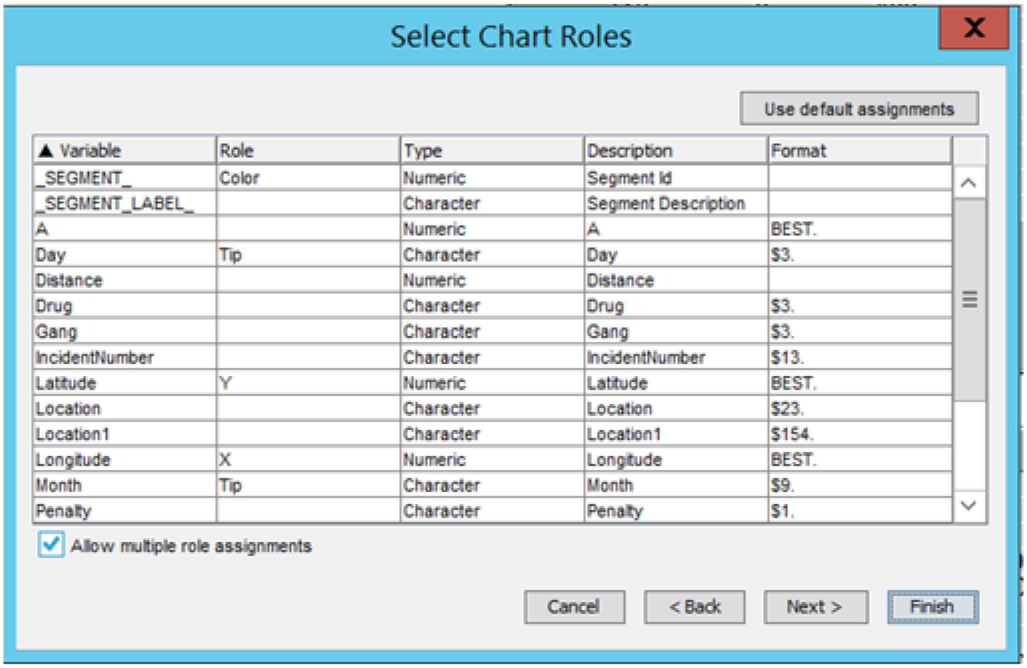
10 clusters are created by this process. However, in order to make sure the number 10 is the optimal number of k, we check the CCC plot.

Figure-17



Based on CCC plot, we can see that there is a peak at 10 and CCC value of 10 is a positive number. Therefore, we can conclude that 10 is the optimal number for clustering.
Although the results show us a lot of information about this analysis, the information is not clear enough. We use Graph Wizard to generate plots of accidents data based on latitude and longitude.

Figure-18

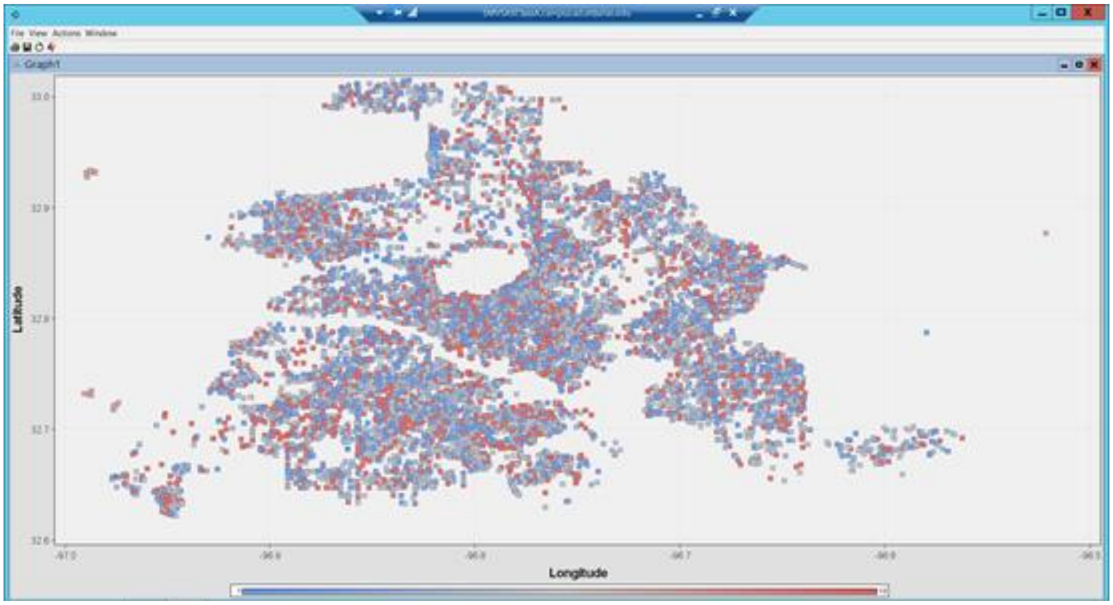


▲ Variable	Role	Type	Description	Format
SEGMENT	Color	Numeric	Segment Id	
_SEGMENT_LABEL_		Character	Segment Description	
A		Numeric	A	BEST.
Day	Tip	Character	Day	$3.
Distance		Numeric	Distance	
Drug		Character	Drug	$3.
Gang		Character	Gang	$3.
IncidentNumber		Character	IncidentNumber	$13.
Latitude	Y	Numeric	Latitude	BEST.
Location		Character	Location	$23.
Location1		Character	Location1	$154.
Longitude	X	Numeric	Longitude	BEST.
Month	Tip	Character	Month	$9.
Penalty		Character	Penalty	$1.

☑ Allow multiple role assignments

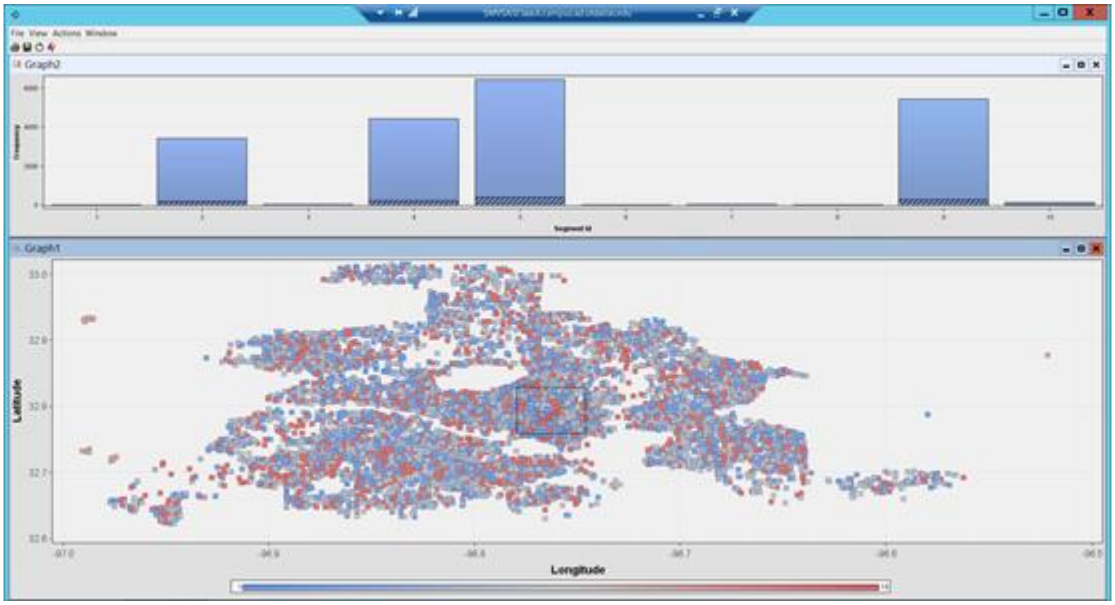
Graph Wizard produces a plot below:

Figure-19



Each square represents an accident and its color shows the cluster segment.
In order to make the scatter plot more interpretable, we add a histogram of the segment and combine it with the plot. By selecting an area on the map, the cluster segments in this area are highlighted in the histogram, which will be easier for us to understand the patterns and insights of the data distribution.

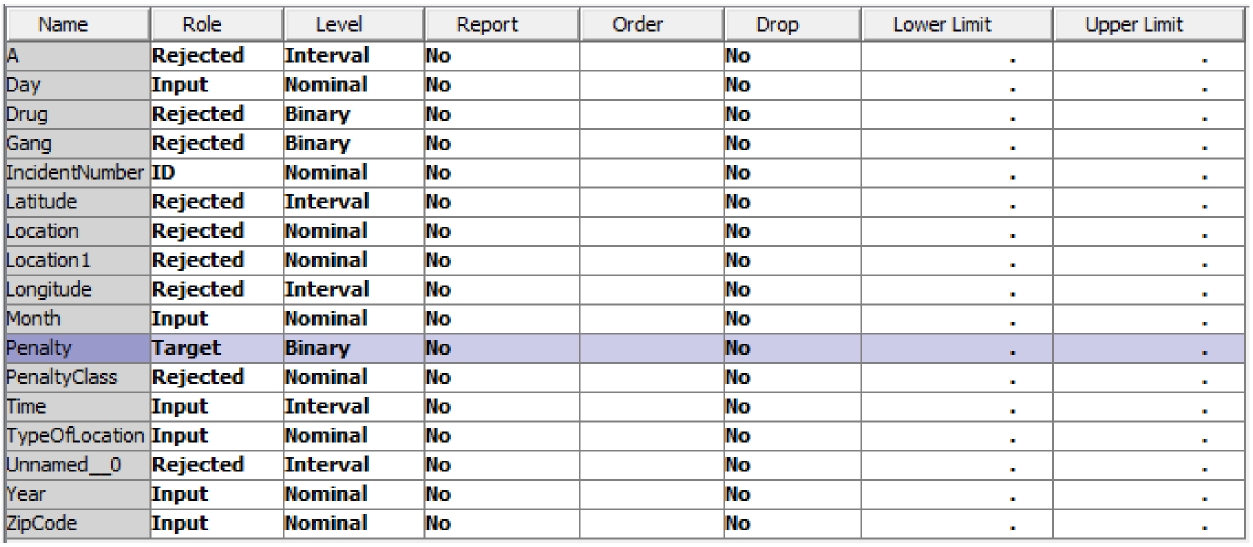
Figure-20



Predictive Analysis

The inputs are the independent variables that are influential to the target which is considered to be the dependent variable. In our case, we aim to find out what factors would lead to a penalty class F and what factors would lead to a penalty class M. Hence, we can predict what kind of crime (target) might have the highest possibility of occurrence under certain circumstance (inputs). In our case, the influential factors would be the year, the month of the year, the day of the week, location (apartment, highway, office building, etc.), and time (Section 1 as 12 am -8 am, Section 2 as 8 am – 4 pm, Section 3 as 4 pm-12 am).

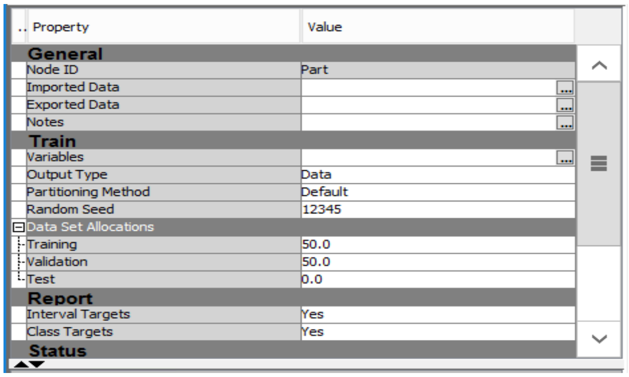
Figure-21



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
A	Rejected	Interval	No		No	.	.
Day	Input	Nominal	No		No	.	.
Drug	Rejected	Binary	No		No	.	.
Gang	Rejected	Binary	No		No	.	.
IncidentNumber	ID	Nominal	No		No	.	.
Latitude	Rejected	Interval	No		No	.	.
Location	Rejected	Nominal	No		No	.	.
Location1	Rejected	Nominal	No		No	.	.
Longitude	Rejected	Interval	No		No	.	.
Month	Input	Nominal	No		No	.	.
Penalty	Target	Binary	No		No	.	.
PenaltyClass	Rejected	Nominal	No		No	.	.
Time	Input	Interval	No		No	.	.
TypeOfLocation	Input	Nominal	No		No	.	.
Unnamed__0	Rejected	Interval	No		No	.	.
Year	Input	Nominal	No		No	.	.
ZipCode	Input	Nominal	No		No	.	.

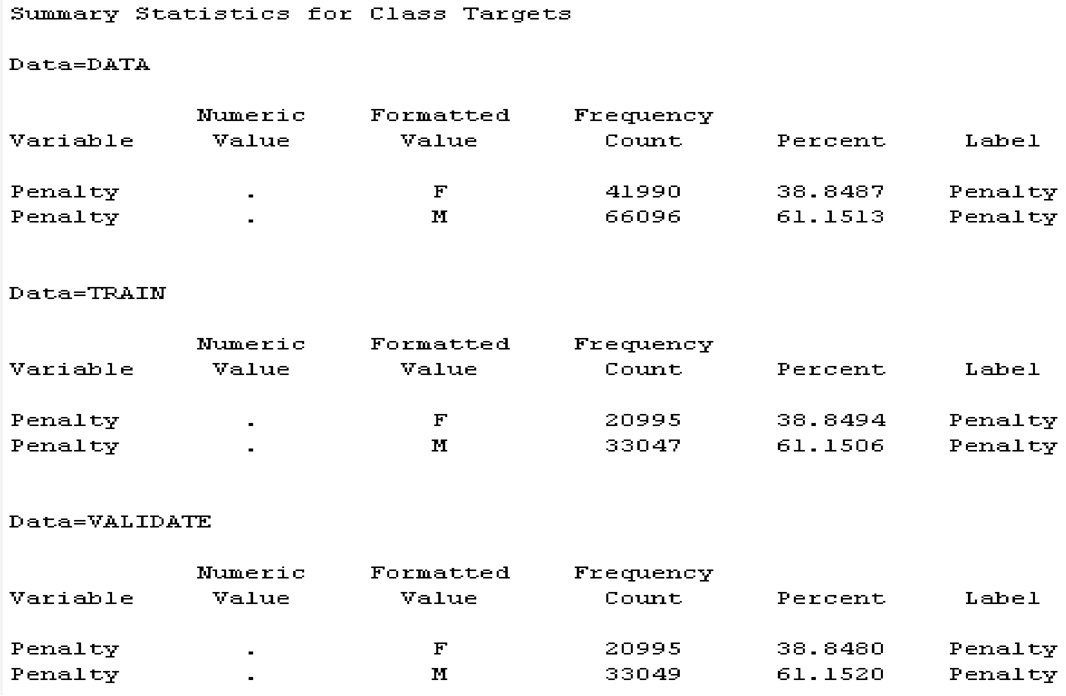
Predictive analysis always starts with a training data set. The analysis that predicts a target from specific inputs can be easily generated by a group of given training data. To avoid overfitting or underfitting, we calibrate the model's performance with a validation data set. At the same time, the test partition, which is used only for calculating the fit statistics after finishing the model selection and modeling process, is regarded as a waste of data. Therefore, we set 50% of the data as the training value and the other 50% as the validation value for the data partition, while leaving the test value as 0%.

Figure-22



Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
⊟Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	

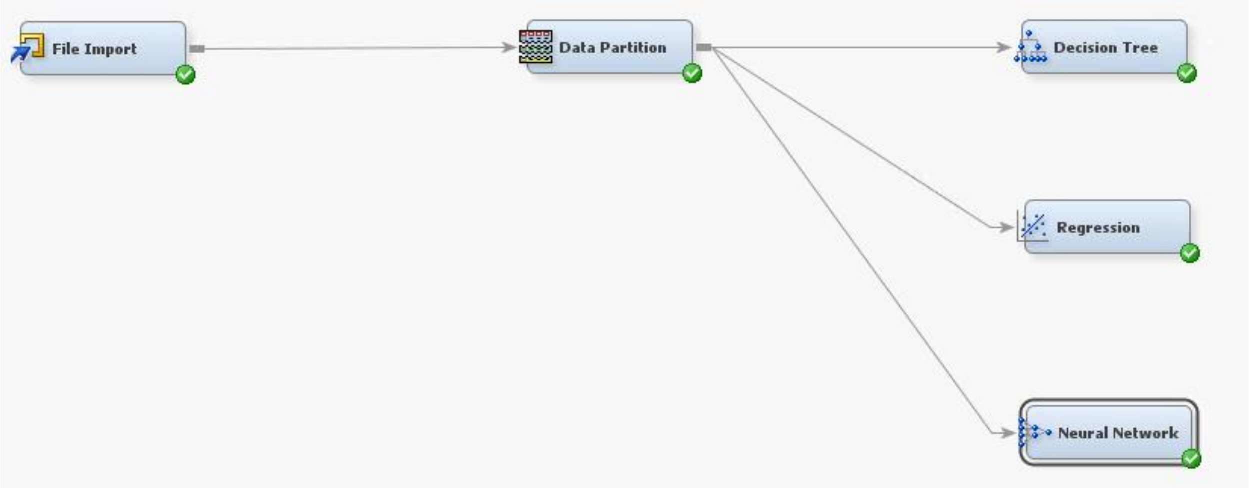
Figure-23



```
Summary Statistics for Class Targets

Data=DATA

              Numeric     Formatted     Frequency
Variable       Value        Value         Count        Percent      Label

Penalty          .            F           41990        38.8487     Penalty
Penalty          .            M           66096        61.1513     Penalty


Data=TRAIN

              Numeric     Formatted     Frequency
Variable       Value        Value         Count        Percent      Label

Penalty          .            F           20995        38.8494     Penalty
Penalty          .            M           33047        61.1506     Penalty


Data=VALIDATE

              Numeric     Formatted     Frequency
Variable       Value        Value         Count        Percent      Label

Penalty          .            F           20995        38.8480     Penalty
Penalty          .            M           33049        61.1520     Penalty
```

Based on the settings above, we imply and compare three different kinds of predictive modeling: Decision Trees, Regressions, and Neural Networks.

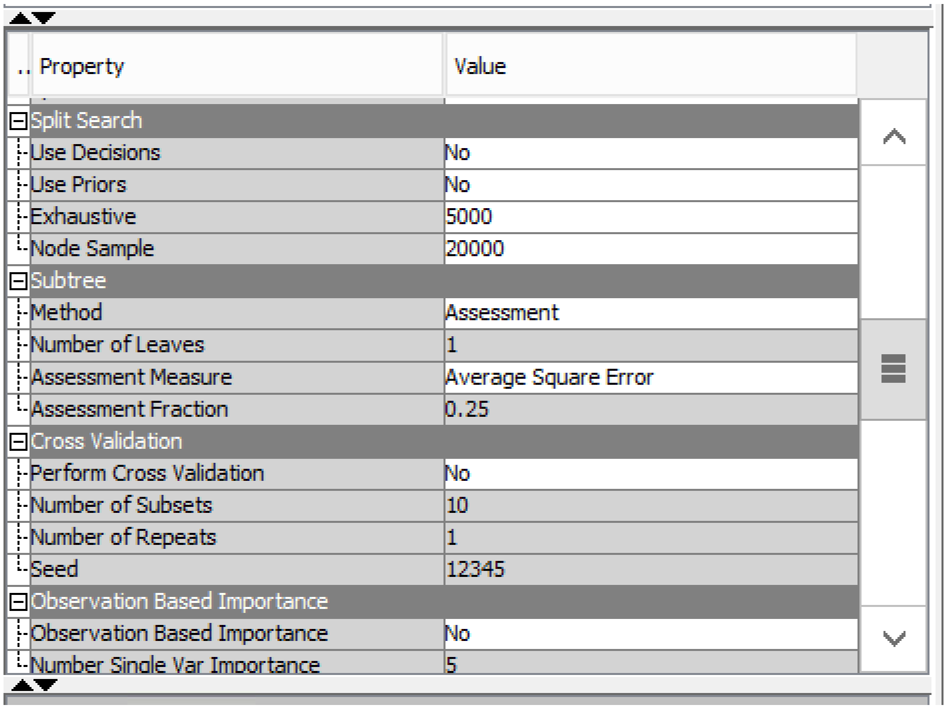
Figure-24



Decision Trees

The original decision tree analysis setting is displayed as Figure-25.

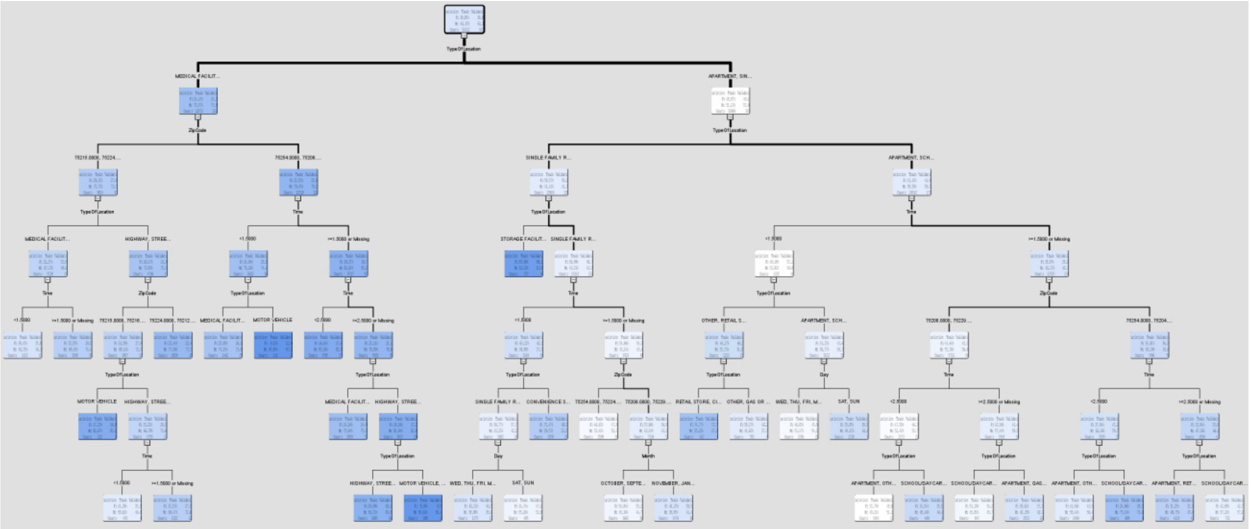
Figure-25



Property	Value
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5

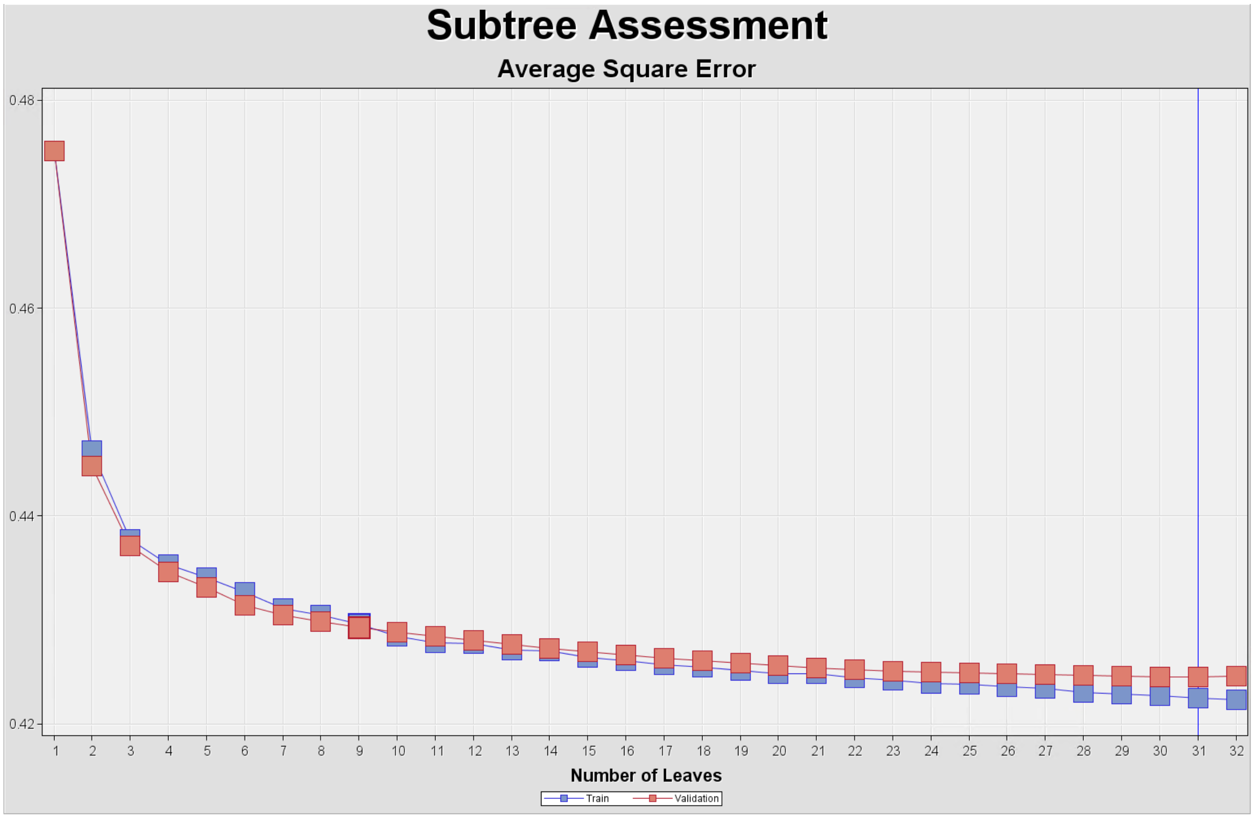
This is the decision tree we get with default settings.

Figure-26



The predicted Penalty Class values are generated by the training data whose main purpose is to select useful input variables for the first predictive model only. A diminishing marginal usefulness of input variables is expected. Therefore, under such circumstances, we need to trim down the tree.

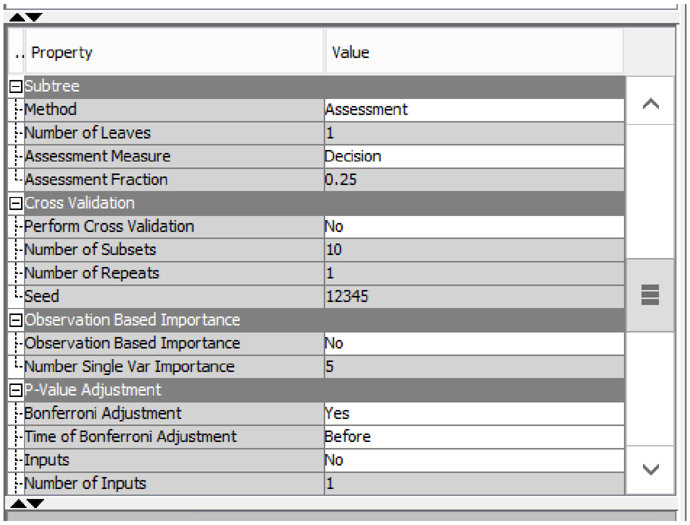
Figure-27



We can tell from the plot chart that the majority of the improvement occurs over the first few splits. There are no significant changes after that. What's more, as we mentioned above, the increasing of the amount of leaves leads to diminishing marginal usefulness.

We change the assessment measure as Decision to prune the branches of the tree.

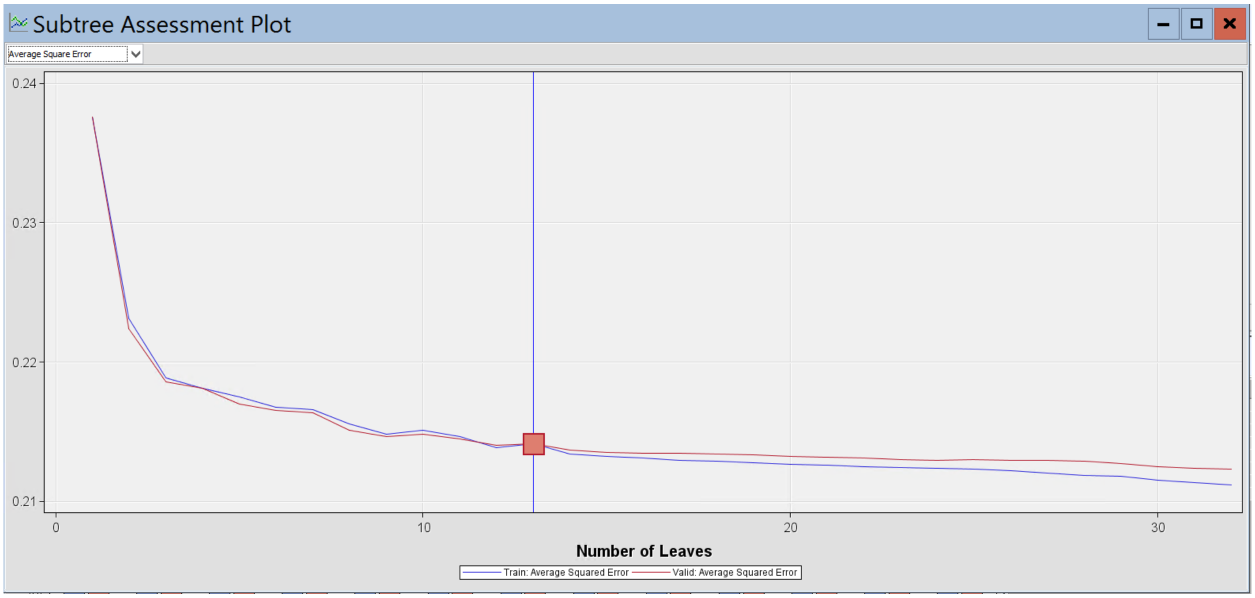
Figure-28



Property	Value
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustment	Before
Inputs	No
Number of Inputs	1

Figure-29 is the plot graph we get through viewing the Subtree Assessment by comparing the misclassification rate:

Figure-29



The plot chart states that while the amount of leaves is limited to 13, we can get the best result. Here below is the updated decision tree as the number of leaves is pruned down to 13.

Figure-30

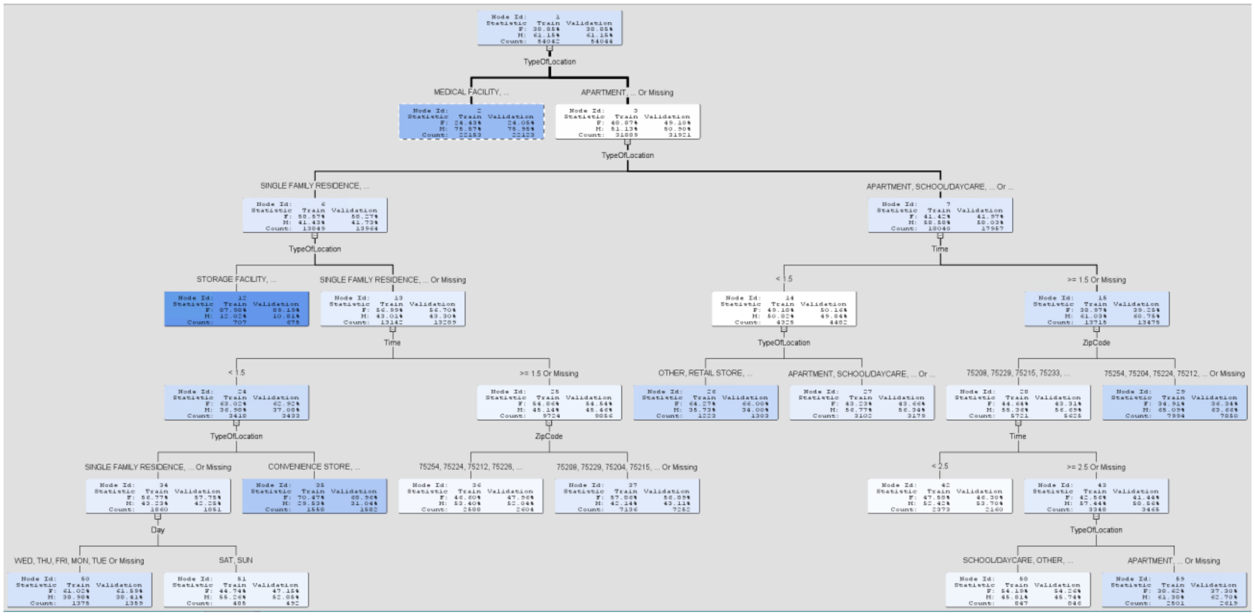
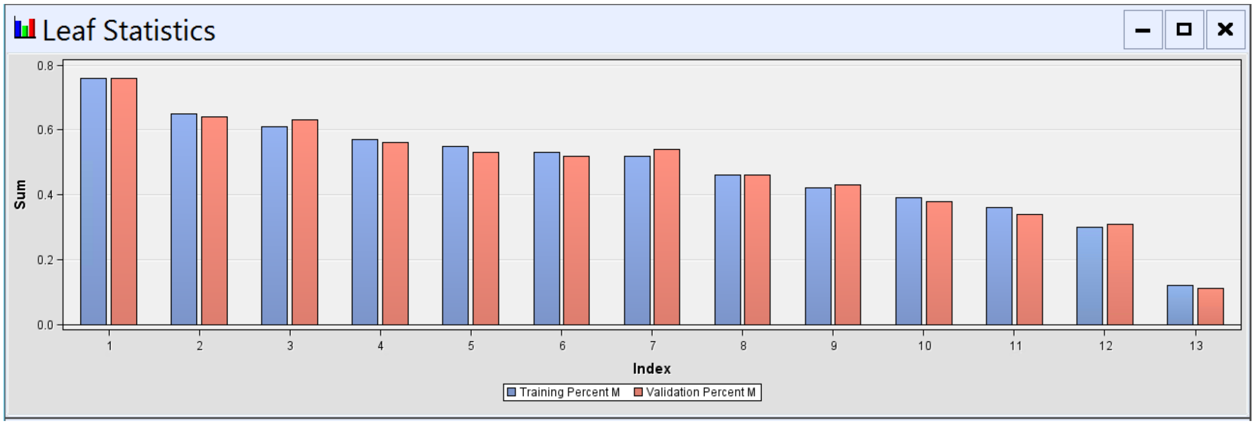

TypeOfLocation
MEDICAL FACILITY, ...
APARTMENT, ... Or Missing
TypeOfLocation
SINGLE FAMILY RESIDENCE, ...
APARTMENT, SCHOOL/DAYCARE, ... Or ...
TypeOfLocation
Time
STORAGE FACILITY, ...
SINGLE FAMILY RESIDENCE, ... Or Missing
< 1.5
>= 1.5 Or Missing
Time
TypeOfLocation
ZipCode
< 1.5
>= 1.5 Or Missing
OTHER, RETAIL STORE, ...
APARTMENT, SCHOOL/DAYCARE, ... Or ...
75208, 75229, 75215, 75233, ...
75254, 75204, 75224, 75212, ... Or Missing
TypeOfLocation
ZipCode
Time
SINGLE FAMILY RESIDENCE, ... Or Missing
CONVENIENCE STORE, ...
75254, 75224, 75212, 75226, ...
75208, 75229, 75204, 75215, ... Or Missing
< 2.5
>= 2.5 Or Missing
Day
TypeOfLocation
WED, THU, FRI, MON, TUE Or Missing
SAT, SUN
SCHOOL/DAYCARE, OTHER, ...
APARTMENT, ... Or Missing

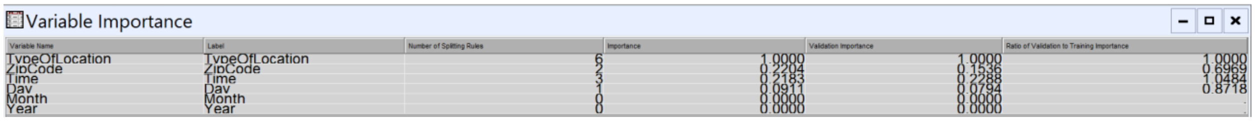
Figure-31 is the Leaf Statistics Bar Chart. If the height difference between training data and validation data is small, we can conclude that the model functions well.

Figure-31



Meanwhile, we can tell from the Variable Importance table that the Type of Location is the most influential factor to the Penalty Class type, followed by the Zip Code and Time. In the Decision Trees analysis, Year has no impacts on the penalty class.

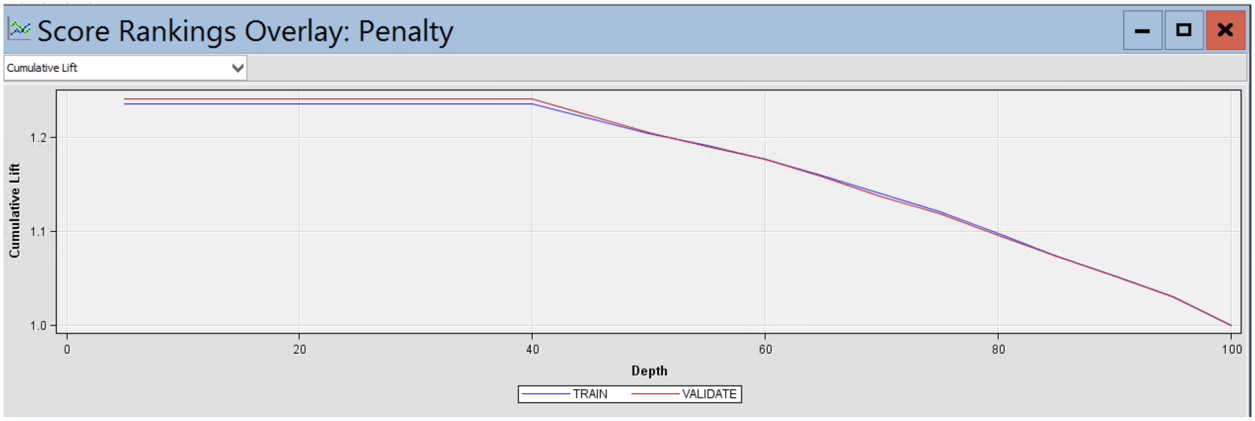
Figure-32


Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
TypeOfLocation	TypeOfLocation	6	1.0000	1.0000	1.0000
ZipCode	ZipCode	2	0.2204	0.1536	0.6969
Time	Time	3	0.2183	0.2288	1.0484
Day	Day	1	0.0911	0.0794	0.8718
Month	Month	0	0.0000	0.0000	.
Year	Year	0	0.0000	0.0000	.

The Score Rankings Overlay plot indicates that this model is a useful model because it shows high cumulative lifts in both the training and validation data.

Figure-33

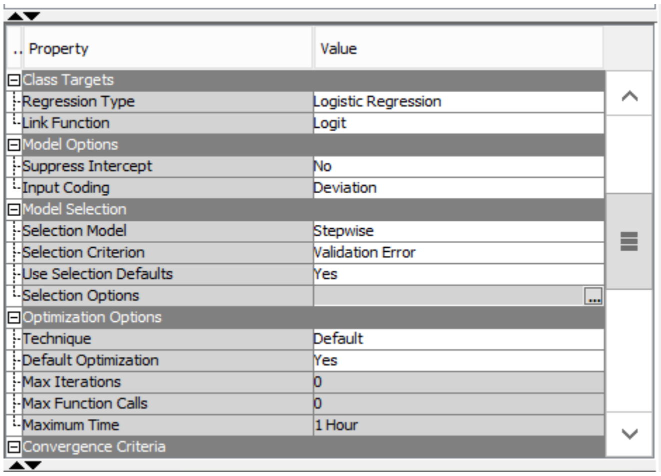


Regressions

Regressions modeling offers a different aspect to predict the results compared to decision trees modeling. Regressions modeling, as a parametric analysis, assumes a specific association structure between inputs and target. However, Decision Trees, as predictive algorithms, do not assume any association structure.

While running the Regressions modeling, we set the Selection Model as Stepwise and the Selection Criterion as Validation Error.

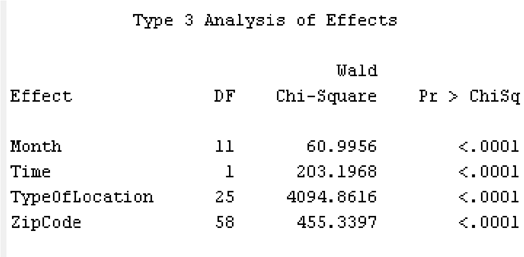
Figure-34



Property	Value
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
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	

The result of Regressions modeling is listed in Figure-35.

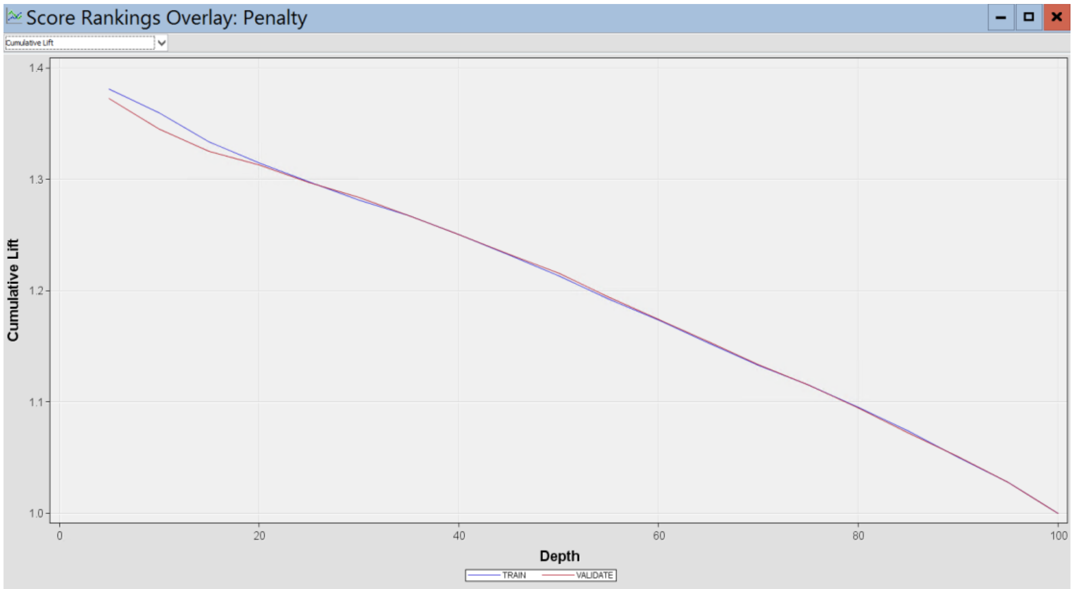
Figure-35



```
             Type 3 Analysis of Effects

                                  Wald
Effect              DF    Chi-Square    Pr > ChiSq

Month               11       60.9956        <.0001
Time                 1      203.1968        <.0001
TypeOfLocation      25     4094.8616        <.0001
ZipCode             58      455.3397        <.0001
```

The Score Rankings Overlay plot indicates that this model is a useful model because it shows high cumulative lifts in both the training and validation data.

Figure-36

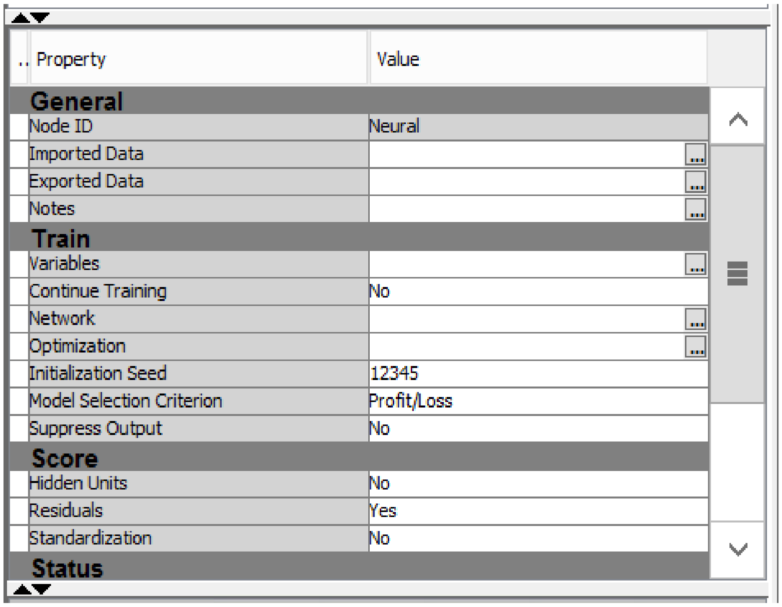


In this case, the result can be used for prediction and it indicates that Month, Time, Type of Location, and Zip Code has decreasing impacts on the penalty class.

Neural Networks

Applying the neural networks modeling under the Stepwise selection model on the inputs to analyze the penalty class, we enable Preliminary Training and take the default maximum iterations as 50 to run the first round of analysis.

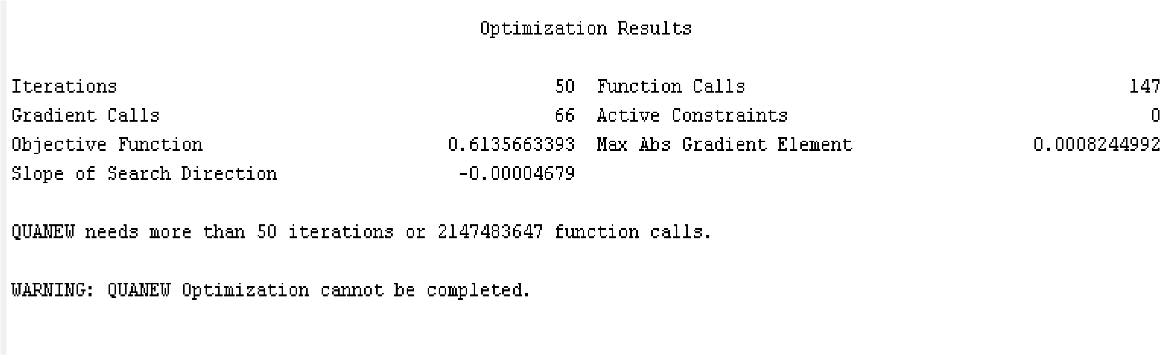
Figure-37



Property	Value
General	
Node ID	Neural
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Continue Training	No
Network	...
Optimization	...
Initialization Seed	12345
Model Selection Criterion	Profit/Loss
Suppress Output	No
Score	
Hidden Units	No
Residuals	Yes
Standardization	No
Status	

Here is what we get.

Figure-38

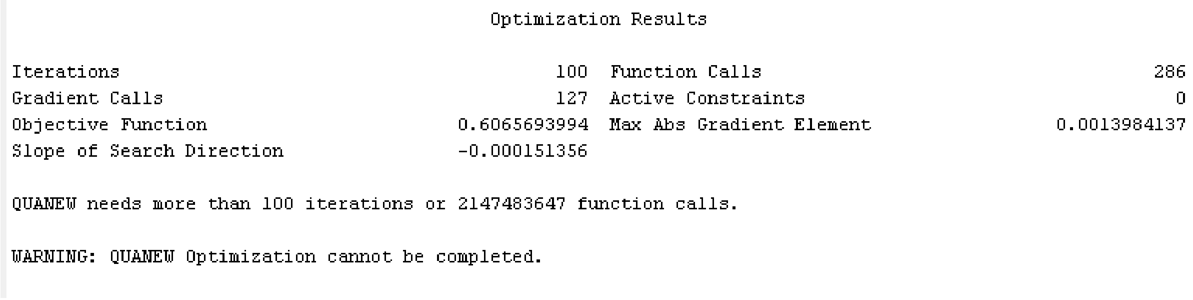


```
                                    Optimization Results

Iterations                                  50  Function Calls                        147
Gradient Calls                              66  Active Constraints                      0
Objective Function                0.6135663393  Max Abs Gradient Element     0.0008244992
Slope of Search Direction          -0.00004679

QUANEW needs more than 50 iterations or 2147483647 function calls.

WARNING: QUANEW Optimization cannot be completed.
```

Notice the warning message. It indicates that data do not fit the model well, because there are too many poorly fitting observations.

In the second round analysis, we set the maximum iterations as 100. However, the same warning appears again as in Figure-39.

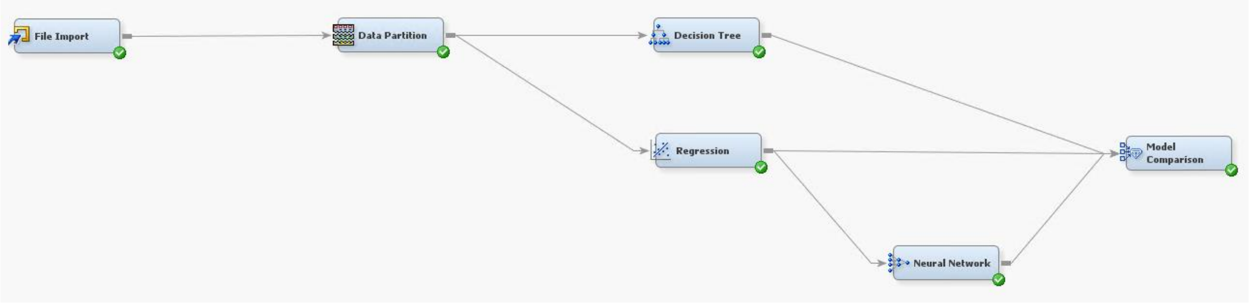
Figure-39



```
                                          Optimization Results

Iterations                                        100  Function Calls                                  286
Gradient Calls                                    127  Active Constraints                                0
Objective Function                       0.6065693994  Max Abs Gradient Element               0.0013984137
Slope of Search Direction                -0.000151356

QUANEW needs more than 100 iterations or 2147483647 function calls.

WARNING: QUANEW Optimization cannot be completed.
```

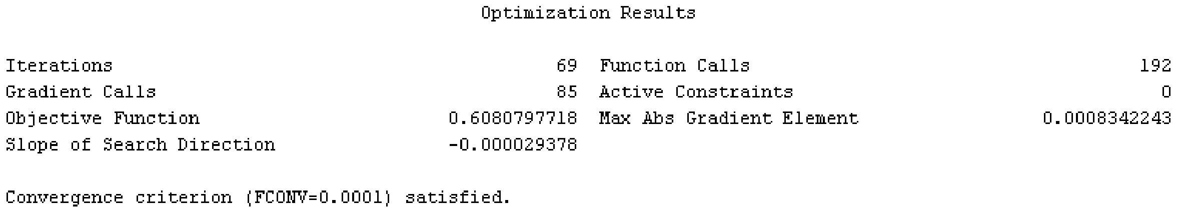
Therefore, we connect the Neural Networks node right after the Regression node in SAS EM. Figure-40 is the new route and the related result we get.

Figure-40



Under such setting, the modeling process functions well, which is shown in Figure-41.

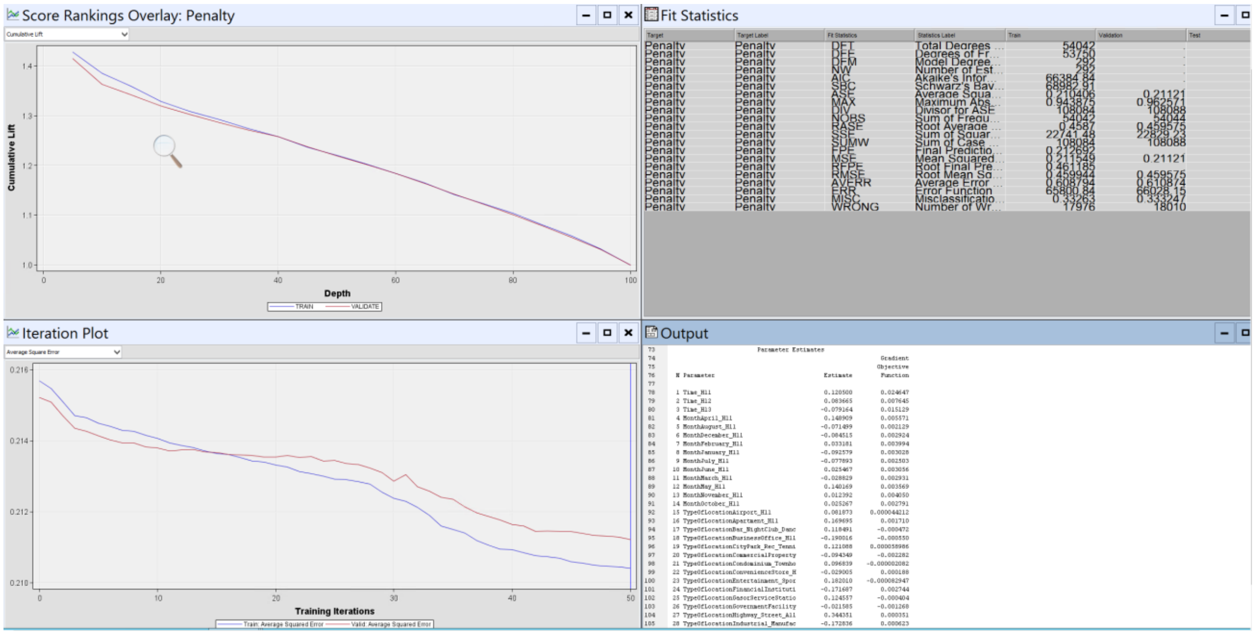
Figure-41



```
                                          Optimization Results

Iterations                                         69  Function Calls                                  192
Gradient Calls                                     85  Active Constraints                                0
Objective Function                       0.6080797718  Max Abs Gradient Element               0.0008342243
Slope of Search Direction                -0.000029378

Convergence criterion (FCONV=0.0001) satisfied.
```

Figure-42 is the result of the Neural Networks analysis.

Figure-42

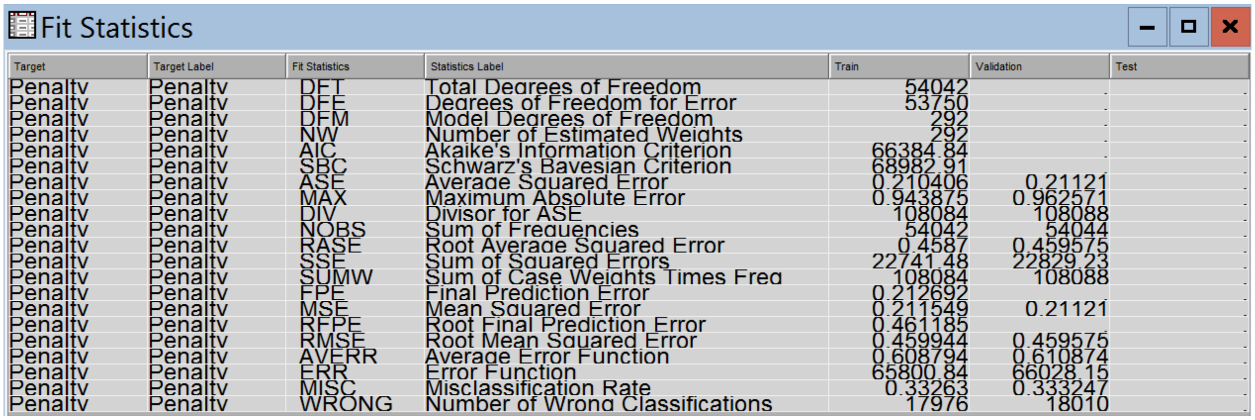


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Penalty	Penalty	DFT	Total Degrees ...	54042	.	
Penalty	Penalty	DFE	Degrees of Fr...	53750	.	
Penalty	Penalty	DFM	Model Degree...	292	.	
Penalty	Penalty	NW	Number of Est...	292	.	
Penalty	Penalty	AIC	Akaike's Infor...	66384.84	.	
Penalty	Penalty	SBC	Schwarz's Bay...	68982.91	.	
Penalty	Penalty	ASE	Average Squa...	0.210406	0.21121	
Penalty	Penalty	MAX	Maximum Abs...	0.943875	0.962571	
Penalty	Penalty	DIV	Divisor for ASE	108084	108088	
Penalty	Penalty	NOBS	Sum of Frequ...	54042	54044	
Penalty	Penalty	RASE	Root Average ...	0.4587	0.459575	
Penalty	Penalty	SSE	Sum of Squar...	22741.48	22829.23	
Penalty	Penalty	SUMW	Sum of Case ...	108084	108088	
Penalty	Penalty	FPE	Final Predictio...	0.212692	.	
Penalty	Penalty	MSE	Mean Squared...	0.211549	0.21121	
Penalty	Penalty	RFPE	Root Final Pre...	0.461185	.	
Penalty	Penalty	RMSE	Root Mean Sq...	0.459944	0.459575	
Penalty	Penalty	AVERR	Average Error ...	0.608794	0.610874	
Penalty	Penalty	ERR	Error Function	65800.84	66028.15	
Penalty	Penalty	MISC	Misclassificatio...	0.33263	0.333247	
Penalty	Penalty	WRONG	Number of Wr...	17976	18010	

```
                              Parameter Estimates

                                                                  Gradient
                                                                 Objective
  N Parameter                                   Estimate          Function

  1 Time_H11                                    0.120500          0.024647
  2 Time_H12                                    0.083665          0.007645
  3 Time_H13                                   -0.079164          0.015129
  4 MonthApril_H11                              0.148909          0.005571
  5 MonthAugust_H11                            -0.071499          0.002129
  6 MonthDecember_H11                          -0.004515          0.002924
  7 MonthFebruary_H11                           0.033181          0.003994
  8 MonthJanuary_H11                           -0.092579          0.003028
  9 MonthJuly_H11                              -0.077893          0.002503
 10 MonthJune_H11                               0.025467          0.003056
 11 MonthMarch_H11                             -0.028829          0.002931
 12 MonthMay_H11                                0.140169          0.003569
 13 MonthNovember_H11                           0.012392          0.004050
 14 MonthOctober_H11                            0.025267          0.002791
 15 TypeOfLocationAirport_H11                   0.081873       0.000044212
 16 TypeOfLocationApartment_H11                 0.169695          0.001710
 17 TypeOfLocationBar_NightClub_Danc             0.118491         -0.000472
 18 TypeOfLocationBusinessOffice_H11           -0.190016         -0.000550
 19 TypeOfLocationCityPark_Rec_Tenni            0.121088       0.000058986
 20 TypeOfLocationCommercialProperty           -0.094349         -0.002282
 21 TypeOfLocationCondominium_Townho            0.096839      -0.000002082
 22 TypeOfLocationConvenienceStore_H           -0.029005          0.000188
 23 TypeOfLocationEntertainment_Spor            0.182010      -0.000082947
 24 TypeOfLocationFinancialInstituti           -0.171687          0.002744
 25 TypeOfLocationGasorServiceStatio            0.124557         -0.000404
 26 TypeOfLocationGovernmentFacility           -0.021585         -0.001268
 27 TypeOfLocationHighway_Street_All            0.344351          0.000351
 28 TypeOfLocationIndustrial_Manufac           -0.172836          0.000623
```

The Fit Statistics window shows a good modeling performance on both the average squared error and misclassification scales.

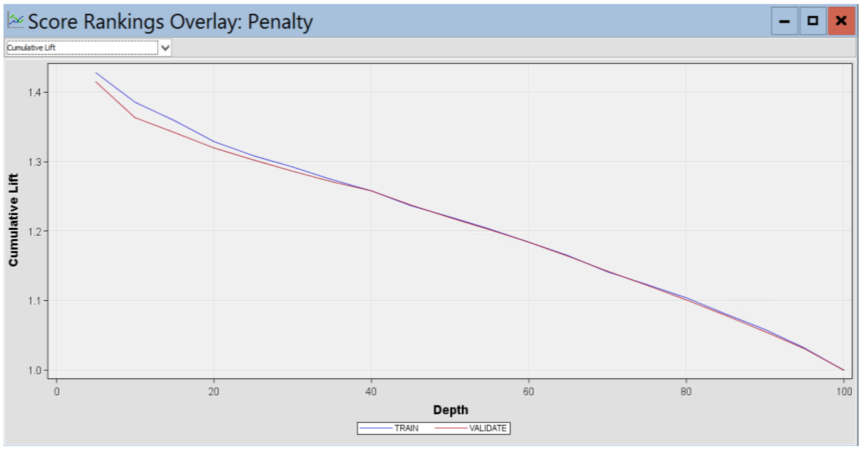
Figure-43



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Penalty	Penalty	DFT	Total Degrees of Freedom	54042	.	.
Penalty	Penalty	DFE	Degrees of Freedom for Error	53750	.	.
Penalty	Penalty	DFM	Model Degrees of Freedom	292	.	.
Penalty	Penalty	NW	Number of Estimated Weights	292	.	.
Penalty	Penalty	AIC	Akaike's Information Criterion	66384.84	.	.
Penalty	Penalty	SBC	Schwarz's Bayesian Criterion	68982.91	.	.
Penalty	Penalty	ASE	Average Squared Error	0.210406	0.21121	.
Penalty	Penalty	MAX	Maximum Absolute Error	0.943875	0.962571	.
Penalty	Penalty	DIV	Divisor for ASE	108084	108088	.
Penalty	Penalty	NOBS	Sum of Frequencies	54042	54044	.
Penalty	Penalty	RASE	Root Average Squared Error	0.4587	0.459575	.
Penalty	Penalty	SSE	Sum of Squared Errors	22741.48	22829.23	.
Penalty	Penalty	SUMW	Sum of Case Weights Times Freq	108084	108088	.
Penalty	Penalty	FPE	Final Prediction Error	0.212692	.	.
Penalty	Penalty	MSE	Mean Squared Error	0.211549	0.21121	.
Penalty	Penalty	RFPE	Root Final Prediction Error	0.461185	.	.
Penalty	Penalty	RMSE	Root Mean Squared Error	0.459944	0.459575	.
Penalty	Penalty	AVERR	Average Error Function	0.608794	0.610874	.
Penalty	Penalty	ERR	Error Function	65800.84	66028.15	.
Penalty	Penalty	MISC	Misclassification Rate	0.33263	0.333247	.
Penalty	Penalty	WRONG	Number of Wrong Classifications	17976	18010	.

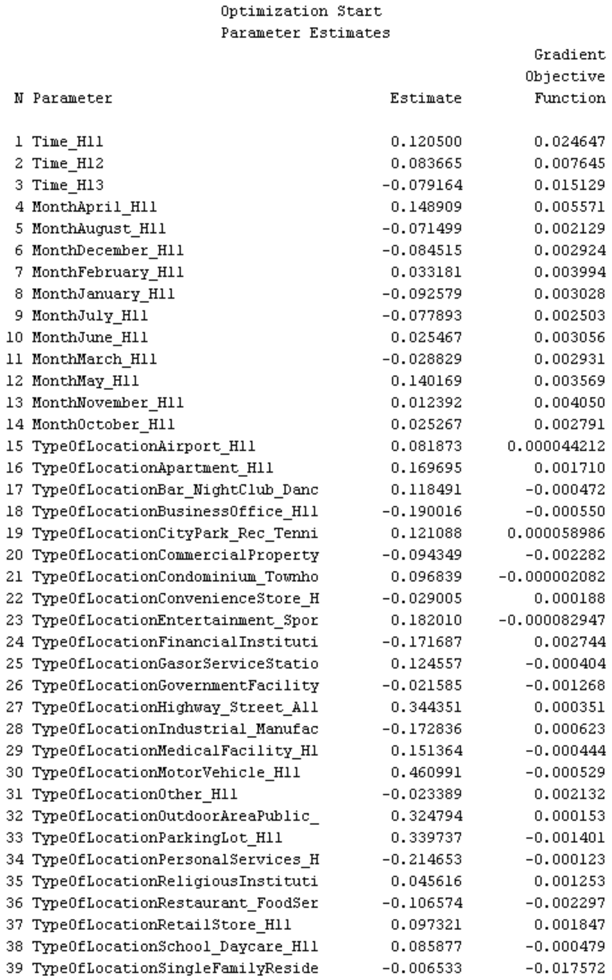
What's more, the Score Rankings Overlay plot indicates that this model is a useful model because it shows high cumulative lifts in both the training and validation data.

Figure-44



In the Output window, the most influential factors are revealed in the sequence as Time, Month, and the Type of Location, which is shown in Figure-45.

Figure-45



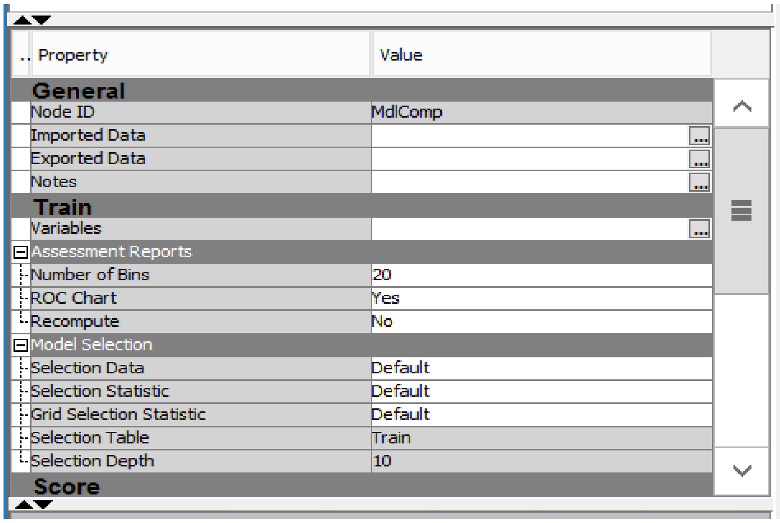
```
                         Optimization Start
                         Parameter Estimates
                                                          Gradient
                                                         Objective
  N Parameter                             Estimate        Function

  1 Time_H11                              0.120500        0.024647
  2 Time_H12                              0.083665        0.007645
  3 Time_H13                             -0.079164        0.015129
  4 MonthApril_H11                        0.148909        0.005571
  5 MonthAugust_H11                      -0.071499        0.002129
  6 MonthDecember_H11                    -0.084515        0.002924
  7 MonthFebruary_H11                     0.033181        0.003994
  8 MonthJanuary_H11                     -0.092579        0.003028
  9 MonthJuly_H11                        -0.077893        0.002503
 10 MonthJune_H11                         0.025467        0.003056
 11 MonthMarch_H11                       -0.028829        0.002931
 12 MonthMay_H11                          0.140169        0.003569
 13 MonthNovember_H11                     0.012392        0.004050
 14 MonthOctober_H11                      0.025267        0.002791
 15 TypeOfLocationAirport_H11             0.081873     0.000044212
 16 TypeOfLocationApartment_H11           0.169695        0.001710
 17 TypeOfLocationBar_NightClub_Danc      0.118491       -0.000472
 18 TypeOfLocationBusinessOffice_H11     -0.190016       -0.000550
 19 TypeOfLocationCityPark_Rec_Tenni      0.121088     0.000058986
 20 TypeOfLocationCommercialProperty     -0.094349       -0.002282
 21 TypeOfLocationCondominium_Townho      0.096839    -0.000002082
 22 TypeOfLocationConvenienceStore_H     -0.029005        0.000188
 23 TypeOfLocationEntertainment_Spor      0.182010    -0.000082947
 24 TypeOfLocationFinancialInstituti     -0.171687        0.002744
 25 TypeOfLocationGasorServiceStatio      0.124557       -0.000404
 26 TypeOfLocationGovernmentFacility     -0.021585       -0.001268
 27 TypeOfLocationHighway_Street_All      0.344351        0.000351
 28 TypeOfLocationIndustrial_Manufac     -0.172836        0.000623
 29 TypeOfLocationMedicalFacility_H1      0.151364       -0.000444
 30 TypeOfLocationMotorVehicle_H11        0.460991       -0.000529
 31 TypeOfLocationOther_H11              -0.023389        0.002132
 32 TypeOfLocationOutdoorAreaPublic_      0.324794        0.000153
 33 TypeOfLocationParkingLot_H11          0.339737       -0.001401
 34 TypeOfLocationPersonalServices_H     -0.214653       -0.000123
 35 TypeOfLocationReligiousInstituti      0.045616        0.001253
 36 TypeOfLocationRestaurant_FoodSer     -0.106574       -0.002297
 37 TypeOfLocationRetailStore_H11         0.097321        0.001847
 38 TypeOfLocationSchool_Daycare_H11      0.085877       -0.000479
 39 TypeOfLocationSingleFamilyReside     -0.006533       -0.017572
```

Comparison

Above all, we have completed three predictive analyzing models, which are Decision Trees, Regressions, and Neural Networks. All of them are logical and useful. We need to conduct a comparative analysis among these models.

The comparison default settings are displayed as Figure-46.

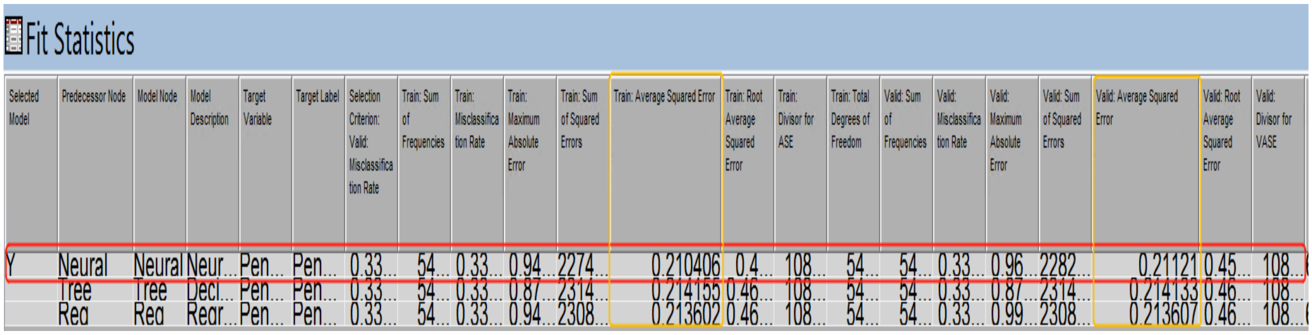
Figure-46



Property	Value
General	
Node ID	MdlComp
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟ Assessment Reports	
Number of Bins	20
ROC Chart	Yes
Recompute	No
⊟ Model Selection	
Selection Data	Default
Selection Statistic	Default
Grid Selection Statistic	Default
Selection Table	Train
Selection Depth	10
Score	

We can find out the best model by comparing the values of Average Square Error. In this comparison, Neural Networks modeling performs the best because of its lowest value of Average Square Error in both Training data and Validation data.

Figure-47



Fit Statistics

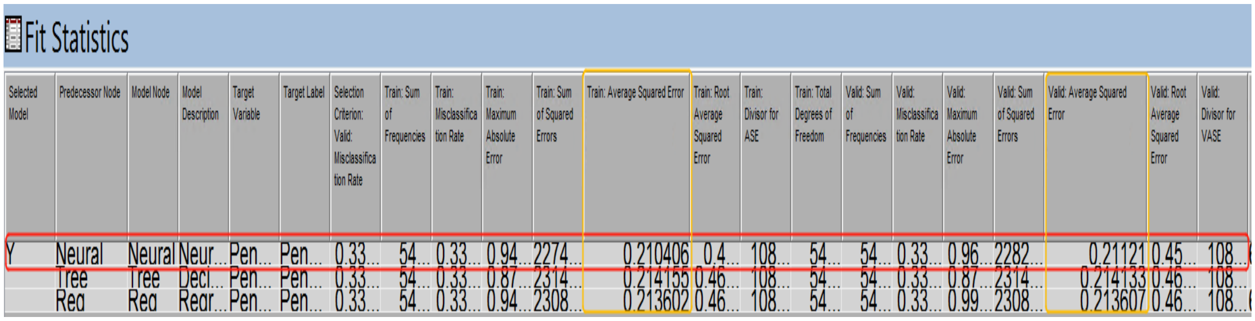
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Average Squared Error	Valid: Root Average Squared Error	Valid: Divisor for VASE
Y	Neural	Neural	Neur...	Pen...	Pen...	0.33...	54...	0.33...	0.94...	2274...	0.210406	0.4...	108...	54...	54...	0.33...	0.96...	2282...	0.21121	0.45...	108...
	Tree	Tree	Deci...	Pen...	Pen...	0.33...	54...	0.33...	0.87...	2314...	0.214155	0.46...	108...	54...	54...	0.33...	0.87...	2314...	0.214133	0.46...	108...
	Reg	Reg	Regr...	Pen...	Pen...	0.33...	54...	0.33...	0.94...	2308...	0.213602	0.46...	108...	54...	54...	0.33...	0.99...	2308...	0.213607	0.46...	108...

Conclusion

Due to algorithm difference, three models are focusing on different points and the results might vary. In our research, the three models demonstrate similar levels of accuracy in predicting violent reconviction. For Neural Networks, the most important indicator is Time, followed by Month factor then followed by the Type of Location. For Regressions analysis, Month is the most important factor toward the crime analysis, and for Decision Trees, the most influential factor is Type of Location.

From the comparison result, we can tell that the neural network performed the best among other methodologies.

Figure-48



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: Misclassification Rate	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Average Squared Error	Valid: Root Average Squared Error	Valid: Divisor for VASE
Y	Neural	Neural	Neur...	Pen...	Pen...	0.33...	54...	0.33...	0.94...	2274...	0.210406	0.4...	108...	54...	54...	0.33...	0.96...	2282...	0.21121	0.45...	108...
	Tree	Tree	Deci...	Pen...	Pen...	0.33...	54...	0.33...	0.87...	2314...	0.214155	0.46...	108...	54...	54...	0.33...	0.87...	2314...	0.214133	0.46...	108...
	Reg	Reg	Regr...	Pen...	Pen...	0.33...	54...	0.33...	0.94...	2308...	0.213602	0.46...	108...	54...	54...	0.33...	0.99...	2308...	0.213607	0.46...	108...

From 5 a.m. the accident occurrence increases gradually and reach a periodic peak at 12 p.m. From 4 p.m. to 12 a.m., accidents occurrence keeps in high level. If you want to travel outside the Dallas County, the wise choice is to get rid of these two time periods.

A reverse relationship between income and crime rate is tested in our research: higher household income areas usually come with lower crime rates. In the future if your financial budget is allowed, renting an apartment located at higher income area could help you get away from crimes. For the police department, the lower household income areas require more police resource especially around 75204 and 75206 districts.

In a word, there are no good or right models among Clustering, Decision Tree and Regressions analysis, and which model will be most helpful to the research is different. In this report, we firmly believe that the Neural Network provides us with the best result.

Take Away

Based on the research of our study, we would like to recommend the following tips for future explorations:

1. Increasing variables: the more variables the raw data has, the more accurate the analysis model will be.
2. Other comparison measures should be considered: Average Square Error is the only method we applied to compare between different models. Other measures such as AUC (area under the curve) value might apply as well.

3. Different methodologies have different advantages and disadvantages, from our perspective the neural network model gave us the best result, but it did not denote that the Neural Network will be working well for any other areas. To figure out which method is the best fit, we suggest the scholar treat the specific topic with the most fitting method by using their own judgments.