Data Mining for Crime Analysis

Julia Holland Chen
Applied Computer Science
University of Colorado Boulder
Boulder, CO USA
julia.chen-1@colorado.edu

ABSTRACT

Although violent crime has generally decreased over the last 20 years [6], the crime rate remains an important issue for Americans [5]. Available crime data continues to increase, as does our need to process and interpret large amounts of raw data. Crime incident data is analyzed using association rule mining and decision trees in order to understand and predict crime incidents in the City of Chicago. In the City of Chicago, crime incidents are most likely to occur in July and August, on the 1st of the month, and noon through night hours. Crime incidents are least likely to occur in February and December, the 31st day of the month, and early morning hours. The most common crime types are theft and battery. Association rules show narcotics is commonly occurring with arrest, which is further supported by decision tree model results when predicting arrest.

1 Introduction

Understanding and predicting the time, location, and type of crime in our cities has become more possible as large amounts of data are available, and the addition of artificial intelligence (AI) in recent times. In 2019, Chicago was determined to be the 31st most dangerous city in America [2], which makes it a useful location for crime analysis. This study examines crime incident data collected from the City of Chicago from 2001 to September 2023 in an effort to understand and predict crime patterns. This study concentrates on patterns in location, time, and type of crime that may help police target their resources where they are needed most and lead to a better understanding of crime patterns in their city. Features in the dataset include the date of incident, primary type of crime, description, police district, and location where the incident occurred. This study focuses on the following areas: generating strong association rules for crime incidents, using models to predict if an arrest will be made given other features in the dataset, and investigate the most common characteristics of crime incidents. The results of this study may be used for strategic planning on how and when to deploy police resources for each district in the City of Chicago.

2 Related Work

Khatun et al. [4] studied how data mining may be used by crime investigation agencies to discover relevant precautionary measures from prediction rates. They used several datasets, including one from the City of Chicago. The authors used decision trees, k-nearest neighbors (KNN), and random forest algorithms in their investigation. Forecasting was performed for the most frequently occurring crimes like robbery, assault, and theft and additionally prediction of arrest. While their paper focused on crime forecasting, this study will also test frequent patterns to see if additional data for crime types may be extracted along with decision trees for arrest prediction. Their study would be useful for police to understand when an arrest is likely, but they did not discuss which features strongly contributed to the arrest prediction. This study will include results on the most important features in determining arrest so police districts may learn what features are currently leading to successful arrests.

Jantan and Jamil [3] explored the relationship between the category of location and type of crime incidents by applying the Apriori algorithm for association rule mining. They focused on the type of crime in specific locations with the goal to produce rules for a pattern visualization system for crime analysis. This study will repeat association rule mining using the Apriori algorithm and expand their work by additionally including classification using decision trees for predicting arrest for an incident.

3 Data Set

The crime incident dataset was published by the City of Chicago [1]. The dataset includes approximately 7.91M rows and 22 features. It spans 23 years, 2001 until September 2023. Note that features which include an exact location are shifted slightly from the true location of the incident to protect privacy, but will still fall on the same block.

3.1 Feature Information

- ID – unique identifier, nominal
- Case Number – police department records division number, nominal
- Date – date of incident, ordinal
- Block – partially redacted address of incident, nominal
- IUCR – Illinois uniform crime reporting code, nominal
- Primary Type – primary description of the IUCR code, nominal
- Description – secondary description of the IUCR code, nominal
- Location Description – description of the incident location, nominal
- Arrest – if arrest was made, Boolean
- Domestic – if incident was domestic-related, Boolean
- Beat – beat where the incident occurred, nominal
- District – police district of incident, nominal
- Ward – City Council district of incident, nominal
- Community Area – community area of incident, nominal
- FBI Code – crime classification as outlined in the FBI National Incident-Based Reporting System (NIBRS), nominal
- X Coordinate – x coordinate of incident location, continuous
- Y Coordinate – y coordinate of incident location, continuous
- Year – year of incident, ordinal
- Updated On – date and time record was last updated, ordinal
- Latitude – latitude of incident location, continuous
- Longitude – longitude of incident location, continuous
- Location – latitude and longitude of incident location formatted for maps, nominal

4 Main Techniques Applied

4.1 Preprocessing

The dataset was converted to a Pandas DataFrame and features were transformed to the correct type. Most of the features of interest had few NaN values, so those rows with NaN values were removed from the dataset. As one of the items of investigation is to investigate crime in different time periods, new features of month, day, and time of day were derived. Time of day was divided into six categories, each spanning four hours. The six categories are: late night (12am-4am), early morning (4am-8am), morning (8am-12pm), noon (12pm-4pm), evening (4pm-8pm), and night (8pm-12am). In examining the data for invalid data, rows corresponding to district numbers 21 and 31 were found in the dataset. Districts 21 and 31 do not currently exist in the City of Chicago. It is possible these were due to clerical errors in data recording. As there was only 251 rows found with this issue, those rows were removed from the dataset.

4.2 Exploratory Data Analysis

Plots were generated to investigate multiple items of interest in the dataset. Incidents per year, month, and day for the entire dataset were examined and additionally for each police district. The most common crime types in the dataset were found as well as the most common crime types for each district. While studying the entire dataset is useful for general reporting on crime in the City of Chicago, this study will repeat every method for each police district in order for individual districts to understand their crime patterns in more detail.

4.3 Association Rule Mining

Using the Apriori algorithm, this study determined frequent itemsets from a subset of the features on the entire dataset. This investigated relationships between the primary type of crime, location of the incident, if the incident was domestic-related, whether an arrest was made, and time of incident. These rules will help police districts understand frequently occurring crime incident details. Strong association rules were then calculated using a support-confidence framework augmented with a correlation measure. Lift was used as the correlation measure where any rules having a lift equal or less than one were removed. This study also generated strong association rules for each district so each district will have information specific to their individual crime patterns.

4.4 Classification

A decision tree algorithm was used for the prediction of arrest, where the two target classes are no arrest and arrest. The decision tree algorithm was chosen for this study as it lends itself to greater interpretability of the results. Prediction of arrest was also investigated in Khatun et al. [4] on the Chicago dataset so this study will replicate, noting that our dataset includes more recent samples and differing features for input data. In addition, although we are using a prediction technique, the goal is to produce a list of features that were most important in the predictive model, unlike Khatun et al. [4]. This feature importance list will be useful to police districts to understand which features tend to lead to arrests. The algorithm was run on the entire dataset as well as for each of the police districts. The classifier model performance was evaluated by computing a confusion matrix and reviewing the values for accuracy, recall, specificity, precision, and F1 score.

5 Key Results

5.1 Exploratory data analysis

As shown in Figure 1, the number of crime incidents steadily decreased from 2001, with the years 2015-19 at a stable level. 2020-2021 show a lower incident count, which is understandable as that was during COVID-19 lockdowns. There is an increase in incident count in 2022 as society was resuming normal activities. The year 2023 has the lowest incident count but that may be because the dataset only includes through September of that year.

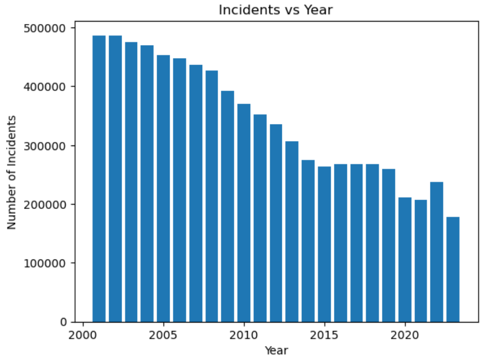

Figure 1: Total incidents reported by year

In Figure 2, the number of incidents does not vary greatly by month, but July and August are slightly elevated for incident counts, while December and February are the lowest.

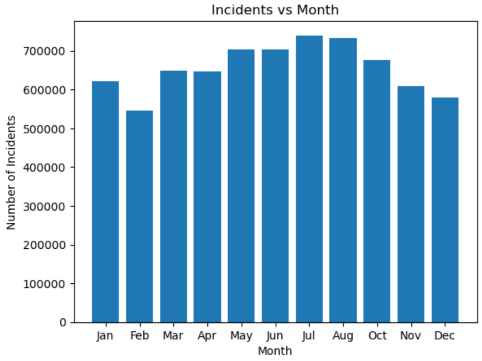

Figure 2: Total incidents reported by month

The first of the month is the most common day for crime incidents, while the 31st day of the month is the lowest (Figure 3). The 31st day of the month numbers may be reduced as not all months have 31 days.

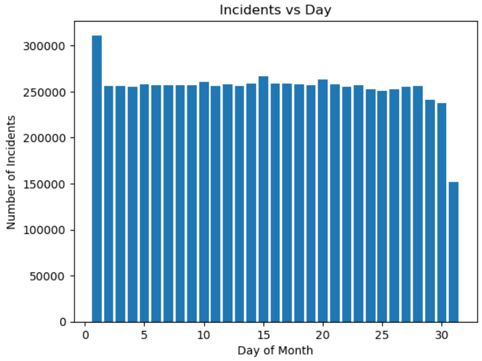

Figure 3: Total incidents reported by day

The lowest time for crime is early morning (4am-8am), while the hours of noon through night (12pm-12am) are the most likely time for crime (Figure 4).

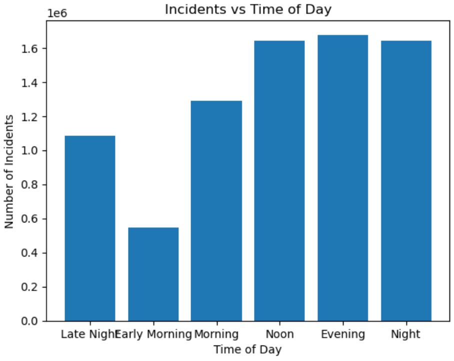

Figure 4: Total incidents by time of day

Figure 5 shows the top three Illinois uniform crime reporting codes (IUCR) in the dataset which are: 0820 (theft – $400 and under), 0486 (battery – domestic battery simple), and 0460 (battery – simple).

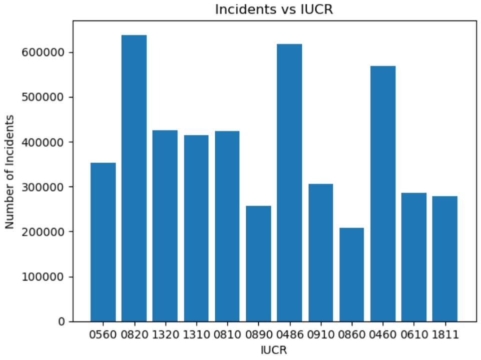

Figure 5: Total incidents reported by IUCR

Districts 8 and 11 have the highest number of incidents (Figure 6). District 20 has the lowest number of incidents. There are no incidents reported from districts 13, 21, and 23, as they do not currently exist.

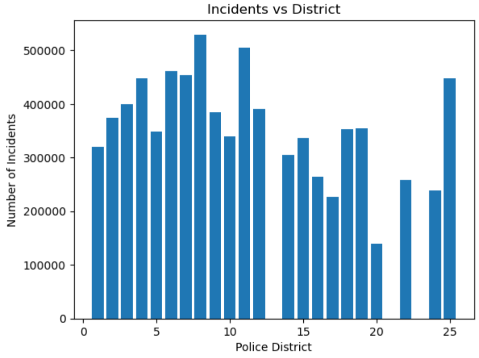

Figure 6: Total incidents by district

5.2 Association Rule Mining

Frequent itemsets and strong rules were calculated using the Apriori algorithm for the entire dataset using the features primary type, location description, domestic, arrest, and time. Rules with a lift value of one or less were removed. The minimum support is 0.06 and minimum confidence is 0.05. Twelve rules were generated (Table 1).

Table 1: Association rules for full dataset

	sup	con	lift
(Type_BATTERY)=>(Domestic)	0.10	0.53	3.1
(Domestic)=>(Type_BATTERY)	0.10	0.56	3.1
(Arrest)=>(Type_NARCOTICS)	0.09	0.36	3.8
(Type_NARCOTICS)=>(Arrest)	0.09	0.99	3.8
(Location_RESIDENCE)=>(Domestic)	0.06	0.38	2.2
(Domestic)=>(Location_RESIDENCE)	0.06	0.37	2.2
(Location_STREET)=>(Arrest)	0.07	0.27	1.0
(Arrest)=>(Location_STREET)	0.07	0.27	1.0
(Location_STREET)=>(Time_Night)	0.07	0.27	1.3
(Time_Night)=>(Location_STREET)	0.07	0.34	1.3
(Arrest)=>(Time_Night)	0.07	0.25	1.2
(Time_Night)=>(Arrest)	0.07	0.31	1.2

Examining a plot of these rules shows there are two connected groups (Figure 7).

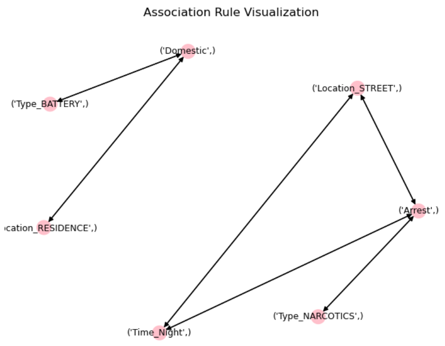

Figure 7: Association rule visualization for dataset

From the visualization, both type of battery and a location of residence are found with a domestic crime. Those rules are not surprising and are expected. The other connected group has type of narcotics and location of street found with arrest. This is interesting as it suggests crime involving narcotics is likely to lead to arrest, and arrests are also likely when the location of the crime is on the street.

5.3 Classification

A decision tree algorithm was used to predict if there was an arrest made for the incident. The dataset does not track if a conviction was successful. The features used in the model are: primary type, location description, district, domestic-related crime, day of incident, and time of incident. The target variable is arrest. The features were one-hot encoded as they are categorical variables. The dataset was split for 70% training and 30% test. The decision tree classifier used entropy as its criterion and a max depth of 15. The model accuracy was approximately 0.88 which is acceptable. Examining the values for precision, recall, and the F1 score suggest the model was moderately successful (Table 2).

Table 2: Performance measures for decision tree (full dataset)

	Precision	recall	F1
0	0.88	0.97	0.92
1	0.87	0.61	0.72

This study is interested in what features were most important in the classification task so additionally the feature importances were calculated and plotted (Figure 8). When the type of crime is narcotics, it contributes significantly to determining if an arrest will be made. If we look back to our association rules (Table 1), this result is consistent.

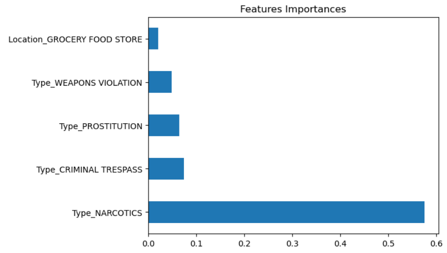

Figure 8: Feature importances for decision tree (full dataset)

This model appears acceptable but our dataset is imbalanced with respect to the arrest feature. There are approximately 26% of rows where arrest is true. In order to test the decision tree algorithm on a balanced dataset we will first oversample the arrest is true rows. The results using this strategy increase the F1 score of class 1, but the model is not significantly improved (Table 3). In addition, the accuracy has dropped to 0.79.

Table 3: Performance measures for balanced dataset (oversampling)

	Precision	recall	F1
0	0.72	0.93	0.81
1	0.91	0.64	0.75

Undersampling was also tested for the arrest prediction model. Random rows from the arrest false category were selected while using almost all of the arrest true rows. This time the split is 80% train and 20% test as we have fewer rows. The accuracy is 0.79. The performance measures do not show a significant difference from the oversampling method (Table 4). As it is easier to work with a smaller dataset, the undersampling method will be used for the decision tree algorithm moving forward.

Table 4: Performance measures for balanced dataset (undersampling)

	Precision	recall	F1
0	0.72	0.94	0.82
1	0.91	0.64	0.75

The features that were the most important in the model using the balanced dataset with undersampling are similar to the original full dataset (Figure 9). The main difference is the primary type battery has replaced the location of grocery food store. The type of narcotics is also less important in the balanced dataset.

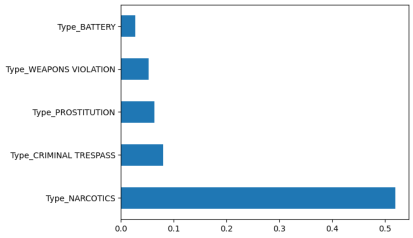

Figure 9: Feature importances for balanced dataset (undersampling)

5.4 District Reports

Summary of exploratory analysis, association rule mining, and classification for each district.

District 1

Crime incidents are most likely in July and August, on the 1st of the month, and noon hours. Crime incidents are least likely in February, the 31st day of the month, and late night through early morning hours. The most common crime types are: theft – $500 and under, theft – from building, and theft – over $500. The association rules generated for District 1 show theft is occurring frequently in the evening and noon hours.

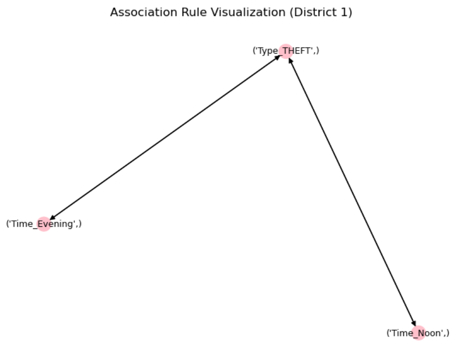

Figure 10: Association rule visualization for District 1

The visualization of association rules (Figure 10) displays the relationship between time of day and the crime of theft. In the classification model predicting arrest, an accuracy of 0.79 was achieved. Additional model performance measurements are in Table 5.

Table 5: Performance measures for District 1

	Precision	recall	F1
0	0.76	0.84	0.80
1	0.82	0.75	0.78

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with criminal trespass second (Figure 11).

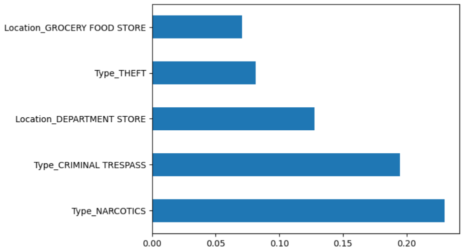

Figure 11: Feature importance for District 1

District 2

Crime incidents are most likely in July and August, on the 1st of the month, and noon and evening hours. Crime incidents are least likely in February, the 31st day of the month, and early morning hours. The most common crime types are: theft – $500 and under, battery – domestic battery simple, and battery – simple. The association rules generated for District 2 show domestic related incidents are occurring frequently with primary type battery, and narcotics are occurring frequently with arrest.

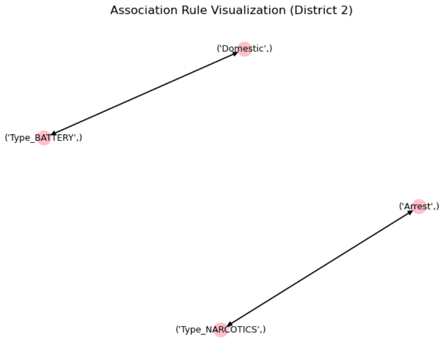

Figure 12: Association rule visualization for District 2

The visualization of association rules (Figure 12) display the relationship between battery and domestic, and narcotics and arrest. In the classification model predicting arrest, an accuracy of 0.79 was achieved. Additional model performance measurements are in Table 6.

Table 6: Performance measures for District 2

	Precision	recall	F1
0	0.72	0.95	0.82
1	0.93	0.63	0.75

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with criminal trespass second (Figure 13).

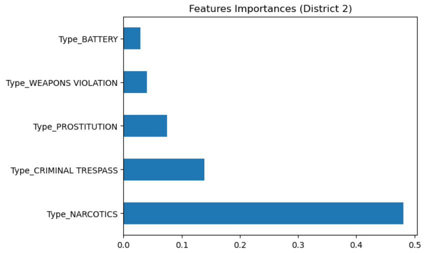

Figure 13: Feature importance for District 2

District 3

Crime incidents are most likely in July and August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February, the 31st day of the month, and early morning hours. The most common crime types are: battery – domestic battery simple, battery – simple, and criminal damage – to property. The association rules generated for District 3 show associations between location of apartment and primary type battery. Domestic related incidents are frequently occurring with location of apartment. Narcotics is frequently appearing with arrest.

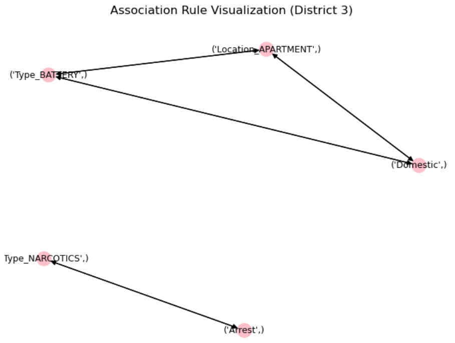

Figure 14: Association rule visualization for District 3

The visualization of association rules (Figure 14) display the relationship between battery, location of apartment and domestic. Narcotics and arrest are also occurring frequently together in a different group. In the classification model predicting arrest, an accuracy of 0.79 was achieved. Additional model performance measurements are in Table 7.

Table 7: Performance measures for District 3

	Precision	recall	F1
0	0.74	0.90	0.81
1	0.86	0.68	0.76

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with criminal trespass second (Figure 15).

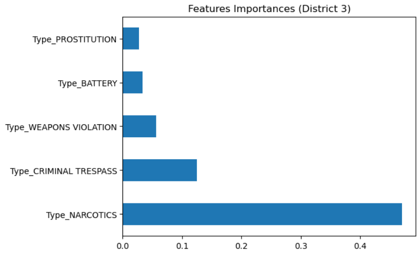

Figure 15: Feature importance for District 3

District 4

Crime incidents are most likely in May through August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February, the 31st day of the month, and early morning hours. The most common crime types are: battery – domestic battery simple, battery – simple, and criminal damage – to property. The association rules generated for District 4 show associations between battery and domestic, and domestic and the location residence.

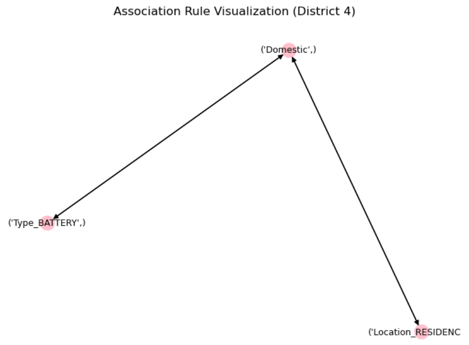

Figure 16: Association rule visualization for District 4

The visualization of association rules (Figure 16) display the relationship between battery, location of residence, and domestic. In the classification model predicting arrest, an accuracy of 0.79 was achieved. Additional model performance measurements are in Table 8.

Table 8: Performance measures for District 4

	Precision	recall	F1
0	0.74	0.96	0.84
1	0.93	0.57	0.71

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with weapons violation second (Figure 17).

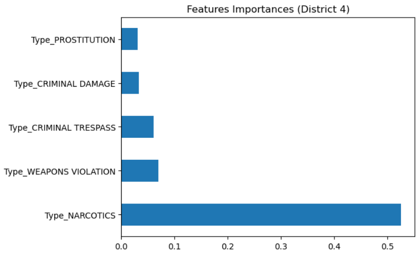

Figure 17: Feature importance for District 4

District 5

Crime incidents are most likely in May through August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February, the 31st day of the month, and early morning hours. The most common crime types are: battery – domestic battery simple, battery – simple, and criminal damage – to property. The association rules generated for District 5 show associations between battery and location residence, battery and domestic, arrest and narcotics, domestic and residence, and arrest and location street.

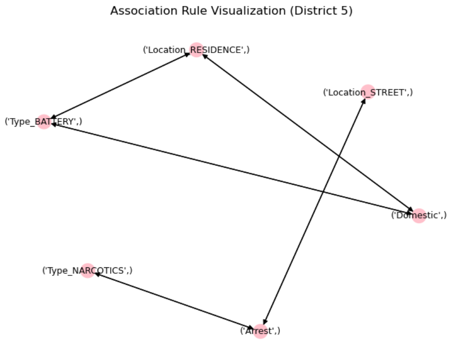

Figure 18: Association rule visualization for District 5

The visualization of association rules (Figure 18) display the various relationships. In the classification model predicting arrest, an accuracy of 0.78 was achieved. Additional model performance measurements are in Table 9.

Table 9: Performance measures for District 5

	Precision	recall	F1
0	0.72	0.94	0.82
1	0.90	0.62	0.73

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with weapons violation second (Figure 19).

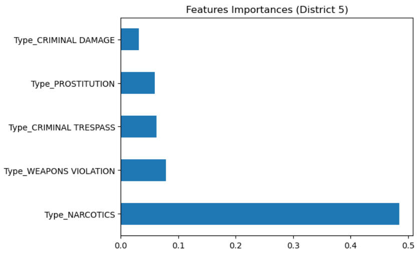

Figure 19: Feature importance for District 5

District 6

Crime incidents are most likely in May through August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February, the 31st day of the month, and early morning hours. The types of crime which are the most frequently reported are: battery – domestic battery simple, theft – $500 and under, and battery – simple. The association rules generated for District 6 show associations between battery and domestic, arrest and narcotics, and domestic and location residence.

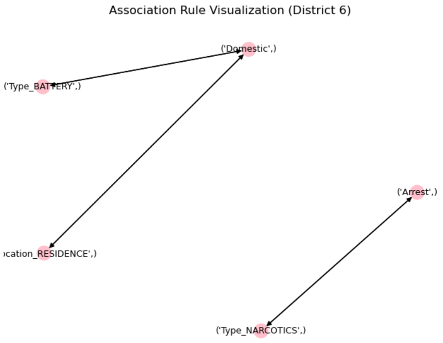

Figure 20: Association rule visualization for District 6

The visualization of association rules (Figure 20) display the various relationships. In the classification model predicting arrest, an accuracy of 0.78 was achieved. Additional model performance measurements are in Table 10.

Table 10: Performance measures for District 6

	Precision	recall	F1
0	0.71	0.96	0.81
1	0.93	0.59	0.73

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with weapons violation second (Figure 21).

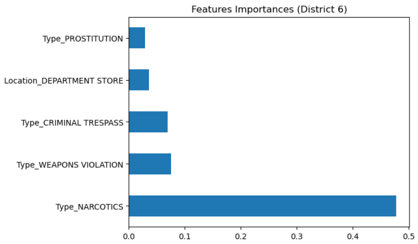

Figure 21: Feature importance for District 6

District 7

Crime incidents are most likely in July and August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February and December, the 31st day of the month, and early morning hours. The most common crime types are: battery – domestic battery simple, battery – simple, and criminal damage – to property. The association rules for District 7 show associations between battery and domestic, arrest and narcotics, domestic and location residence, and arrest and location street.

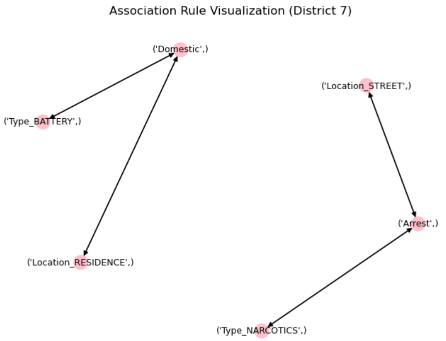

Figure 22: Association rule visualization for District 7

The visualization of association rules (Figure 22) display the various relationships. In the classification model predicting arrest, an accuracy of 0.80 was achieved. Additional model performance measurements are in Table 11.

Table 11: Performance measures for District 7

	Precision	recall	F1
0	0.72	0.95	0.82
1	0.93	0.66	0.77

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with weapons violation second (Figure 23).

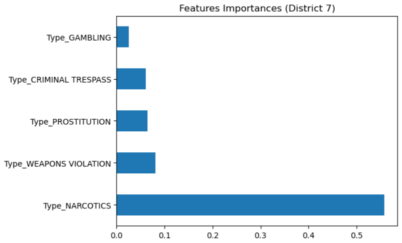

Figure 23: Feature importance for District 7

District 8

Crime incidents are most likely in May through August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February, the 31st day of the month, and early morning hours. The most common crime types are: battery – domestic battery simple, theft –$500 and under, and criminal damage – to property. The association rules generated for District 8 show associations between battery and domestic, and domestic and location residence.

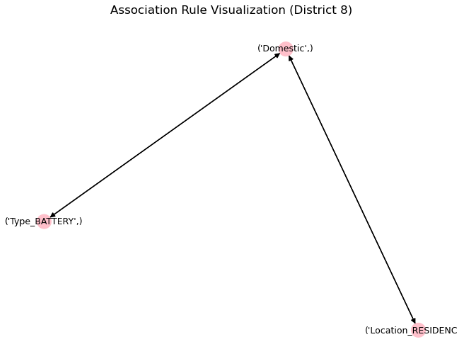

Figure 24: Association rule visualization for District 8

The visualization of association rules (Figure 24) display the various relationships. In the classification model predicting arrest, an accuracy of 0.79 was achieved. Additional model performance measurements are in Table 12.

Table 12: Performance measures for District 8

	Precision	recall	F1
0	0.76	0.90	0.82
1	0.84	0.65	0.73

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with type prostitution second (Figure 25).

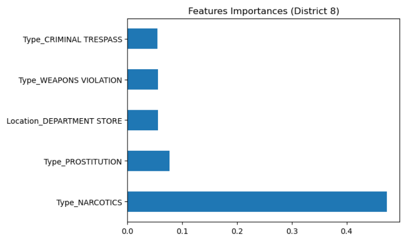

Figure 25: Feature importance for District 8

District 9

Crime incidents are most likely in July and August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February and December, the 31st day of the month, and early morning hours. The most common crime types are: battery – domestic battery simple, theft –$500 and under, and battery – simple. The association rules generated for District 9 show associations between battery and domestic, arrest and narcotics, location street and arrest, and location street and time night.

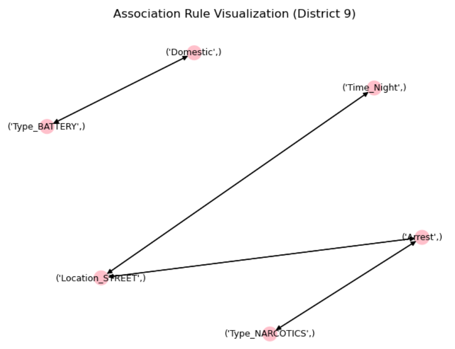

Figure 26: Association rule visualization for District 9

The visualization of association rules (Figure 26) display the various relationships. In the classification model predicting arrest, an accuracy of 0.77 was achieved. Additional model performance measurements are in Table 13.

Table 13: Performance measures for District 9

	Precision	recall	F1
0	0.69	0.95	0.80
1	0.93	0.59	0.72

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with type prostitution second (Figure 27).

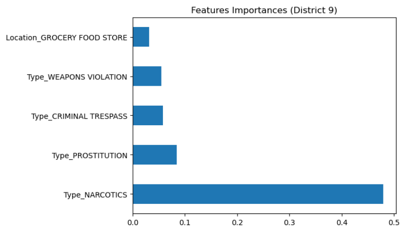

Figure 27: Feature importance for District 9

District 10

Crime incidents are most likely in July and August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February and December, the 31st day of the month, and early morning hours. The most common crime types are: battery – domestic battery simple, battery – simple, and theft –$500 and under. The association rules generated for District 10 show associations between battery and domestic, arrest and narcotics, arrest and location sidewalk, and arrest and time of night.

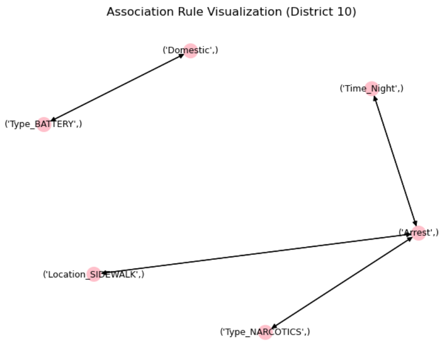

Figure 28: Association rule visualization for District 10

The visualization of association rules (Figure 28) display the various relationships. In the classification model predicting arrest, an accuracy of 0.79 was achieved. Additional model performance measurements are in Table 14.

Table 14: Performance measures for District 10

	Precision	recall	F1
0	0.71	0.91	0.79
1	0.91	0.7	0.80

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with type weapons violation second (Figure 29).

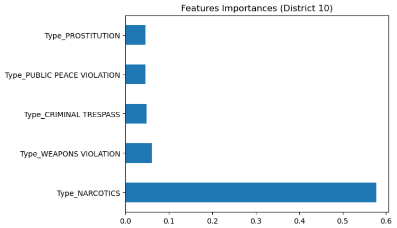

Figure 29: Feature importance for District 10

District 11

Crime incidents are most likely in May through August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February and December, the 31st day of the month, and early morning hours. The most common crime types are: battery – domestic battery simple, narcotics – possess heroin white, and battery – simple. There were 32 association rules generated for District 11. Narcotics was a common item found with arrest, location sidewalk, and location of street.

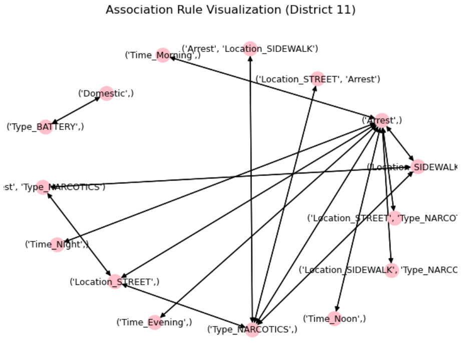

Figure 30: Association rule visualization for District 11

The visualization of association rules (Figure 30) displays the various relationships. In the classification model predicting arrest, an accuracy of 0.86 was achieved. Additional model performance measurements are in Table 15.

Table 15: Performance measures for District 11

	Precision	recall	F1
0	0.70	0.97	0.82
1	0.98	0.80	0.88

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with type prostitution second (Figure 31).

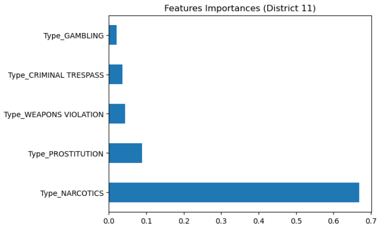

Figure 31: Feature importance for District 11

District 12

Crime incidents are most likely in July and August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February, the 31st day of the month, and early morning hours. The most common crime types are: theft – $500 and under, theft – over $500, and battery – simple. The association rules generated for District 12 show associations between location street and theft, and location street and time of night.

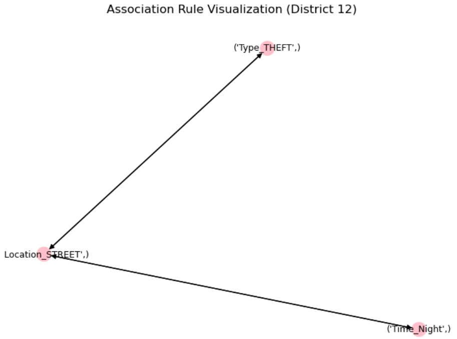

Figure 32: Association rule visualization for District 12

The visualization of association rules (Figure 32) display the various relationships. In the classification model predicting arrest, an accuracy of 0.78 was achieved. Additional model performance measurements are in Table 14.

Table 16: Performance measures for District 12

	Precision	recall	F1
0	0.72	0.97	0.83
1	0.94	0.55	0.69

The F1 score for class 1 is under 0.7 and therefore does not suggest a good fit by the model. The most important feature leading to a prediction of arrest is the primary type of narcotics with type criminal trespass second (Figure 33).

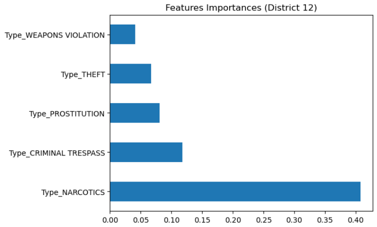

Figure 33: Feature importance for District 12

District 14

Crime incidents are most likely in July and August, on the 1st of the month, and evening through night hours. Crime incidents are least likely in February and December, the 31st day of the month, and early morning hours. The most common crime types are: theft – $500 and under, theft – over $500, and battery – simple. The association rules generated for District 14 show associations between location street and theft, and location street and time of night.

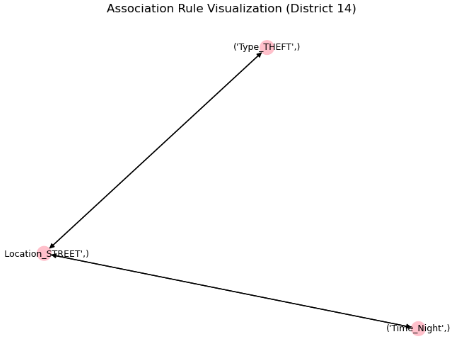

Figure 34: Association rule visualization for District 14

The visualization of association rules (Figure 34) display the various relationships. In the classification model predicting arrest, an accuracy of 0.78 was achieved. Additional model performance measurements are in Table 17.

Table 17: Performance measures for District 14

	Precision	recall	F1
0	0.75	0.93	0.83
1	0.86	0.57	0.69

The F1 score for class 1 is under 0.7 and therefore does not suggest a good fit by the model. The most important feature leading to a prediction of arrest is the primary type of narcotics with type prostitution second (Figure 35).

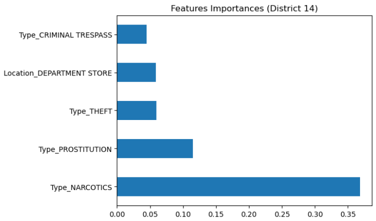

Figure 35: Feature importance for District 14

District 15

Crime incidents are most likely in May, July and August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February, the 31st day of the month, and early morning hours. The most common crime types are: battery – domestic battery simple, battery – simple, and narcotics – possess cannabis 30 grams or less. There were 24 association rules generated for District 15. One of the more interesting relationships is between arrest, location sidewalk, and narcotics.

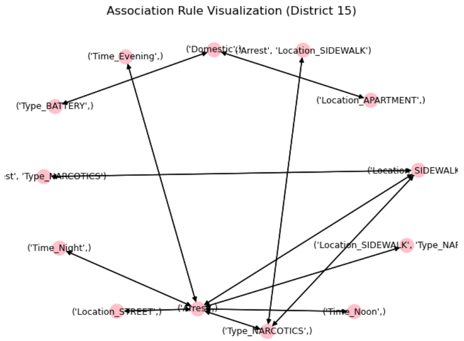

Figure 36: Association rule visualization for District 15

The visualization of association rules (Figure 36) display the various relationships. In the classification model predicting arrest, an accuracy of 0.78 was achieved. Additional model performance measurements are in Table 18.

Table 18: Performance measures for District 15

	Precision	recall	F1
0	0.70	0.93	0.80
1	0.95	0.78	0.86

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics (Figure 37).

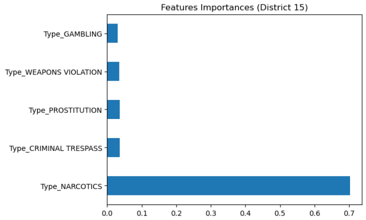

Figure 37: Feature importance for District 15

District 16

Crime incidents are most likely in July and August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February, the 31st day of the month, and early morning hours. The most common crime types are: theft – $500 and under, theft – over $500, and criminal damage – to vehicle. The association rules generated for District 16 show associations between domestic and location residence.

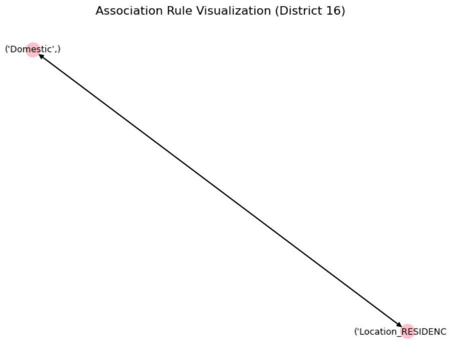

Figure 38: Association rule visualization for District 16

The visualization of association rules (Figure 38) display the various relationships. In the classification model predicting arrest, an accuracy of 0.78 was achieved. Additional model performance measurements are in Table 19.

Table 19: Performance measures for District 16

	Precision	recall	F1
0	0.77	0.91	0.84
1	0.81	0.58	0.67

The F1 score for class 1 is under 0.7 and therefore does not suggest a good fit by the model. The most important feature leading to a prediction of arrest is the primary type of narcotics with type criminal trespass second (Figure 39).

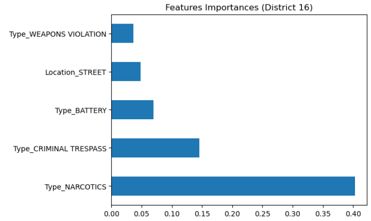

Figure 39: Feature importance for District 16

District 17

Crime incidents are most likely in July and August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February, the 31st day of the month, and early morning hours. The most common crime types are: theft – $500 and under, criminal damage – to vehicle, and battery – simple. The association rules generated for District 17 show associations between battery and domestic, and location street and time of night.

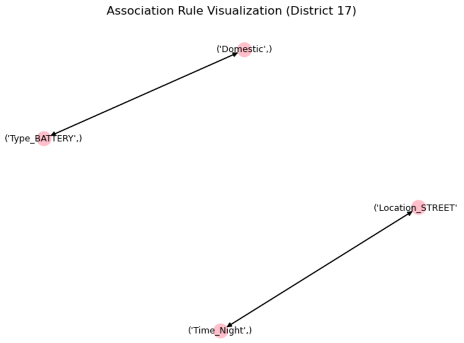

Figure 40: Association rule visualization for District 17

The visualization of association rules (Figure 40) display the various relationships. In the classification model predicting arrest, an accuracy of 0.77 was achieved. Additional model performance measurements are in Table 20.

Table 20: Performance measures for District 17

	Precision	recall	F1
0	0.75	0.89	0.81
1	0.79	0.60	0.68

The F1 score for class 1 is under 0.7 and therefore does not suggest a good fit by the model. The most important feature leading to a prediction of arrest is the primary type of narcotics with type battery second (Figure 41).

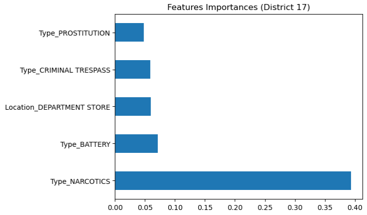

Figure 41: Feature importance for District 17

District 18

Crime incidents are most likely in July and August, on the 1st of the month, and noon through evening hours. Crime incidents are least likely in February, the 31st day of the month, and early morning hours. The most common crime types are: theft – from building, theft – over $500, and theft – $500 and under. The association rules generated for District 18 show associations between beat 1834 and theft, time of evening and theft, and time of noon and theft.

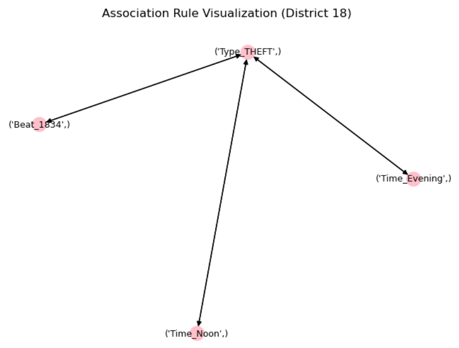

Figure 42: Association rule visualization for District 18

The visualization of association rules (Figure 42) display the various relationships. In the classification model predicting arrest, an accuracy of 0.79 was achieved. Additional model performance measurements are in Table 21.

Table 21: Performance measures for District 18

	Precision	recall	F1
0	0.77	0.88	0.82
1	0.83	0.69	0.76

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with type criminal trespass second (Figure 43).

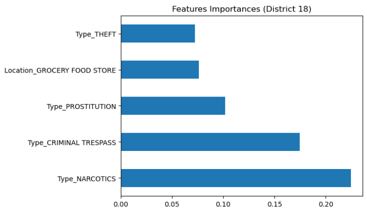

Figure 43: Feature importance for District 18

District 19

Crime incidents are most likely in July and August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February and December, the 31st day of the month, and early morning hours. The most common crime types are: theft – $500 and under, theft – over $500, and battery – simple. The association rules generated for District 19 show associations between location street and theft, and time of evening and theft.

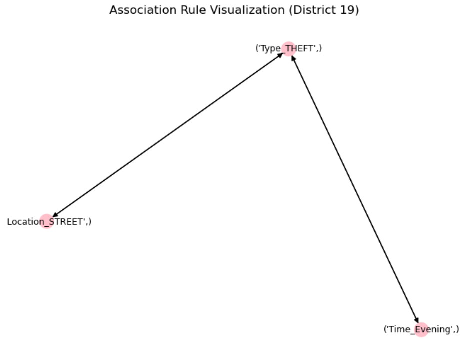

Figure 44: Association rule visualization for District 19

The visualization of association rules (Figure 44) display the various relationships. In the classification model predicting arrest, an accuracy of 0.79 was achieved. Additional model performance measurements are in Table 22.

Table 22: Performance measures for District 19

	Precision	recall	F1
0	0.81	0.83	0.82
1	0.77	0.75	0.76

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with type criminal trespass second (Figure 45).

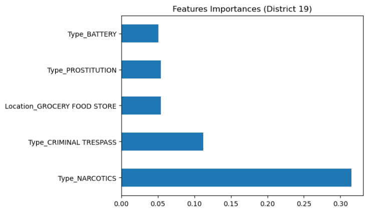

Figure 45: Feature importance for District 19

District 20

Crime incidents are most likely in May through August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in

February and December, the 31st day of the month, and early morning hours. The most common crime types are: theft – $500 and under, battery – simple, and theft – over $500. There were no association rules generated with the specified minimum support and confidence. In the classification model predicting arrest, an accuracy of 0.76 was achieved. Additional model performance measurements are in Table 23.

Table 23: Performance measures for District 20

	Precision	recall	F1
0	0.71	0.95	0.81
1	0.90	0.52	0.66

The F1 score for class 1 is under 0.7 and therefore does not suggest a good fit by the model. The most important feature leading to a prediction of arrest is the primary type of narcotics with type criminal trespass second (Figure 46).

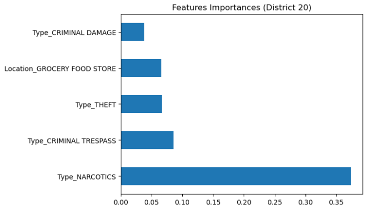

Figure 46: Feature importance for District 20

District 22

Crime incidents are most likely in May through October, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February and December, the 31st day of the month, and early morning hours. The most common crime types are: theft – $500 and under, battery – domestic battery simple, and criminal damage – to property. The association rules generated for District 22 show associations between battery and domestic, and domestic and location residence.

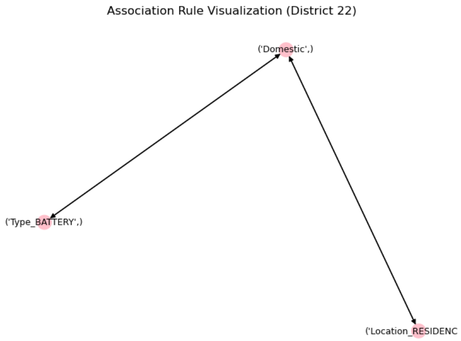

Figure 47: Association rule visualization for District 22

The visualization of association rules (Figure 47) display the various relationships. In the classification model predicting arrest, an accuracy of 0.80 was achieved. Additional model performance measurements are in Table 24.

Table 24: Performance measures for District 22

	Precision	recall	F1
0	0.75	0.96	0.84
1	0.91	0.58	0.71

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with type weapons violation second (Figure 48).

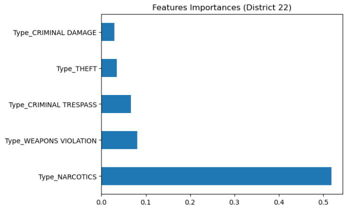

Figure 48: Feature importance for District 22

District 24

Crime incidents are most likely in July and August, on the 1st of the month, and evening through night hours. Crime incidents are least likely in February and December, the 31st day of the month, and early morning hours. The most common crime types are: theft – $500 and under, battery – simple, and battery – domestic battery simple. The association rules generated for District 24 show an interesting association between battery and domestic.

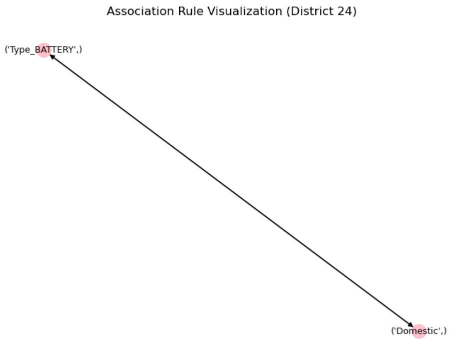

Figure 49: Association rule visualization for District 24

The visualization of association rules (Figure 49) display the various relationships. In the classification model predicting arrest, an accuracy of 0.78 was achieved. Additional model performance measurements are in Table 25.

Table 25: Performance measures for District 24

	Precision	recall	F1
0	0.73	0.96	0.83
1	0.91	0.55	0.68

The F1 score for class 1 is under 0.7 and therefore does not suggest a good fit by the model. The most important feature leading to a prediction of arrest is the primary type of narcotics with type criminal trespass second (Figure 50).

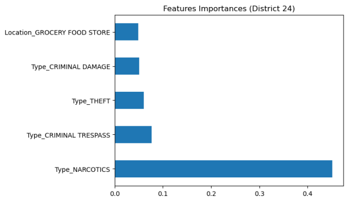

Figure 50: Feature importance for District 24

District 25

Crime incidents are most likely in July and August, on the 1st of the month, and noon through night hours. Crime incidents are least likely in February and December, the 31st day of the month, and early morning hours. The most common crime types are: battery – domestic battery simple, theft - $500 and under, and battery – simple. The association rules generated for District 25 show associations between battery and domestic, arrest and narcotics, and location street and arrest.

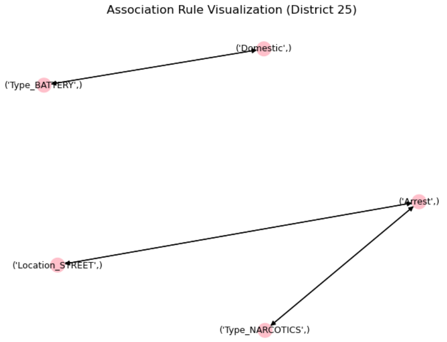

Figure 51: Association rule visualization for District 25

The visualization of association rules (Figure 51) display the various relationships. In the classification model predicting arrest, an accuracy of 0.79 was achieved. Additional model performance measurements are in Table 26.

Table 26: Performance measures for District 25

	Precision	recall	F1
0	0.74	0.90	0.81
1	0.87	0.69	0.77

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important feature leading to a prediction of arrest is the primary type of narcotics with type prostitution second (Figure 52).

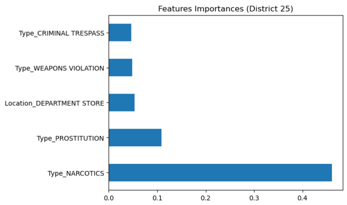

Figure 52: Feature importance for District 25

6 Applications

The results of this study will inform police districts in the City of Chicago when and where to deploy resources for crime prevention and response. The classification results also allow police districts to see which types of incidents have successfully led to an arrest and where they can improve. The data mining techniques described in this study may be applied to a visualization system where police and city administrators would have the ability to explore the data in more detail and automatically import the most recent data available.

REFERENCES

[1] Data.gov. 2023. Crimes - 2001 to Present. Retrieved from https://catalog.data.gov/dataset/crimes-2001-to-present.

[2] Elisha Fieldstadt. 2020. The most dangerous cities in America, ranked. Retrieved from https://www.cbsnews.com/pictures/the-most-dangerous-cities-in-america.

[3] Hamidah Jantan, Aina Zalikha Mohd Jamil. 2019. Association Rule Mining Based Crime Analysis using Apriori Algorithm. International Journal of Advanced Trends in Computer Science and Engineering Vol.8, No.1.5, 2019. DOI: 10.30534/ijatcse/2019/0581.52019.

[4] Most. Rokeya Khatun, Safial Islam Ayon, Md. Rahat Hossain, Md. Jaber Alam. 2021. Data mining technique to analyse and predict crime using crime categories and arrest records. Indonesian Journal of Electrical Engineering and Computer Science Vol.22, No.2, May 2021, DOI: 10.11591/ijeecs.v22.i2.pp1052-1060.

[5] Domenico Montanaro. 2023. Poll: Dangers for both parties on the economy, crime and transgender rights. Retrieved from https://www.npr.org/2023/03/29/1166486046/poll-economy-inflation-transgender-rights-republicans-democrats-biden.

[6] Wikipedia contributors. 2023. Crime in the United States. Wikipedia, The Free Encyclopedia. Retrieved December 12, 2023 from https://en.wikipedia.org/w/index.php?title=Crime_in_the_United_States&oldid=1186104526.