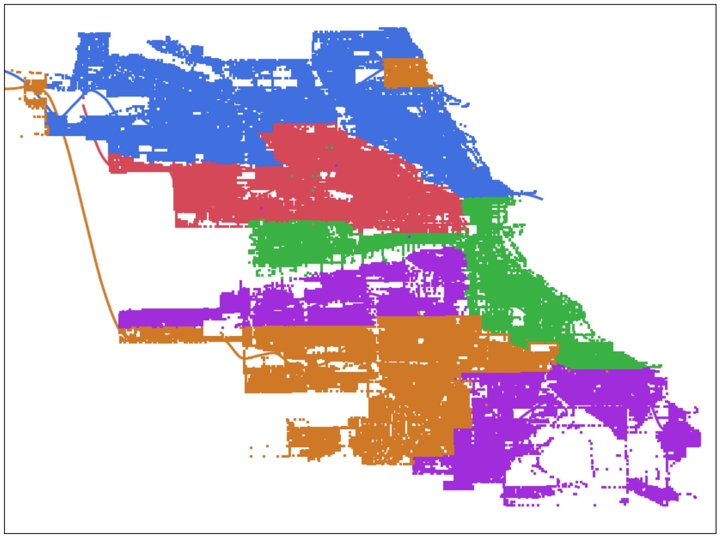
OPIM5604-SECB12
Group Project Final Report
Instructor: Iva Stricevic

CRIMES IN CHICAGO



Team Members:
Swati Arora
Ankita Paunikar
Siddharth Rai
Anoop Ramathirtha
Kees Van Haasteren

SUMMARY

The dataset can be found on https://www.kaggle.com/currie32/crimes-in-chicago .It contains 1,456,714 records/rows and 23 variables to describe the data. The variables contain some of the important information about the crime in Chicago like the date of crime, location, type of crime and various other description.

The objective is to analyze the data and find possible patterns between various variables and predict whether the criminal will be arrested for criminal offense or not depending on the provided dataset. The offense rate variation based on time is also explored. With such understanding, authorities can proactively take measures to prevent some of the potential crimes. In addition, any other observation, unrelated to the prediction goal, will be recorded and summarized in the paper.

Objectives

I. Data description
II. Data visualization and pattern discovery
III. Data pre-processing
IV. Data Distribution
V. Model Evaluation
VI. Conclusion

I. Data description

This dataset analyzes the criminal behavior in the city of Chicago in a variety of different ways. It is a compilation of reported criminal activities within the city limits to accurately depict each individual crime, the aspects surrounding the crime, and the activities. With this data, we hope to be able to spot trends in the crimes being committed and the arresting patterns of officers to determine where the strengths and weakness of the Chicago PD lies. Is there an area or district within the city which has a low arrest rate and requires more officers? We hope to be able to accurately predict whether a crime, given the nature of the crime, the location of the crime, and other aspects of the report, will result in an arrest and justice for those who have been aggrieved.

The data includes 1,456,714 rows (individual crimes) which have taken place from 2012 to the early months of 2017. The data set recognizes 23 facets of the crime, which will serve as our columns in the predictive model. These columns can be divided into four categories. The first are identifiers, the data point is unique and allows people studying the crime to find individual crimes to focus on. The second are descriptors of the crime type, which dictate what the crime was, and the severity. The third are location identifiers, showing the location of the crime, whose jurisdiction it fall under. The fourth category is a grab bag of the remainder, including the result (arrest/no arrest), the update information and the time of the crime.

Identifiers: Column 1 is a number based solely in this dataset, simply a running tally of the crimes. ID and Case Number were assigned by Chicago PD, to allow each case a unique identification to easier track and solve the crimes. All identifiers must not be included for the model when it is created, as they are unique numbers which are not continuous.

Crime Type: There are 4 columns, each identifying the type of crime. IUCR and FBI Code are each based on assigning different crimes and severities a number, so that each crime can be grouped on a state or nationwide level. Primary Type divides the crimes into police officer described categories. These will need to be grouped in order to use them effectively. Description involves the severity of the crime, whether weapons were used, etc. and were also described by the police officer reporting. Because of the large variety of reports, it becomes difficult to group or label crimes together in this category.

Location: There are 11 different columns describing some variation of the location of the crime. Ward, District, Beat and Community Area, are all different levels of establishing which officers are patrolling which area. The city is divided into 3 districts, which are subdivided into 77 Community areas and 50 Wards (Precincts or Police Stations) and each of those are divided into beats, describing where each team of officers is scheduled to patrol. The other type of location column are exact descriptors, the city block, longitude and latitude or a description of the crime scene (street or residence, for example).

Other Columns: Date and year will describe when the crime occurred, we are able to break the time into hour of day and month to show more accurate cycles of crime throughout a day or year.

Predicted Value: The column arrest is a binary choice, whether or not an arrest was made in this particular crime. We will attempt to project a yes or no binary outcome (shown as 1 for arrest, 0 for no arrest) in this model.

II. Data visualization and pattern discovery

1. Top ten communities with most Arrest and least number of Arrest

Most Arrest

	Community Area	Arrest	N Rows
1	25	True	35356
2	23	True	18122
3	29	True	17018
4	26	True	15140
5	67	True	12821
6	68	True	11385
7	8	True	11366
8	27	True	11267
9	43	True	11263
10	71	True	11147

Least Arrest

	Community Area	Arrest	N Rows
1	0	True	4
2	•	True	6
3	0	False	9
4	•	False	31
5	9	True	189
6	12	True	287
7	47	True	457
8	74	True	485
9	18	True	515
10	36	True	558

2. The Primary Crime Types versus primary crimes which lead to Arrest

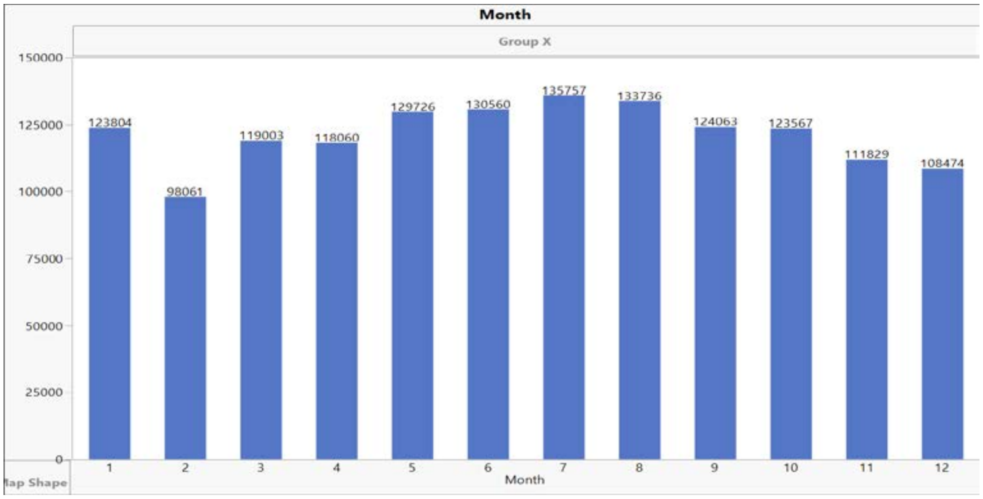
Primary Crime Types

	Primary Type	N Rows
1	THEFT	329450
2	BATTERY	263684
3	CRIMINAL DAMAGE	155449
4	NARCOTICS	135232
5	ASSAULT	91284
6	OTHER OFFENSE	87868
7	BURGLARY	83395
8	DECEPTIVE PRACTICE	75491
9	MOTOR VEHICLE THEFT	61132
10	ROBBERY	57310

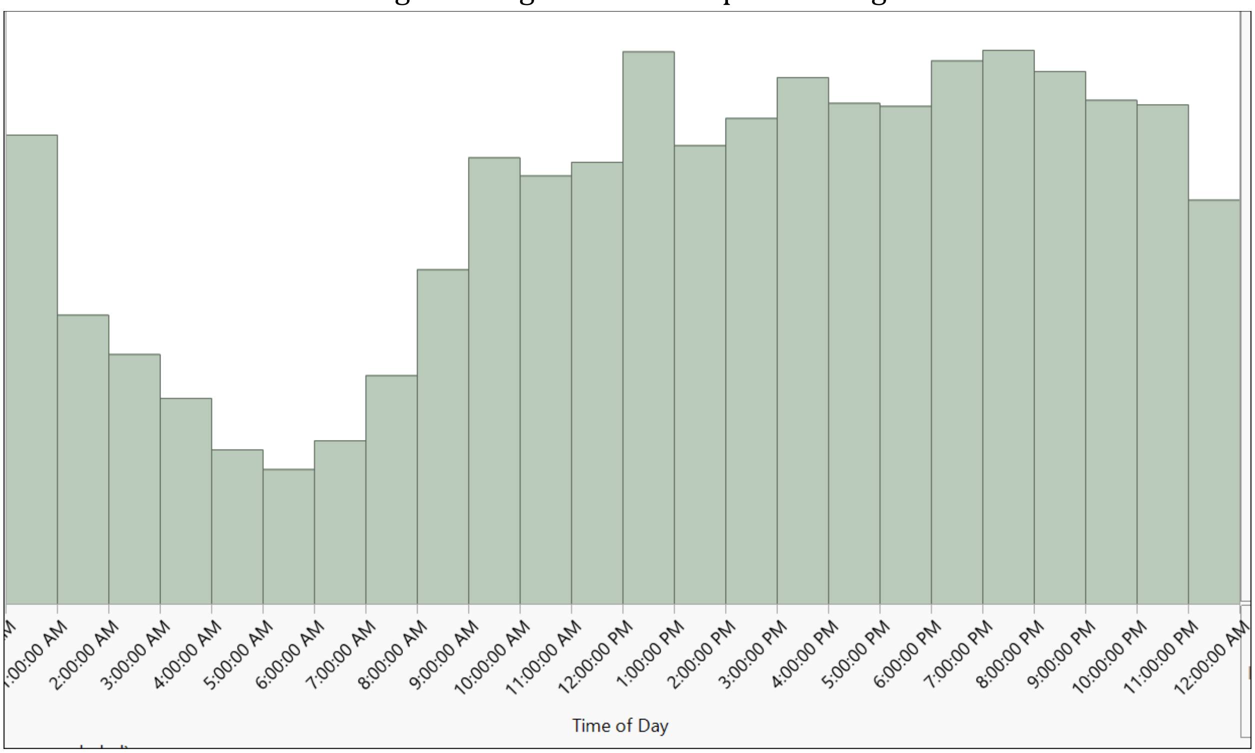
Primary Crime Types with most Arrest

	Primary Type	Arrest	N Rows
1	NARCOTICS	True	134309
2	BATTERY	True	60500
3	THEFT	True	36673
4	CRIMINAL TRESPASS	True	25926
5	ASSAULT	True	21347
6	OTHER OFFENSE	True	18572
7	WEAPONS VIOLATION	True	13745
8	CRIMINAL DAMAGE	True	10165
9	PUBLIC PEACE VIOLATION	True	9947
10	DECEPTIVE PRACTICE	True	8917

3. Most of the crimes are committed during Summer (July, August, June, May), whereas less criminal activities during winters (February, December, November and April)

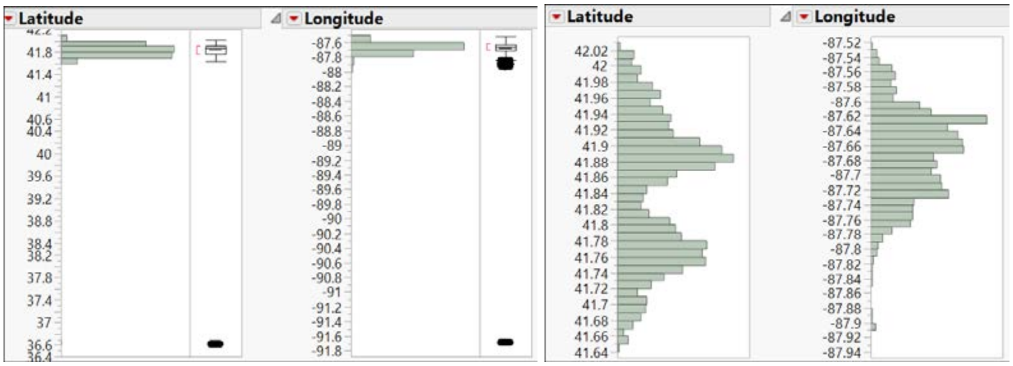


4. Less number of crimes during morning hours as compared to night.



III. Data pre-processing

1. Datatype: The datatype of all the variables are correct.
2. Outliers: By plotting the distribution of the variables and selecting outlier boxplot the Latitude and Longitude variables has 77 outlier values for same rows in the dataset. These values are too less and can be removed.



3. Missing Data: Used missing data pattern to check the missing values in the dataset. X coordinate, Y coordinate, latitude and longitude had some missing values. The total number of rows with missing values were 38,349 which is less than 5% of the total data. So, we can remove the rows with missing values because the variables which have missing values are not useful for our classification.

	Count	Number of columns missing	Patterns	Co
1	1418365	0	0000000000000000000000	
2	36635	5	0000000000000001100111	
3	25	1	0000000000000100000000	
4	15	6	0000000000000101100111	
5	14	1	0000000000001000000000	
6	1	1	0000000000010000000000	
7	1226	1	0000000100000000000000	
8	432	6	0000000100000001100111	
9	1	6	0010000000000001100111	

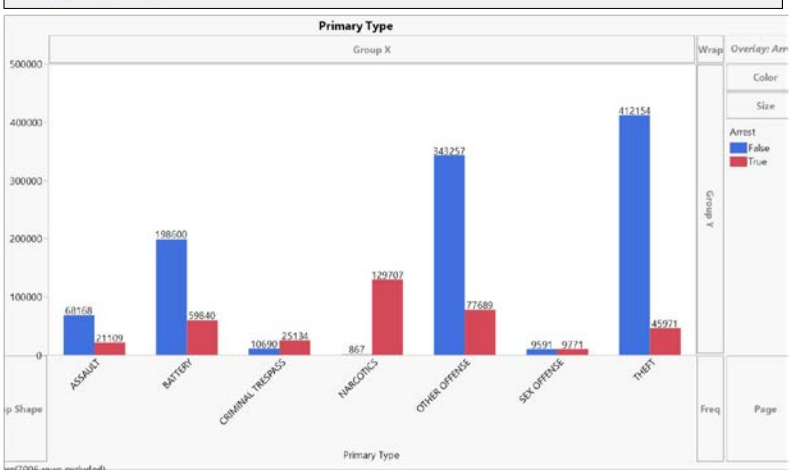
All rows	1,456,714
Selected	38,349
Excluded	77
Hidden	77
Labelled	0

4. Recode: There are various columns which have nominal datatype. To predict the Arrest, we need to reduce the number of values in those columns, so that the model can provide us the best results. With our understanding of the dataset we recoded the values to reduce the number of option for a column.

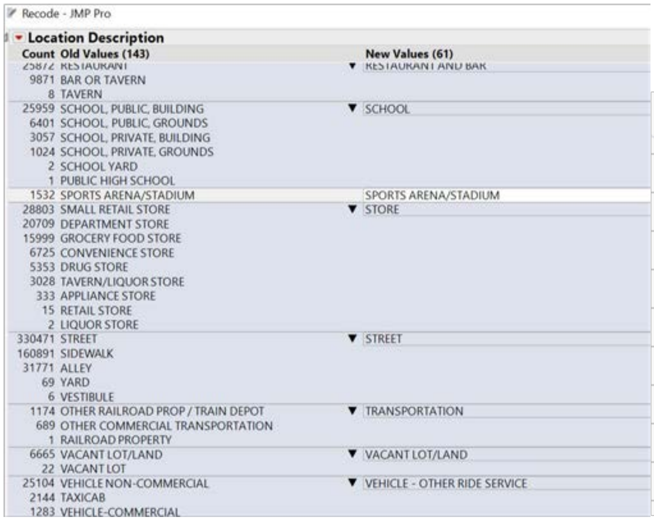
- Primary Type: From 29 to 7 values

Primary Type

Count	Old Values (29)	New Values (7)
89507	ASSAULT	ASSAULT
258929	BATTERY	BATTERY
36428	CRIMINAL TRESPASS	CRIMINAL TRESPASS
131203	NARCOTICS	NARCOTICS
152811	CRIMINAL DAMAGE	OTHER OFFENSE
85356	OTHER OFFENSE	
68352	DECEPTIVE PRACTICE	
59852	MOTOR VEHICLE THEFT	
17066	WEAPONS VIOLATION	
13013	PUBLIC PEACE VIOLATION	
10590	OFFENSE INVOLVING CHILDREN	
6136	INTERFERENCE WITH PUBLIC OFFICER	
2590	HOMICIDE	
2211	GAMBLING	
2175	ARSON	
1928	LIQUOR LAW VIOLATION	
1075	KIDNAPPING	
643	INTIMIDATION	
122	NON-CRIMINAL	
7567	PROSTITUTION	SEX OFFENSE
6298	CRIM SEXUAL ASSAULT	
4491	SEX OFFENSE	
774	STALKING	
169	OBSCENITY	
61	PUBLIC INDECENCY	
20	HUMAN TRAFFICKING	
322423	THEFT	THEFT
81671	BURGLARY	
56093	ROBBERY	

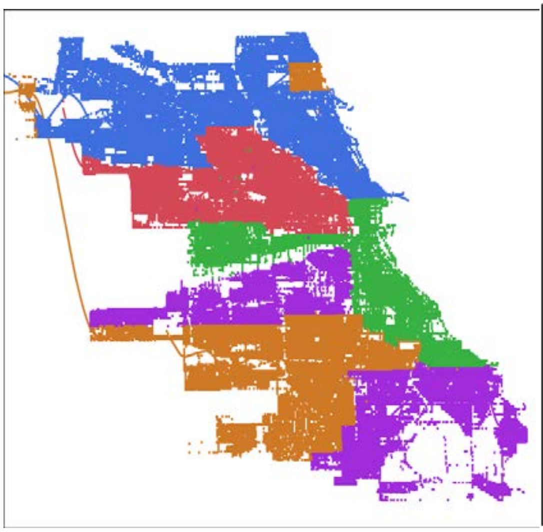


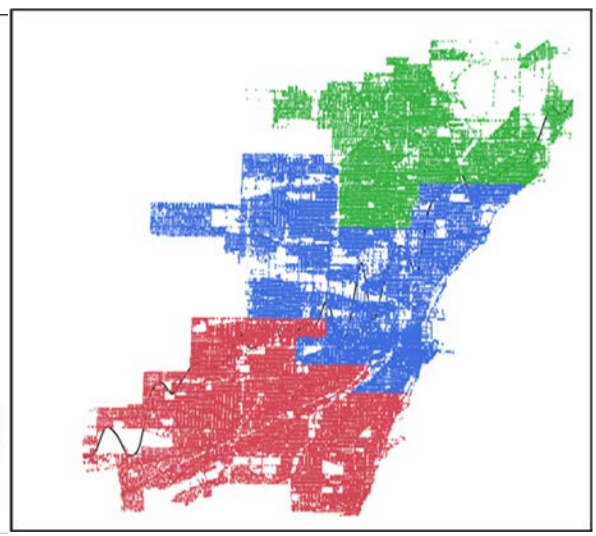
● Location Description : From 143 to 9 values.



	Location Description	N Rows
1	GAS STATION	15030
2	OTHER	121145
3	PARK PROPERTY	12104
4	RESIDENCE	492820
5	RESTAURANT/HOTEL	40985
6	SCHOOL/UNIVERSITY	37151
7	STORE	79416
8	STREET	514737
9	VEHICLE - OTHER RIDE SERVICE	99160

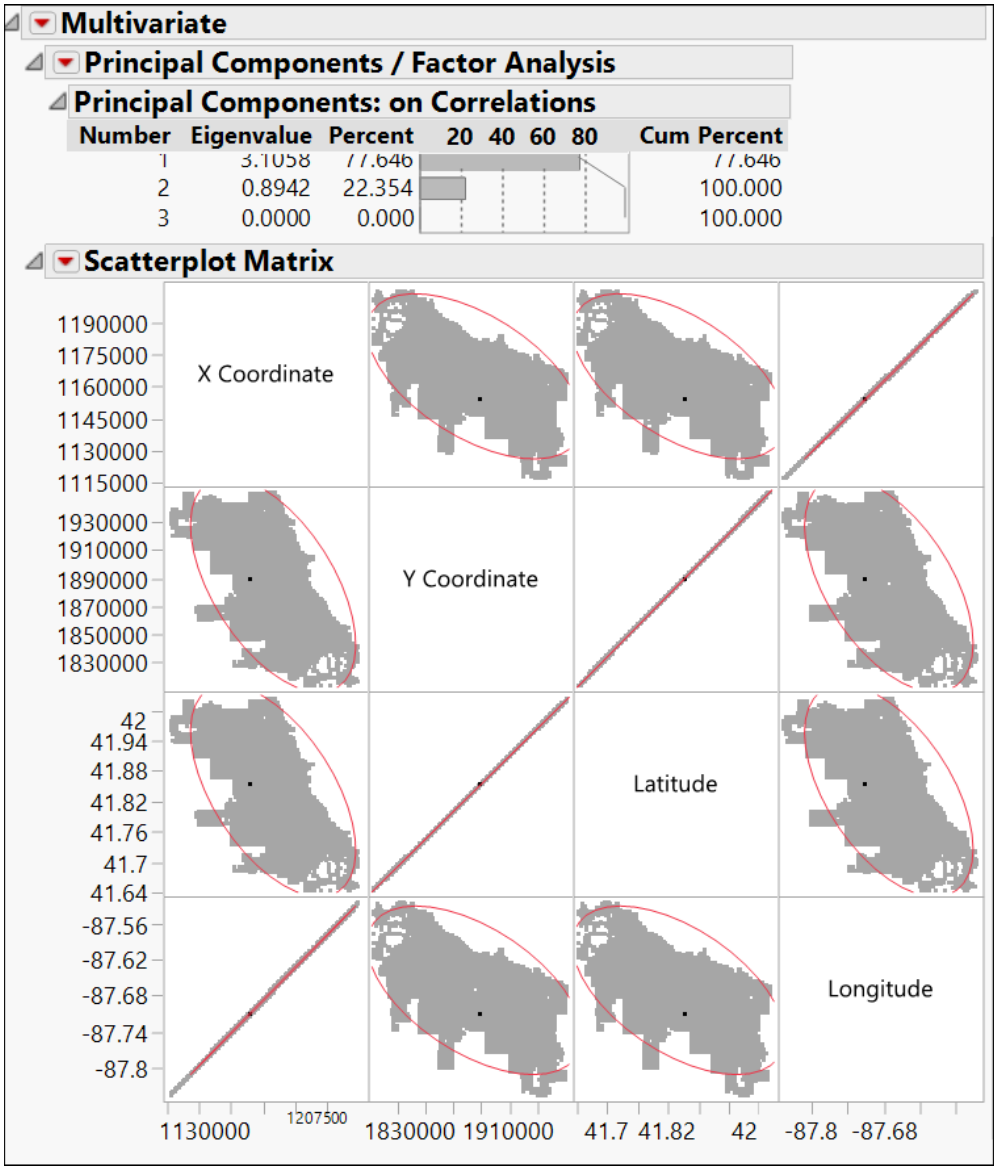
● Districts of Chicago PD, binned to Zones : From 25 to 3 values.





Pattern Discovery

Principal Component Analysis: X and Y coordinates are correlated with the Latitudes and Longitudes. Hence, only the Latitudes and Longitudes were considered for prediction. Other variables like ward, beat etc are not correlated but are a measure of Location with varying degrees of land area coverage. Only the the largest of these i.e. District was considered, and bined further into three zones – North, Central and South – to arrive at a better predictive model.



IV. Data Distribution

We tried to create the model using various distribution methods like: Stratified and random stratified. For the final model, the data is distributed into training, validation and test part (60:30:10).

	Arrest	Validation	N Rows
1	False	Training	625996
2	False	Validation	312998
3	False	Test	104333
4	False	•	0
5	True	Training	221533
6	True	Validation	110766
7	True	Test	36922
8	True	•	0

	Validation 2	Arrest	N Rows
1	Training	True	275830
2	Training	False	275765
3	Validation	True	91567
4	Validation	False	91586

V. Model Evaluation:

Our initial strategy is to select the most effective model be running all three primary model types, and comparing and contrasting the benefits of each, specifically analyzing the RSquared values, which is the percentage of the predicted value in the model which can be explained by the predictor values. We will also compare the values of the misclassification rate, which, in a binary model, tells the percentage of rows which were predicted to be true (the crime results in an arrest) but are actually false (no arrest made) or vice versa. The three different types of models are logistic regression models, classification models, and neural networks. The results of each of these models are displayed below.

- **Logistic Model**

Nominal Logistic Fit for Arrest

Converged in Gradient, 7 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	157746.83	31	315493.7	<.0001*
Full	329152.26			
Reduced	486899.09			

RSquare (U)	0.3240
AICc	658369
BIC	658741
Observations (or Sum Wgts)	847518

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.3240	0.3236	0.3249	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4551	0.4546	0.4562	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3884	0.3886	0.3878	∑ -Log(ρ[j])/n
RMSE	0.3496	0.3496	0.3491	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2429	0.2428	0.2425	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1641	0.1640	0.1637	∑ (ρ[j]≠ρMax)/n
N	847518	423755	141252	n

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	778063	320087.42	640174.8
Saturated	778094	9064.84	Prob>ChiSq
Fitted	31	329152.26	1.0000

Parameter Estimates

Effect Wald Tests

Confusion Matrix

Training

Actual	Predicted	
Arrest	True	False
True	100240	121288
False	17799	608191

Validation

Actual	Predicted	
Arrest	True	False
True	50014	60751
False	8743	304247

Test

Actual	Predicted	
Arrest	True	False
True	16756	20165
False	2964	101367

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	99.9126557	4.4116211	512.91	<.0001*
Primary Type[ASSAULT]	-1.1499989	0.011697	9665.9	<.0001*
Primary Type[BATTERY]	-1.0813531	0.0096387	12586	<.0001*
Primary Type[CRIMINAL TRESPASS]	0.94983517	0.0152616	3873.4	<.0001*
Primary Type[NARCOTICS]	5.00603611	0.0383484	17041	<.0001*
Primary Type[OTHER OFFENSE]	-1.4056312	0.008769	25695	<.0001*
Primary Type[SEX OFFENSE]	0.1598694	0.0176999	81.58	<.0001*
Location Description[GAS STATION]	0.1263734	0.0252544	25.04	<.0001*
Location Description[OTHER]	-0.5674832	0.0118873	2279.0	<.0001*
Location Description[PARK PROPERTY]	0.06181045	0.0291637	4.49	0.0341*
Location Description[RESIDENCE]	-0.7751152	0.0083098	8700.7	<.0001*
Location Description[RESTAURANT/HOTEL]	-0.0905189	0.0168789	28.76	<.0001*
Location Description[SCHOOL/UNIVERSITY]	-0.0736965	0.0164818	19.99	<.0001*
Location Description[STORE]	1.54915004	0.0112846	18846	<.0001*
Location Description[STREET]	-0.1884474	0.0076321	609.67	<.0001*
Domestic[False]	-0.0565718	0.0047724	140.52	<.0001*
District[Central]	-0.0233722	0.0045076	26.88	<.0001*
District[North]	-0.0181219	0.0064845	7.81	0.0052*
Ward	-0.0001434	0.0002952	0.24	0.6271
Community Area	0.00228555	0.0002151	112.93	<.0001*
Year	-0.0495813	0.0021906	512.26	<.0001*
Time of Day	3.96095e-6	1.3239e-7	895.19	<.0001*
Month[10-1]	-0.0702285	0.0155611	20.37	<.0001*
Month[11-10]	-0.0402459	0.0157491	6.53	0.0106*
Month[12-11]	-0.0163316	0.0164159	0.99	0.3198
Month[2-12]	0.18976407	0.0166434	130.00	<.0001*
Month[3-2]	-0.0312496	0.0159851	3.82	0.0506
Month[4-3]	-0.0279397	0.0152658	3.35	0.0672
Month[5-4]	-0.0235755	0.014907	2.50	0.1138
Month[6-5]	-0.0262847	0.0145323	3.27	0.0705
Month[7-6]	-0.0109483	0.0145117	0.57	0.4506
Month[8-7]	-0.0312949	0.0145644	4.62	0.0317*
Month[9-8]	0.02354701	0.0149395	2.48	0.1150

The above chart shows the best available logistic regression model, which displays disappointingly low values for the Entropy RSquared rate at only .3240, meaning only 32% of the model (which is still highly descriptive according to the Chi Squared value at the top) can be explained by the predictor value which we used in the model. The model is misclassifying rows at 16.41%, particularly concerning is the rate at which crimes were predicted that no arrest would be made but one actually was (false positive misclassification). After looking at the results of the other available models, logistic regression will be discarded as the least effective model in this particular study.

- **Neural Net**

Neural

Validation Column: Validation

Model Launch

Model NTanH(3)NTanH2(3)

Training — **Arrest**

Measures	Value
Generalized RSquare	0.4890324
Entropy RSquare	0.3537976
RMSE	0.3402572
Mean Abs Dev	0.2315451
Misclassification Rate	0.1548687
-LogLikelihood	314635.38
Sum Freq	847518

Confusion Matrix

Actual Arrest	Predicted True	Predicted False
True	103212	118316
False	12938	613052

Confusion Rates

Actual Arrest	Predicted Rate True	Predicted Rate False
True	0.466	0.534
False	0.021	0.979

Validation — **Arrest**

Measures	Value
Generalized RSquare	0.4891725
Entropy RSquare	0.3539216
RMSE	0.340162
Mean Abs Dev	0.2315096
Misclassification Rate	0.1546837
-LogLikelihood	157287.38
Sum Freq	423755

Confusion Matrix

Actual Arrest	Predicted True	Predicted False
True	51692	59073
False	6475	306515

Confusion Rates

Actual Arrest	Predicted Rate True	Predicted Rate False
True	0.467	0.533
False	0.021	0.979

Test — **Arrest**

Measures	Value
Generalized RSquare	0.4879391
Entropy RSquare	0.3528223
RMSE	0.3402089
Mean Abs Dev	0.2314153
Misclassification Rate	0.1546952
-LogLikelihood	52517.954
Sum Freq	141252

Confusion Matrix

Actual Arrest	Predicted True	Predicted False
True	17260	19661
False	2190	102141

Confusion Rates

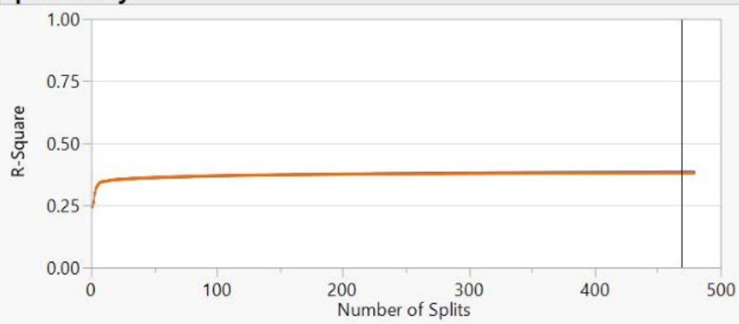
Actual Arrest	Predicted Rate True	Predicted Rate False
True	0.467	0.533
False	0.021	0.979

Above is the neural net model. After some trial and error, our team decided to use the Hyperbolic Tangent method, using two layers of three nodes apiece. Despite this being the best available neural networking model, it was not much more effecting than the logistic regression model above, boasting the entropy RSquared of 35% percent, not nearly as descriptive as we would have hoped that our model would be. The model is good in that the validation and test statistics are equally as low as the training rate, but few positives can be found. Misclassification rate is still around 15.4% (as compared to 16.4% above) but it is still consistently classifying false positives.

- **Classification Tree Model**

Partition for Arrest

Split History



Validation Data in Red
Test Data in Orange

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.3834	0.3800	0.3791	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5217	0.5180	0.5170	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3542	0.3562	0.3567	∑ -Log(ρ[j])/n
RMSE	0.3319	0.3330	0.3330	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2204	0.2211	0.2211	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1491	0.1502	0.1500	∑ (ρ[j]≠ρMax)/n
N	847529	423764	141255	n

Confusion Matrix

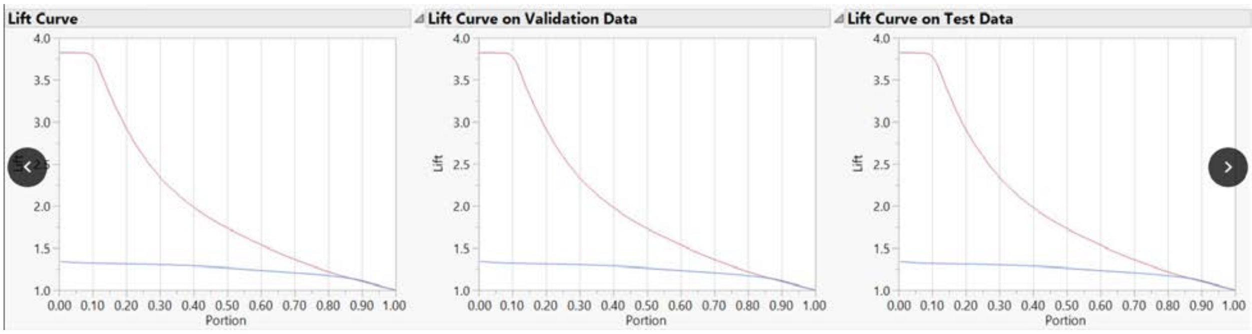
Training: Actual Arrest	Predicted True	Predicted False
True	112479	109054
False	17320	608676

Validation: Actual Arrest	Predicted True	Predicted False
True	56024	54742
False	8891	304107

Test: Actual Arrest	Predicted True	Predicted False
True	18672	18250
False	2934	101399

Finally, above is the classification model, showing the efficacy of the best available model for this particular dataset. The model boasts close to 500 individual splits, approximately 470 after pruning, however any number of splits more than about 50 would have returned virtually identical results in terms of our peak Entropy RSquare value of .3834. By definition, this means that less than 40% of our results are explained by the data in our model, rendering the dataset less than effective at predicting whether an arrest will be made in a particular datapoint. The classification rate remains quite high at 14.9%, and for the first time in our model, false positives are outnumbered by accurate arrest predictions.

The chart displayed below shows the lift ratio for our model, demonstrating the final test of efficiency for the final model we chose (the classification model). At lower portions, we see that the model accurately predicts the arrest rates nearly 4 times better than other methods. This may be viewed as a tepid success because of this result.



Total Number of Rows 423764 - Validation Data Set

423764

Logistic

Predicted	Actual 0	Actual 1
0	304,247	60,751
1	8,743	50,014

Accuracy of the Model 83.60%

Predicted 1		58,757
Accuracy of 1	0.851201	85.12%

Lift Ratio for the Models 3.256

Classification Tree

Predicted	Actual 0	Actual 1
0	304,107	54,742
1	8,891	56,024

Accuracy 84.98%

Predicted 1		64,915
Accuracy of 1	0.863036	86%

Baseline 0.261387

Lift Ratio for the Models 3.302

Neural Nets

Predicted	Actual 0	Actual 1
0	306,515	59,073
1	6,475	51,692

Accuracy 84.53%

Predicted 1		58,167
Accuracy of 1	0.888682586	89%

Lift Ratio for the Models 3.400

VI. Conclusion

Given a very limited data set in terms of descriptive characteristics regarding each crime, with three quarters of data points for any given crime describing the location or time of the crime. This limiting dataset tended to hold back our model in many respects as many of the location figures were so highly correlated or not particularly useful predictors. It is remarkably clear that we need different types of information in order to produce a model which can explain 80 percent of the variance or more.

Simply put, there is a lot more to police work than the location of the crime and the type of crime. Some interesting data points that may have been useful include who was reporting the crime, whether it be the victim, a witness or the police officer him/herself. You might imagine that this would show high predictive power. Or potentially response time to the reported crime may affect the arrest rate. Much of the probability of making an arrest can hinge on factors down to the competency level of the officer investigating.

Despite a low explanation rate or RSquared and a high misclassification rate, we can still learn much from the valuable work done in creating this model. Firstly, we determined that much of this data did have real effective predictive power, as is evidenced by the lift rate of nearly 4 times. We can also take knowledge away from the visualization process, which showed many cases where Chicago PD were having success, such as the high arrest rate in the field of narcotics. We can also take note of the cyclical nature of criminal activity, particularly noting the drop in outdoor crimes during winter and fall months, and the drop-in crimes that occur between 2 and 10 AM, requiring smaller active forces during those times.