FINAL PROJECT
OPIM 5604 – SPRING 2017

Predicting the nature of a crime in San Francisco

The work contained and presented here is our work and our work alone

TEAM #3

CLASS SECTION – EVENING

SECB12 – 1173

TEAM MEMBERS:

Sapanjeet Singh Chatwal
Suhas Nadiga
Sreekanth Kyatham
Varnika Yertha

San Francisco as a city is known more for its technology than its criminal background. But, with the increase seen in wealth inequality, housing shortages, and increase in expensive digital toys driving BART to work, there is an increase in crime in the city by the bay**.**

This project was conducted on the dataset obtained from crime incidents derived from the San Francisco Police Department Crime Reporting System. The dataset depicts crimes that have taken place in San Francisco at a specific time and location. Records considered range from 1/1/2003 to 5/13/2015.

The goal is to predict whether a crime is violent given its time and location.

Objectives

I. Data pre-processing
II. Pattern Discovery and Visualization
III. Modelling
    a. Model Trials
    b. Model results and comparison
    c. Final Implementation
IV. Conclusions
V. Appendix

Data Pre-processing

Following is the list of columns that we have in the original dataset as well as derived columns. We have listed the description as well as the changes made to each column.

1. Dates - Timestamp of the crime incident
2. Category – The type of crime that occurred (Assault, Theft etc)
3. Description - detailed description of the crime incident
4. PdDistrict - name of the Police Department District
5. Resolution - how the crime incident was resolved
6. Address - the approximate street address of the crime incident
7. X – (changed to **Longitude**) – Modified the column format by changing the data format to Longitude DDD under Geographic Formats
8. Y – (changed to **Latitude**) - Modified the column format by changing the data format to Latitude DDD under Geographic Formats

Derived columns

9. DayOfWeek - the day of the week
10. **Crime Class** – This is the dependent variable (Violent/Non violent).

Stratification:

We stratified the dataset on Date, PdDistrict and Crime Class variables by allotting 60% as training, 40% as validation.

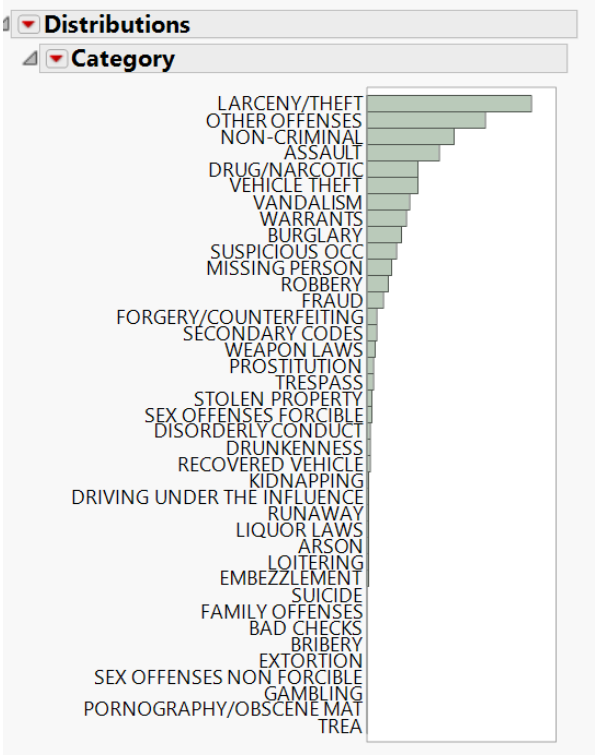
Predictors:

From the dataset, we observed that variables like Category, Description and Resolution are updated only after the incident occurs, and hence should not be used in the prediction. Latitude and Longitude are not continuous variables and could only be used for visualization in their current format. Hence, PdDistrict and Date were the main variables that gave some explanatory value. The reason for Latitude and Longitude not continuous variables is because the computation used to calculate latitude and Longitude involves more of spatial mathematics (Haversian Formula) ignoring ellipsoidal effect. The variables as classified by jmp are not correct and we prefer to use them for visualizing the data Hence, we ignored latitude and longitude as predictors.

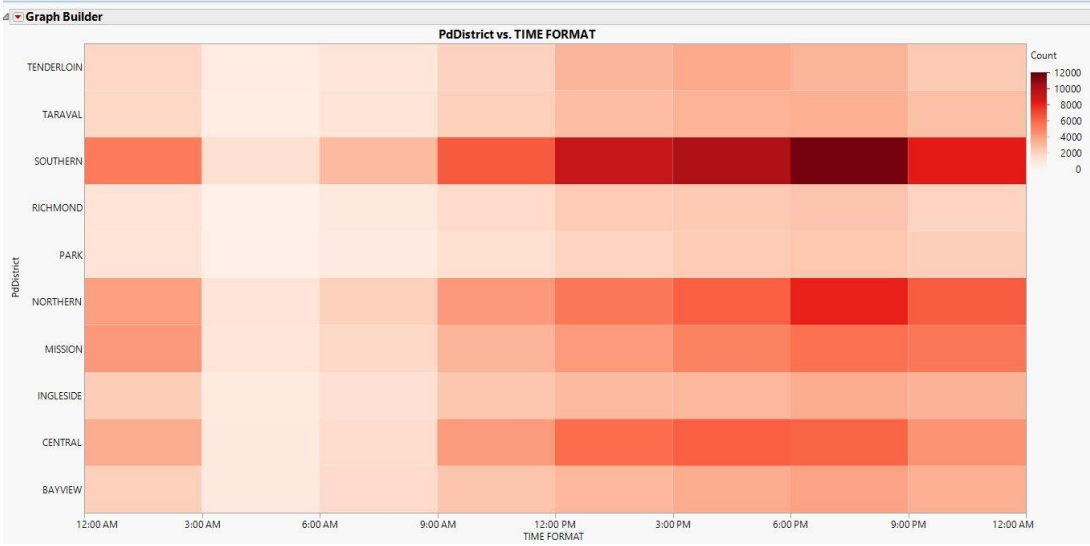
Pattern Discovery and Visualization

The initial dataset contained ~800,00 records spread across 39 unique categories of crime and 10 police districts. We extracted Year, Month, Day and Hour as individual columns to explore patterns at this level. However, variation was not meaningful at year or month level.

1. The top 5 crime categories overall were Larceny/Theft, Other Offenses, Non-Criminal, Assault and Drug/Narcotics.

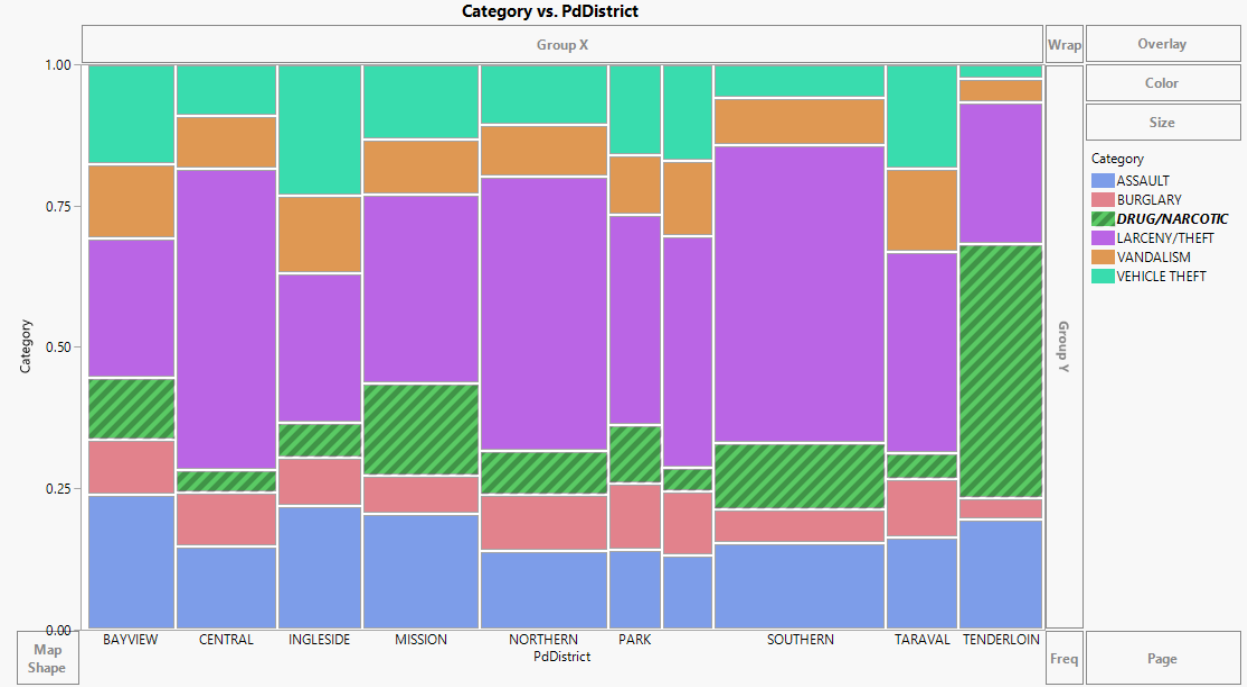


2. The PD District Southern had the highest number of crimes overall, followed by Northern. After normalizing for the total amount of crime, crime activity by time is similar across all districts. We see relatively less crime overall in the early hours of 3:00 AM – 9:00 AM.

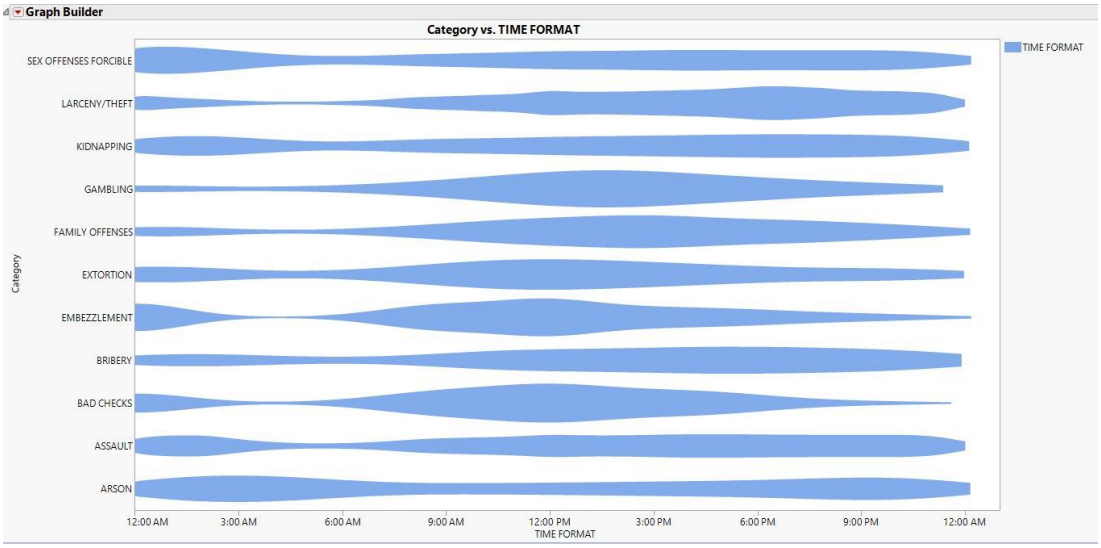


At the aggregate level, variability in time and location is not seen clearly. For visualizing and uncovering patterns we chose to depict only some categories of interest with a high number of records - Assault, Larceny/Theft, Drug/Narcotic, Vandalism, Burglary and Vehicle Theft.

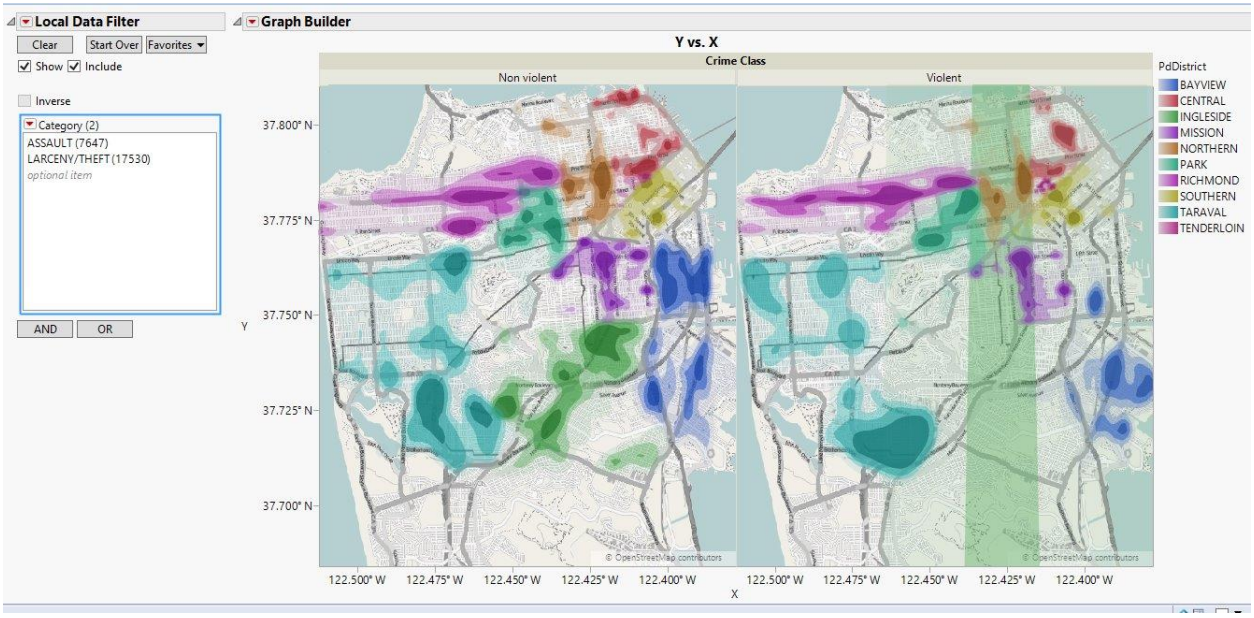
3. We clearly see Tenderloin emerge as a dominant neighbourhood with respect to drug related crimes and assault. However, non-violent crimes like vehicle theft, vandalism and burglary are low here.



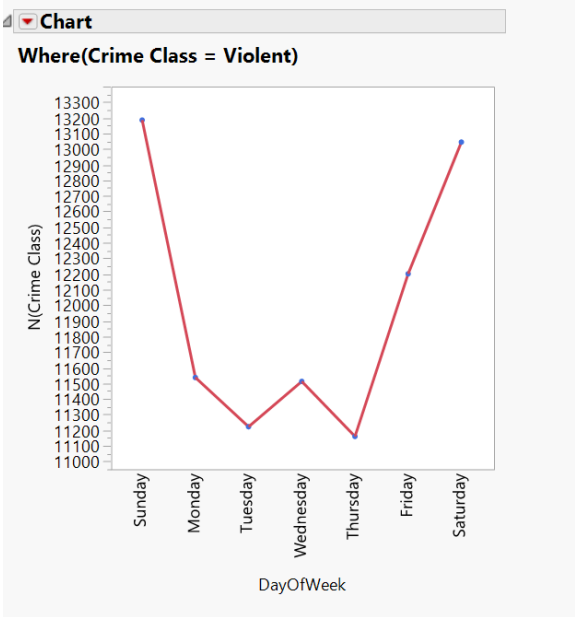
4. Comparing some violent categories vs non-violent categories, we can see how peak time vary. Sex Offenses forcible, Arson, Assault peak in between 6:00 PM to 6:00 AM. Gambling, Embezzlement, Fraud, Bad Checks all peak during the daylight hours. However, Larceny which is categorized as a non-violent crime, understandably occurs after dark. This shows that while time can be a good predictor for some categories, it can reduce the overall variability between Violent and non violent crime.

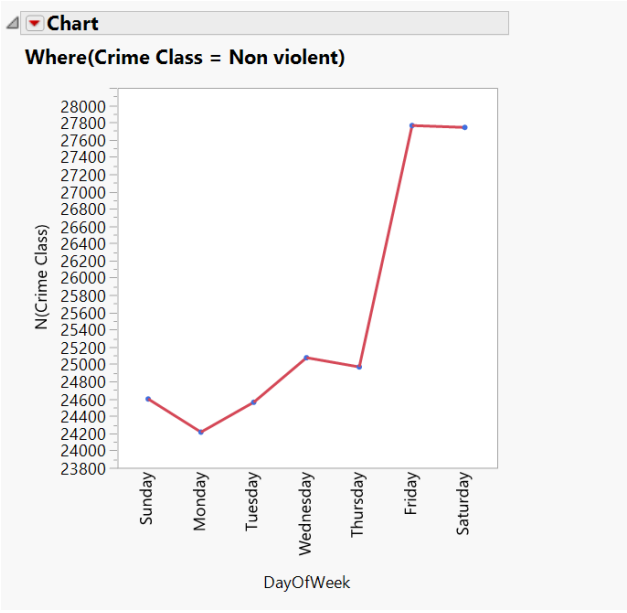


5. Assault and Larceny/Theft contribute the maximum to the violent and non-violent crime classes respectively. Comparing the density contour plot of both categories, we can see some clear variation in the data within a single district (for ex – density centres in Taraval and Bayview lie on streets far away). Precise latitude, longitude information or the street data would be a good differentiator at category level.



6. Aggregate variation by day of week – Crimes tend to occur more frequently on Fridays and Saturdays with an increase seen in both violent and non violent crimes. The differentiator at the aggregate level is Sunday, where we see a relatively higher percentage of violent crimes.





Predictive Modelling

After splitting the data into training and validation sets, we chose to run 3 different models (Decision trees, Logistic Regression and Neural Networks). However, while the overall accuracy was high, the sensitivity of the model to violent crimes was low. From the dataset, we see that the number of non- violent crimes reported (75%) by the police department received are a lot higher than the number of crimes categorized as non-violent.

Model results without stratification:

- Area under ROC Curve is approx 0.651
- Misclassification Rate is 0.2642, majorly contributed by errors in predicting violent crimes. (Accuracy – 78%)
- Sensitivity ratio (violent crimes) is approximately 0.116. As, we have taken predicting violent crimes as the category of interest, we chose to stratify the dataset in order to improve sensitivity.

We then stratified the dataset to have a more proportional balance of 1's and 0's and then created our model using this random sample. We then extrapolated the results of this to the entire dataset with much better results.

In the sections below, we will go through the ROC and Lift curves of the 3 options attempted and then evaluate the best fit.

Model Trials

Logistic Regression

Effect Summary

Source	LogWorth	PValue
PdDistrict	1080.258	0.00000
Hour[Dates]	513.753	0.00000
DayOfWeek	5.596	0.00000

Remove Add Edit FDR

Converged in Gradient, 4 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	3812.057	38	7624.114	<.0001*
Full	65941.423			
Reduced	69753.480			

RSquare (U)	0.0547
AICc	131961
BIC	132332
Observations (or Sum Wgts)	100633

Measure	Training	Validation	Definition
Entropy RSquare	0.0547	0.0515	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.0973	0.0918	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.6553	0.6575	∑ -Log(p[j])/n
RMSE	0.4811	0.4821	√ ∑(y[j]-p[j])²/n
Mean Abs Dev	0.4631	0.4639	∑ |y[j]-p[j]|/n
Misclassification Rate	0.3801	0.3830	∑ (p[j]≠pMax)/n
N	100633	67089	n

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	1641	1375.928	2751.857
Saturated	1679	64565.495	Prob>ChiSq
Fitted	38	65941.423	<.0001*

We see an improvement in overall performance of the model with accuracy at 62%. The AIC and BIC values also reduced from the model run without stratification which indicates an improvement in the model fit.

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	0.18195403	0.0084005	469.16	<.0001*
DayOfWeek[Sunday]	0.07806375	0.0160145	23.76	<.0001*
DayOfWeek[Monday]	0.01406421	0.0163474	0.74	0.3896
DayOfWeek[Tuesday]	0.01505101	0.0164263	0.84	0.3595
DayOfWeek[Wednesday]	-0.0012367	0.0162508	0.01	0.9393
DayOfWeek[Thursday]	-0.0208775	0.0163555	1.63	0.2018
DayOfWeek[Friday]	-0.0449479	0.0157119	8.18	0.0042*
PdDistrict[BAYVIEW]	0.70078126	0.0211961	1093.1	<.0001*
PdDistrict[CENTRAL]	-0.5422836	0.0186758	843.13	<.0001*
PdDistrict[INGLESIDE]	0.56064108	0.0215849	674.64	<.0001*
PdDistrict[MISSION]	0.2419008	0.017844	183.78	<.0001*
PdDistrict[NORTHERN]	-0.4854804	0.017431	775.71	<.0001*
PdDistrict[PARK]	-0.1644886	0.0269298	37.31	<.0001*
PdDistrict[RICHMOND]	-0.3957294	0.026973	215.25	<.0001*
PdDistrict[SOUTHERN]	-0.4593164	0.0149193	947.82	<.0001*
PdDistrict[TARAVAL]	0.0366572	0.0227641	2.59	0.1073
Hour[Dates][0]	0.24557855	0.0289757	71.83	<.0001*
Hour[Dates][1]	0.63339354	0.0351024	325.59	<.0001*
Hour[Dates][2]	0.95288401	0.0401449	563.40	<.0001*
Hour[Dates][3]	0.69505775	0.0521604	177.57	<.0001*
Hour[Dates][4]	0.63024499	0.0650128	93.98	<.0001*
Hour[Dates][5]	0.31558118	0.0680471	21.51	<.0001*
Hour[Dates][6]	0.05137608	0.0570385	0.81	0.3677
Hour[Dates][7]	0.08990892	0.046496	3.74	0.0532
Hour[Dates][8]	0.00077176	0.0363897	0.00	0.9831
Hour[Dates][9]	-0.0056926	0.0342075	0.03	0.8678
Hour[Dates][10]	-0.12326	0.0319906	14.85	0.0001*
Hour[Dates][11]	-0.1762784	0.0308315	32.69	<.0001*
Hour[Dates][12]	-0.1684762	0.0272451	38.24	<.0001*
Hour[Dates][13]	-0.2113445	0.0291764	52.47	<.0001*
Hour[Dates][14]	-0.2743507	0.0287311	91.18	<.0001*
Hour[Dates][15]	-0.1770981	0.0274337	41.67	<.0001*
Hour[Dates][16]	-0.2771542	0.027573	101.04	<.0001*
Hour[Dates][17]	-0.332947	0.02661	156.55	<.0001*
Hour[Dates][18]	-0.5221562	0.0259452	405.03	<.0001*
Hour[Dates][19]	-0.4821264	0.0265489	329.78	<.0001*
Hour[Dates][20]	-0.3034289	0.0275072	121.68	<.0001*
Hour[Dates][21]	-0.1540919	0.028195	29.87	<.0001*
Hour[Dates][22]	-0.1984521	0.0282124	49.48	<.0001*

The parameter estimates not significant in detecting a violent crime are Taraval district, all weekdays except Friday and hours 6 – 9AM of the day which have higher p-value.

Confusion Matrix

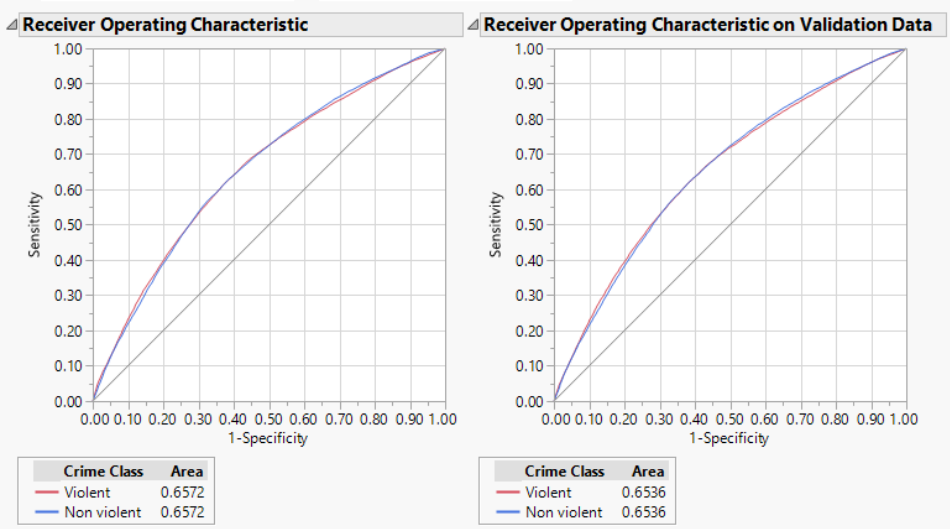
Training

Actual	Predicted	
Crime Class	Violent	Non violent
Violent	29893	20424
Non violent	17824	32492

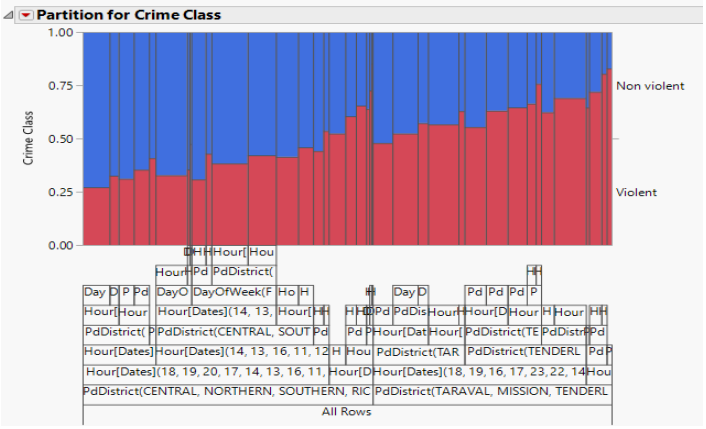
Validation

Actual	Predicted	
Crime Class	Violent	Non violent
Violent	19766	13778
Non violent	11915	21630

Area under the ROC curve is 0.6536. The sensitivity ratio has increased to around 58.90%.



Decision Trees

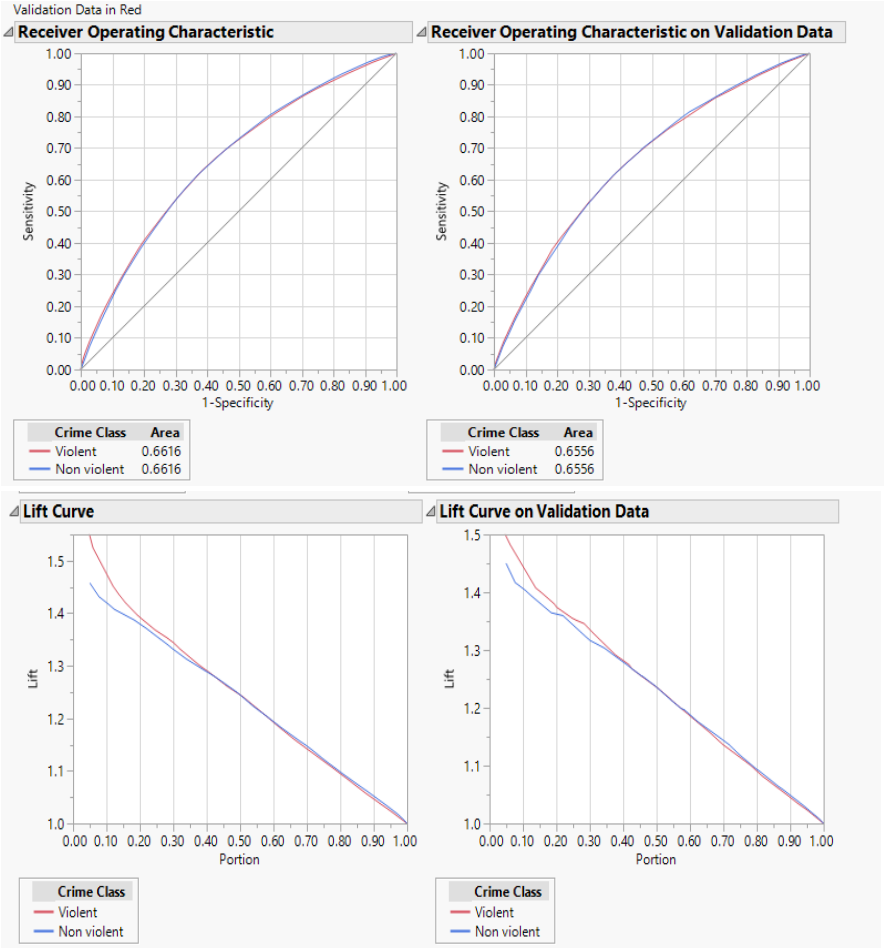


37 splits were chosen as optimal.

Leaf Report

Response Prob

Leaf Label	Violent	Non violent
^^^&PdDistrict(NORTHERN, CENTRAL, SOUTHERN, RICHMOND)&Hour[Dates](18, 19)&DayOfWeek(Thursday, Friday, Monday, Wednesday, Tuesday)	0.2714	0.7286
^^^&PdDistrict(NORTHERN, CENTRAL, SOUTHERN, RICHMOND)&Hour[Dates](18, 19)&DayOfWeek(Saturday, Sunday)	0.3257	0.6743
^^^^&Hour[Dates](20, 17)&PdDistrict(CENTRAL, NORTHERN)	0.3110	0.6890
^^^^&Hour[Dates](20, 17)&PdDistrict(SOUTHERN, RICHMOND)	0.3536	0.6464
^^&Hour[Dates](18, 19, 20, 17)&PdDistrict(PARK)	0.4078	0.5922
^^^&PdDistrict(CENTRAL, SOUTHERN, NORTHERN, RICHMOND)^&DayOfWeek(Saturday, Sunday)&Hour[Dates](11, 13, 15, 14, 16, 12, 10, 23)	0.3275	0.6725
^^^&PdDistrict(CENTRAL, SOUTHERN, NORTHERN, RICHMOND)^^&Hour[Dates](22)&DayOfWeek(Saturday)	0.3546	0.6454
^^^&PdDistrict(CENTRAL, SOUTHERN, NORTHERN, RICHMOND)^^&Hour[Dates](22)&DayOfWeek(Sunday)	0.4740	0.5260
^^^^^&DayOfWeek(Friday, Monday, Wednesday, Thursday, Tuesday)&PdDistrict(CENTRAL)&Hour[Dates](14, 11, 12, 13, 16, 15)	0.3081	0.6919
^^^^^&DayOfWeek(Friday, Monday, Wednesday, Thursday, Tuesday)&PdDistrict(CENTRAL)&Hour[Dates](10, 23, 22)	0.4287	0.5713
^^^^^&DayOfWeek(Friday, Monday, Wednesday, Thursday, Tuesday)&PdDistrict(SOUTHERN, NORTHERN, RICHMOND)&Hour[Dates](22, 23, 14, 13, 16)	0.3833	0.6167
^^^^^&DayOfWeek(Friday, Monday, Wednesday, Thursday, Tuesday)&PdDistrict(SOUTHERN, NORTHERN, RICHMOND)&Hour[Dates](10, 11, 15, 12)	0.4217	0.5783
^^^&PdDistrict(CENTRAL, SOUTHERN, NORTHERN, RICHMOND)^&Hour[Dates](21, 9)	0.4137	0.5863
^^^&PdDistrict(CENTRAL, SOUTHERN, NORTHERN, RICHMOND)^&Hour[Dates](8, 7, 6)	0.4584	0.5416
^^^&PdDistrict(PARK)&Hour[Dates](22, 21, 12, 14, 13, 8, 16, 23)	0.4407	0.5593
^^^&PdDistrict(PARK)&Hour[Dates](10, 6, 15, 11, 9, 7)	0.5357	0.4643
PdDistrict(CENTRAL, NORTHERN, SOUTHERN, RICHMOND, PARK)^&Hour[Dates](0, 5)	0.5237	0.4763
^^^&PdDistrict(RICHMOND, NORTHERN, PARK, SOUTHERN)&Hour[Dates](1, 4)	0.6045	0.3955
^^^&PdDistrict(RICHMOND, NORTHERN, PARK, SOUTHERN)&Hour[Dates](3, 2)	0.6548	0.3452
^^&Hour[Dates](4, 1, 3, 2)&PdDistrict(CENTRAL)&DayOfWeek(Wednesday, Thursday, Friday, Tuesday, Monday)	0.6394	0.3606
^^^&PdDistrict(CENTRAL)&DayOfWeek(Sunday, Saturday)&Hour[Dates](4, 3, 1)	0.7251	0.2749
^^^&PdDistrict(CENTRAL)&DayOfWeek(Sunday, Saturday)&Hour[Dates](2)	0.8498	0.1502
^^^&Hour[Dates](18, 19, 21,17, 23, 22, 16, 20,6, 10, 8)&PdDistrict(TARAVAL)	0.4786	0.5214
^^^&Hour[Dates](18, 19, 21, 17, 23, 22, 16, 20, 6, 10, 8)&PdDistrict(MISSION)&DayOfWeek(Tuesday, Monday, Thursday, Wednesday, Friday)	0.5234	0.4766
^^^&Hour[Dates](18, 19, 21, 17, 23, 22, 16, 20, 6, 10, 8)&PdDistrict(MISSION)&DayOfWeek(Saturday, Sunday)	0.5720	0.4280
^^&PdDistrict(TARAVAL, MISSION)^&Hour[Dates](15, 11, 12, 13, 14, 7, 9)	0.5656	0.4344
^^&PdDistrict(TARAVAL, MISSION)^&Hour[Dates](0)	0.6288	0.3712
^^^^&Hour[Dates](16, 14, 19, 15, 18, 17, 13, 12, 20)&PdDistrict(TENDERLOIN)	0.5533	0.4467
^^^^&Hour[Dates](16, 14, 19, 15, 18, 17, 13, 12, 20)&PdDistrict(INGLESIDE)	0.6307	0.3693
^^^^&Hour[Dates](11, 10, 21, 23, 22, 0, 9, 6, 8, 7)&PdDistrict(INGLESIDE)	0.6468	0.3532



Area under the ROC curve is 0.6556, with a lift ratio of 1.4 on the top 10% of the data.

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.0591	0.0539	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.1048	0.0960	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.6522	0.6558	∑ -Log(ρ[j])/n
RMSE	0.4797	0.4814	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.4603	0.4618	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.3774	0.3820	∑ (ρ[j]≠ρMax)/n
N	100633	67089	n

Confusion Matrix

Training

Actual	Predicted	
Crime Class	Violent	Non violent
Violent	31712	18605
Non violent	19374	30942

Validation

Actual	Predicted	
Crime Class	Violent	Non violent
Violent	20885	12659
Non violent	12970	20575

- The results from the recursive partition model has sensitivity improved up to **63%** from 14.5% and the misclassification rate **0.37** which indicates a predictive accuracy of **63%** The issue of overfitting is resolved with RMSE reduced for validation dataset.
- The sensitivity is 63% with number of true positive(violent crimes) predicted to be approximately 20,885 in this model and the number of (false negative)non-violent crimes predicted to be approximately 12,659.
- The leaf report indicates that there is probability of **83%** for violent crimes to occur in Tenderloin district at hours of 4 A.M., 3 A.M., 5 A.M. and 2 A.M. during the early start of the day.
- More non-violent crimes are observed in the Taraval and Mission District.

Neural Networks

Model Launch

Model NTanH(6)

Training

Crime Class

Measures	Value
Generalized RSquare	0.1068659
Entropy RSquare	0.0602643
RMSE	0.4792796
Mean Abs Dev	0.4595237
Misclassification Rate	0.3768247
-LogLikelihood	65549.838
Sum Freq	100633

Confusion Matrix

Actual	Predicted	
Crime Class	Violent	Non violent
Violent	31188	19129
Non violent	18792	31524

Confusion Rates

Actual	Predicted Rate	
Crime Class	Violent	Non violent
Violent	0.620	0.380
Non violent	0.373	0.627

Validation

Crime Class

Measures	Value
Generalized RSquare	0.0963137
Entropy RSquare	0.0540846
RMSE	0.4813327
Mean Abs Dev	0.4613344
Misclassification Rate	0.3801965
-LogLikelihood	43987.481
Sum Freq	67089

Confusion Matrix

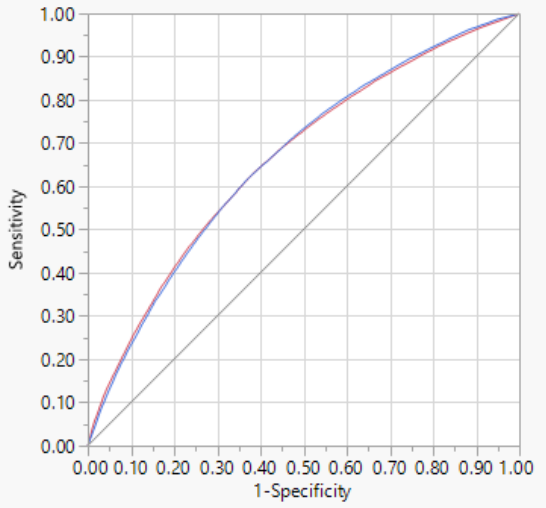
Actual	Predicted	
Crime Class	Violent	Non violent
Violent	20611	12933
Non violent	12574	20971

Confusion Rates

Actual	Predicted Rate	
Crime Class	Violent	Non violent
Violent	0.614	0.386
Non violent	0.375	0.625

Training

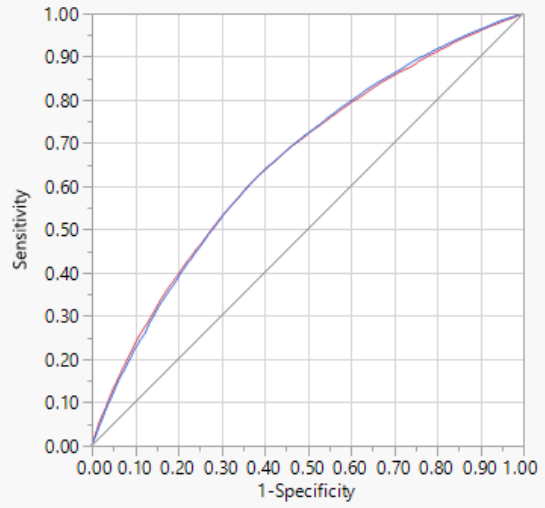
Receiver Operating Characteristic



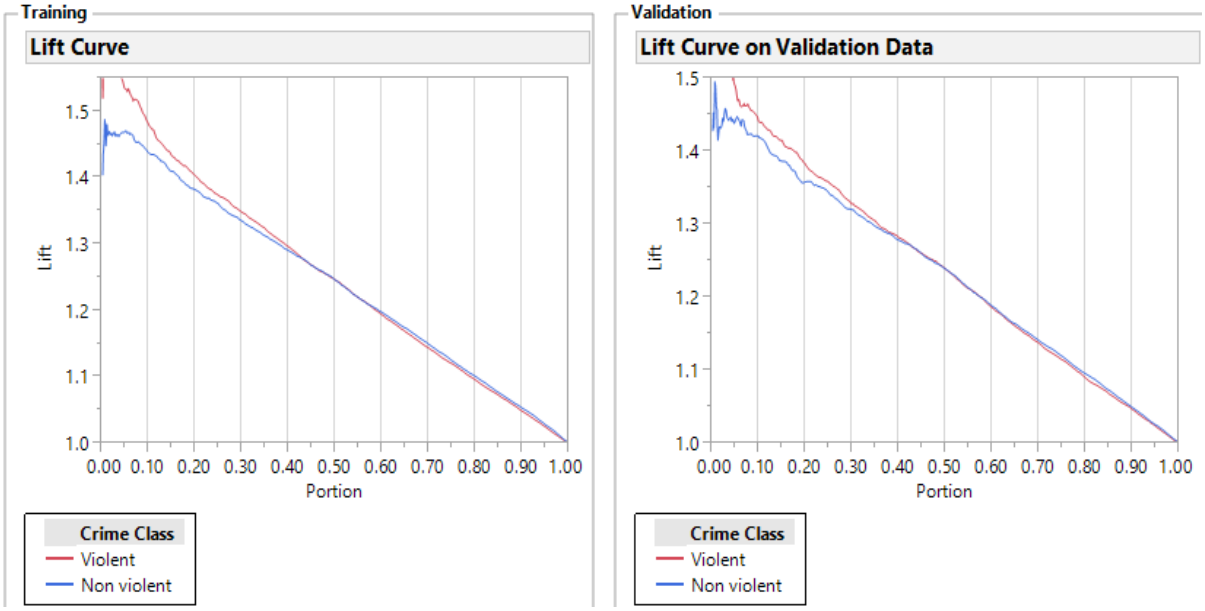
Crime Class	Area
Violent	0.6641
Non violent	0.6641

Validation

Receiver Operating Characteristic on Validation Data



Crime Class	Area
Violent	0.6563
Non violent	0.6563


Training
Lift Curve
Lift
Portion
1.5
1.4
1.3
1.2
1.1
1.0
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
Crime Class
Violent
Non violent
Validation
Lift Curve on Validation Data
Lift
Portion
1.5
1.4
1.3
1.2
1.1
1.0
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
Crime Class
Violent
Non violent

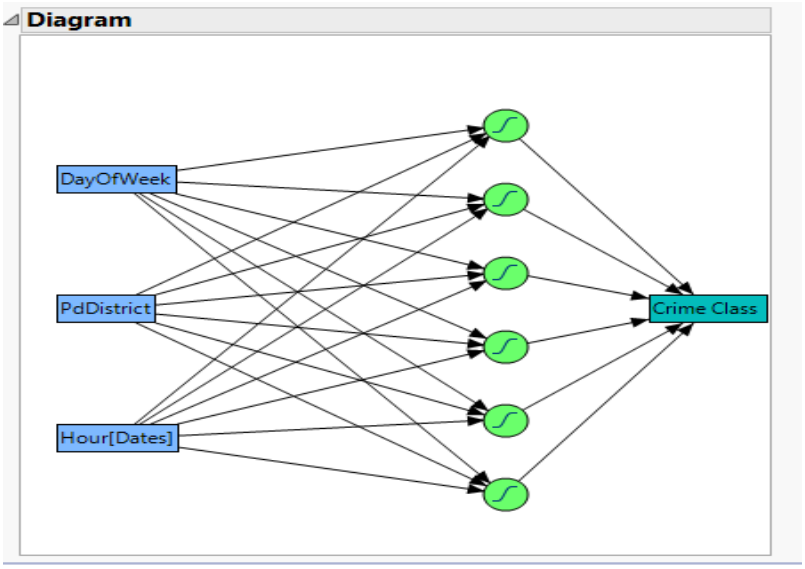


Diagram
DayOfWeek
PdDistrict
Hour[Dates]
Crime Class

- The neural networks indicate the misclassification rate of 0.3767 and predictive accuracy of 62.23%.
- The ROC curve provides better fit for predicting violent crimes and represents an area of 0.6563.
- The sensitivity is 0.6144 or 61.44% with number of True positive(violent crimes) predicted to be approximately 20,611 in this model and the number of (False Negative)non-violent crimes predicted to be approximately 12,933.
- The lift curve is good with around 40% of the data showing lift ratio of approximately 1.35 for violent crimes and 1.3 for non-violent crimes.
- The number of nodes selected for neural networks is n=6 because the other nodes show problems of overfitting and there are higher misclassification rates observed for n=4,10,15. Hence we are selecting n=6 for the final model run.
- The overall improvement in the model observed with predictors PdDistrict, Hour Dates and Day of Week is better after stratification of data.

Evaluation of the Models:

Total Number of Rows 67089 - Validation Data Set

Logistic

	Actual	
Predicted	Non-Violent	Violent
Non-Violent	21,630	13,778
Violent	11,915	19,766

Accuracy of the Model 61.70%

Cost - 0.4		
Cost - 12		

Predicted 1		31,681
Accuracy of 1	0.623907074	62.39%
% Rows		396%

Lift Ratio for the Models 1.248

Classification Tree

	Actual	
Predicted	Non-Violent	Violent
Non-Violent	20,575	12,659
Violent	12,970	20,885

Accuracy 61.80%

Predicted 1		33,855
Accuracy of 1	0.616895584	62%
% Rows		423%

Baseline 0.499992547

Lift Ratio for the Models 1.234

Neural Nets

	Actual	
Predicted	Non-Violent	Violent
Non-Violent	20,971	12,933
Violent	12,574	20,611

Accuracy 61.98%

Predicted 1		33,185
Accuracy of 1	0.621093868	62%
% Rows		415%

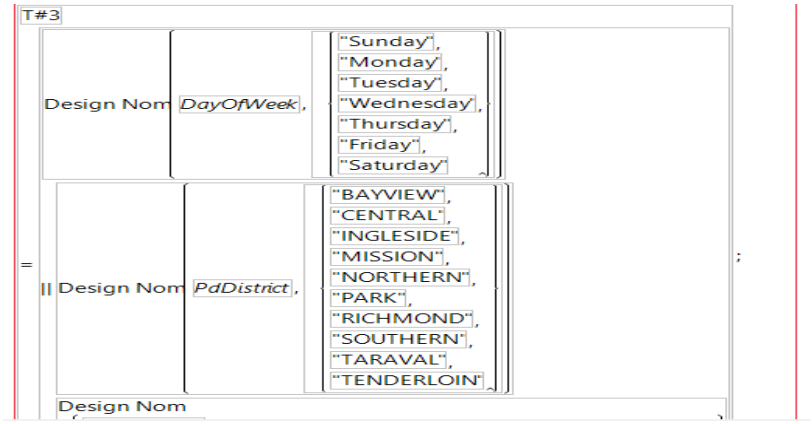
Lift Ratio for the Mod 1.242

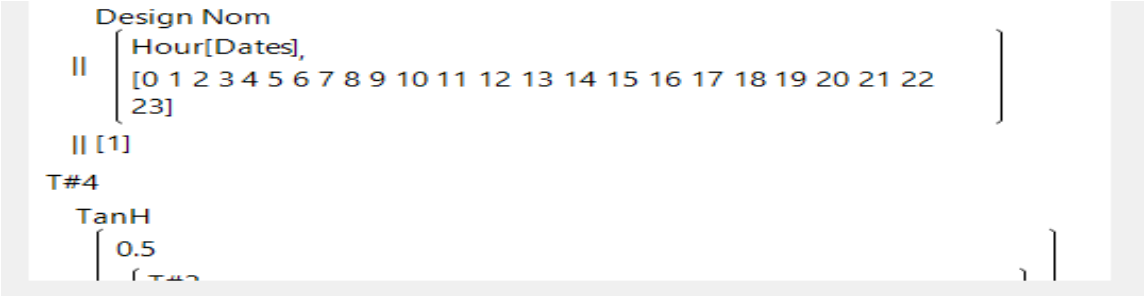
Models	Lift Ratio	Accuracy	Sensitivity	ROC
Logistics	1.248	61.7%	58.9%	0.6536
Decision Trees	1.238	61.8%	63%	0.6556
Neural Networks	1.242	61.98%	61.44%	0.6563

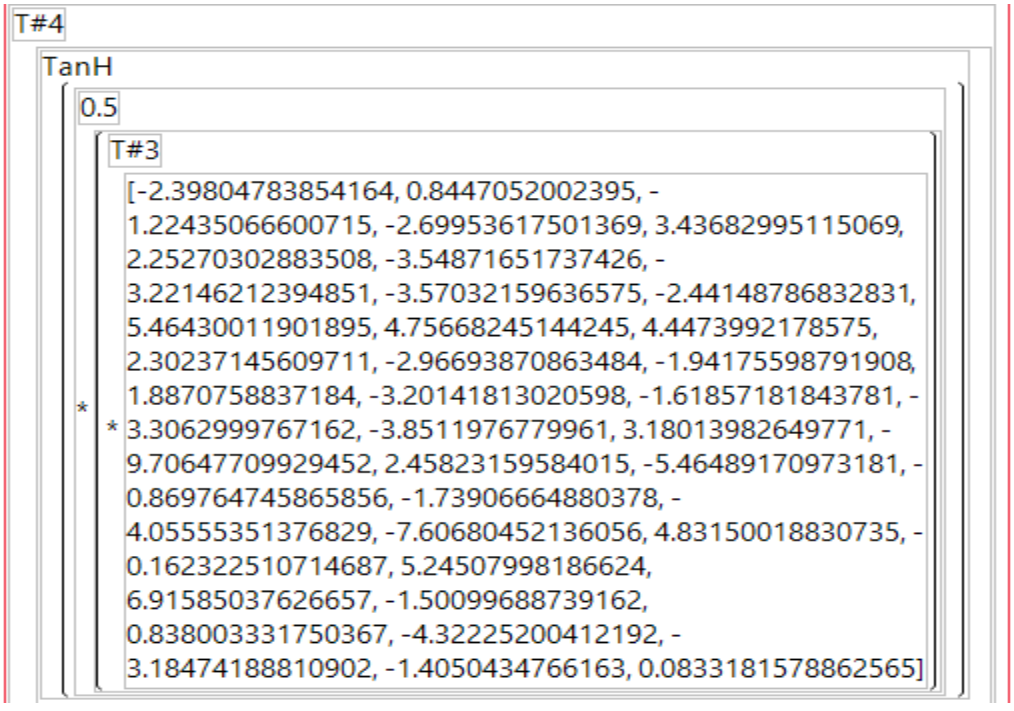
- With a constant ROC, we have considered the trade-off between Accuracy and Sensitivity keeping ROC constant and found Neural Networks to be the best fit model.
- The overall fit for the Neural Networks Models is better in terms of the Lift Ratio, Sensitivity and ROC as compared to other models.
- The logistic models does not show a good sensitivity in predicting violent crimes over non-violent. Hence, considering the class of interest, we rejected it in comparison to Neural Networks and Decision Trees.
- If we look at the RMSE values for validation throughout the three models, it is consistent in comparison to training data so there is no problem of overfitting.
- The misclassification rate is a little higher for Logistic Regression and Decision Tree models in comparison to Neural Network Model.
- We see that Neural Network Model has a decently low false positive rate (37.5%) and high accuracy (61.44%). The decision tree has a high true positive Rate as well as very high False positive rate. This suggests that it is overpredicting 1's. The cost of false positive rate in our scenario is considerate because of the limit of police resources in real life.

Final Implementation:

We have selected the Neural Networks model based on our results and observation of the dataset. After running the models and interpreting the accuracy we have arrived at this final model. The reason being the intangible loss of predicting violent is low in comparison to non-violent with respect to other models. The final Formula to be implemented using Neural Network Analysis to predict crimes in San Francisco is given below:


T#3
Design Nom
DayOfWeek,
"Sunday",
"Monday",
"Tuesday",
"Wednesday",
"Thursday",
"Friday",
"Saturday"
=
|| Design Nom
PdDistrict,
"BAYVIEW",
"CENTRAL",
"INGLESIDE",
"MISSION",
"NORTHERN",
"PARK",
"RICHMOND",
"SOUTHERN",
"TARAVAL",
"TENDERLOIN"
Design Nom


Design Nom
|| Hour[Dates],
[0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23]
|| [1]
T#4
TanH
0.5


T#4
TanH
0.5
T#3
[-2.39804783854164, 0.8447052002395, -1.22435066600715, -2.69953617501369, 3.43682995115069, 2.25270302883508, -3.54871651737426, -3.22146212394851, -3.57032159636575, -2.44148786832831, 5.46430011901895, 4.75668245144245, 4.4473992178575, 2.30237145609711, -2.96693870863484, -1.94175598791908, 1.8870758837184, -3.20141813020598, -1.61857181843781, -3.3062999767162, -3.8511976779961, 3.18013982649771, -9.70647709929452, 2.45823159584015, -5.46489170973181, -0.869764745865856, -1.73906664880378, -4.05555351376829, -7.60680452136056, 4.83150018830735, -0.162322510714687, 5.24507998186624, 6.91585037626657, -1.50099688739162, 0.838003331750367, -4.32225200412192, -3.18474188810902, -1.4050434766163, 0.0833181578862565]

Conclusion:

The locations of **Ingleside, Tenderloin** have high occurrence of violent crimes whereas, the neighbourhoods of Southern and Central Richmond, have high occurrence of non-violent crimes during the weekdays Monday, Tuesday and Wednesday. The most prominent time for violent crimes like kidnapping, assault etc. is around 2 A.M to 3 A.M. in the morning, with likely days of the week being Friday, Saturday and Sunday.

As we are not able to use longitude/latitude in the current dataset to predict and are also forced to fit the original 39 categories into 2 classes for a classification model, the scope for accuracy is reduced. This is also a dataset related to human behaviour and combining this data with census records, social media, education and employment records etc will give some more variability and useful information to find patterns related to crime.

Appendix

We have not used Clustering, Principle Component Analysis as all the predictors are categorical variables. These methods calculate distance function between the variables which is not possible with categorical variables.

Sources for Crime Classification:

We have used the San Francisco Police Department to understand the difference between what is considered a violent crime and non-violent crime.

https://hilo.hawaii.edu/security/CaseDescriptions.php

https://www.neighborhoodscout.com/ca/san-francisco/crime

https://www.trulia.com/real_estate/San_Francisco-California/crime/

https://www.trulia.com/blog/trends/trulia-local/

https://www.fastcodesign.com/1664491/infographic-of-the-day-when-do-criminals-prowl-the-streets