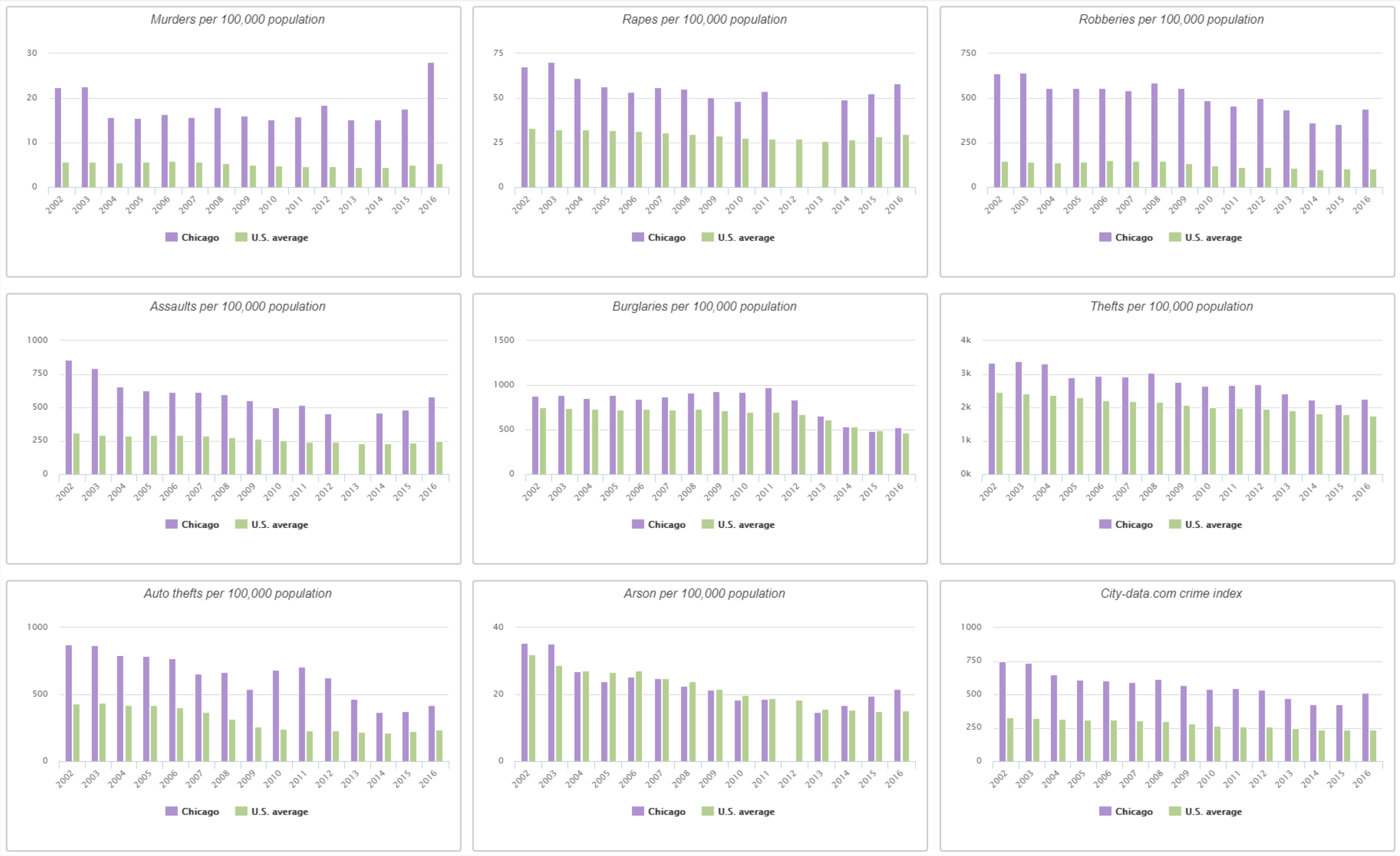
A SPATIAL AND TEMPORAL ANALYSIS OF CRIMES IN CHICAGO

Anita Nallapareddy

Intermediate Data Science

PROBLEM AND INTENT OF RESEARCH



Murders per 100,000 population

Rapes per 100,000 population

Robberies per 100,000 population

Assaults per 100,000 population

Burglaries per 100,000 population

Thefts per 100,000 population

Auto thefts per 100,000 population

Arson per 100,000 population

City-data.com crime index

Crime in Chicago compared to the U.S. average (City-Data, 2018)

Problem:

- Chicago has had crime rates above the U.S. average for 2002-2016
  - Makes the city unsafe
  - Could possibly put more police officers at risk for injury or worse

Intent:

- Discover any spatial/temporal factors that be useful in predicting the type of crime that may occur
  - Useful for Chicago police officers
  - Better anticipate the type of crime that may occur at a certain location/time
  - Perhaps prevent crimes from occurring with more police presence at certain locations/times
  - An adequate number of officers could be dispatched to patrol areas that are hotspots for certain crimes

PROCEDURE

- Cleaned crime report data
- Added additional temporal/spatial features
    - e.g. holiday, season, distance from closest bus stop…
- Performed EDA to find features that may have some relationship with the type of crime
- Evaluated several models with the chosen features
- Examined the most important features based on the coefficients produced by the best model

CRIME DATA

- Crime reports in Chicago from 2001 into 2018 provided by the City of Chicago
  - Original dataset contained 6,726,718 reports
  - Contained the following useful features for each report:
    - Date/Time
    - Block
    - Primary Type of Crime
    - Location Description
    - Domestic
    - Police Beat
    - Police District
    - Ward
    - Community
    - Latitude/Longitude
    - Y/X Coordinates (Projected Latitude/Longitude)

ADDITIONAL DATASETS

- Chicago train (‘L’) stops provided by the City of Chicago
- Chicago bus stops provided by the City of Chicago
- Chicago business licenses provided by the City of Chicago
  - Extracted liquor stores from this dataset
  - Used businesses whose names contained: liquor, spirits, wine, or alcohol
- Chicago police stations provided by the City of Chicago
- U.S. federal holidays provided by Kaggle
- Chicago community area boundaries provided by the City of Chicago
  - Contained coordinates for all Chicago community boundaries
- Chicago ward boundaries provided by the City of Chicago
  - Contained coordinates for all Chicago ward boundaries

CLEANING THE CRIME DATA

- Removed duplicate reports
- Checked the syntax of all features
  - Removed extra spaces and unnecessary characters(e.g. quotes and accents)
  - Made everything uppercase
  - Made sure syntax was consistent for the block and date/time
- Police Districts
  - Examined the crime counts for each district
  - Districts 21 and 31 had much lower counts
    - These districts no longer existed so I made them null

- Community
  - Examined the crime counts for each community
  - Community 0 had a much lower count
    - This community does not exist so I made it null
- Latitude, Longitude, Y/X Coordinates
  - Created scatterplots and looked for irregular positions
  - Several positions found far southwest of Chicago
    - Appeared to be a default location that may have been input when the actual location was not noted
    - Made these locations null

MISSING DATA

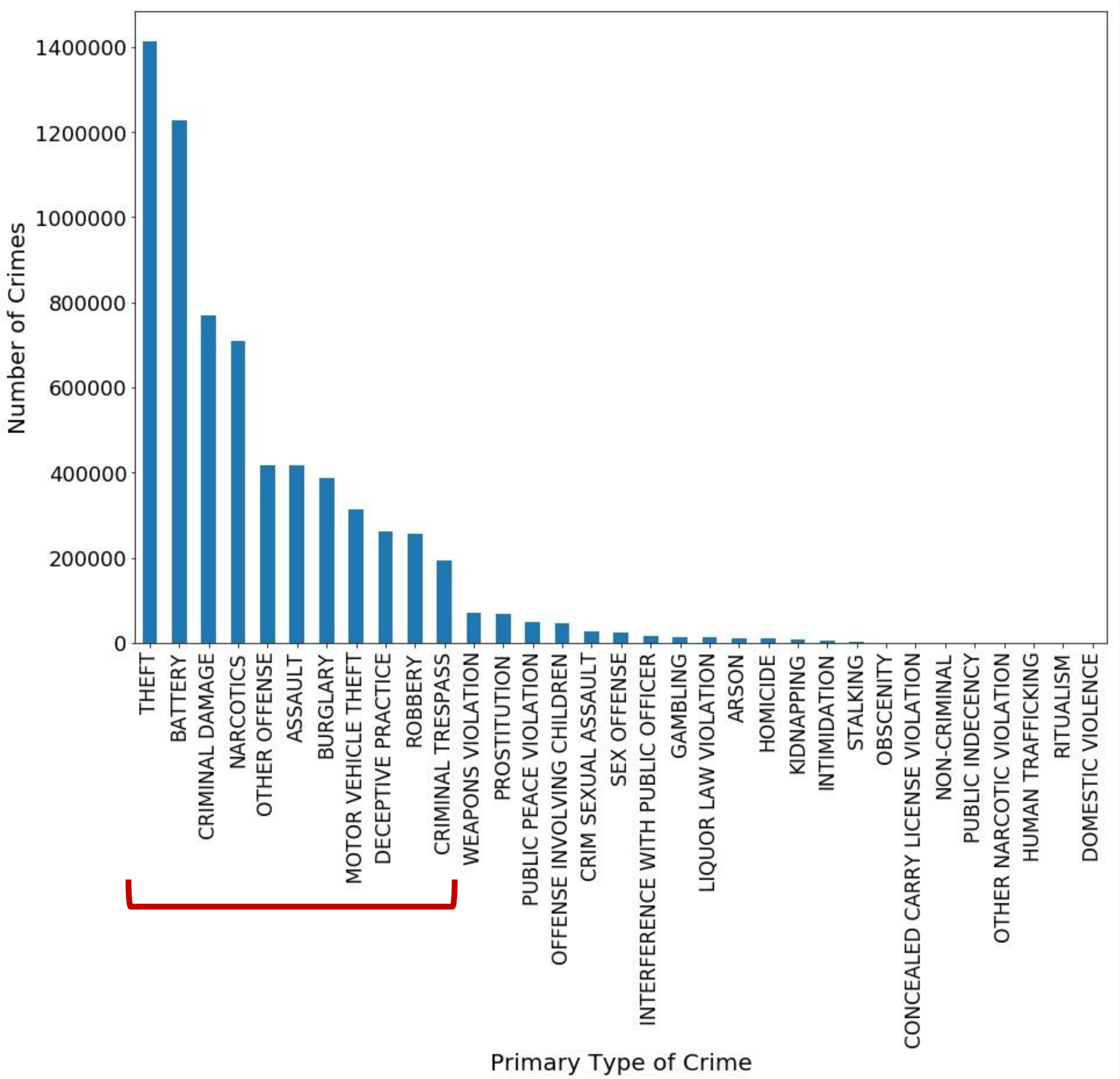
- **Block**
  - 2 entries listed as 'XX UNKNOWN'
  - Left as is as no other location data was available
- **Location Description**
  - 3974 null values
  - For crimes missing the location description and were domestic, filled in the location description as RESIDENCE
    - Checked the counts of each location description for domestic crimes and saw that RESIDENCE was the most common location description
    - Only removed 1 null value, left the rest as is
- **Police District**
  - 198 null values
  - Police beat boundaries fit neatly within police district boundaries
    - Created a dictionary with police beat as the key and police district as the value
    - Filled in all missing police districts based on what the police beat was
- **Location (Latitude/Longitude and Y/X Coordinate)**
  - 60,336 null values
  - Made a list of unique blocks that were missing the location
  - For each of these blocks, looked up all reports with non-null locations with the same block and chose a location at random
  - Filled in missing locations using the location linked to the block
  - Reduced the null values to 2,464
- **Ward**
  - 614,846 null values
  - Used ward boundary coordinates to create polygons
  - For each known latitude/longitude, used the Shapely package to figure out which ward polygon it was located in
  - Reduced the null values to 2,952
- **Community**
  - 616,113 null values
  - Used same procedures as ward
  - Reduced the null values to 2,710

ADDED FEATURES

- Using date/time, created columns for:
  - Month
  - Season
  - Quarter of the year
  - Day of the month
  - Third of the month
  - Day of the week
  - Day type (weekday/weekend)
  - Hour
  - Time of day

- Distance from Chicago city center
  - Used the haversine formula to calculate the distance of each reported crime from the city center (41.88°N 87.62°W)
- Distance from closest bus stop, train stop, liquor store, and police station
  - Used scikit-learn's ball tree query with haversine distance metric
- Federal holidays
  - Column for which federal holiday or no holiday
  - Boolean column for if there is any federal holiday that day

NUMBER OF CRIMES FOR EACH PRIMARY TYPE OF CRIME



- Used the top 11 primary types of crime in my study

MODEL EVALUATION

- 6,357,103 non-null reports for the top 11 primary types of crime
- Chose at random a sample of 3 million reports
- Features used to evaluate models:
  - Ward
  - Police district
  - Police beat
  - Community
  - Location description
  - Latitude
  - Longitude
  - Distance from city center of Chicago
  - Square root of distance from closest police station
  - Distance from closest train stop
  - Closest train line
  - Square root of distance from closest bus stop
  - Square root of distance from closest liquor store
  - Season
  - Month
  - If it is a weekday or the weekend
  - If it is a federal holiday
  - What federal holiday it is
  - Day of the week
  - Time of day
  - Hour
- Converted categorical features to dummy variables
- Split data
  - 75% to train
  - 25% to test

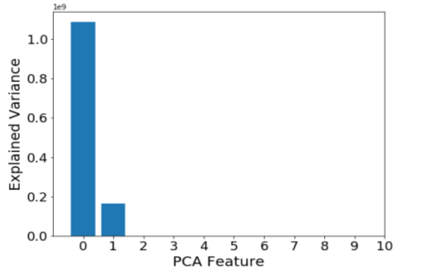
MODEL EVALUATION

Model Name	Adjusted Model Parameters	Shortened Model Name
SGD Classifier		SGDa
SGD Classifier	class_weight='balanced'	SGDb

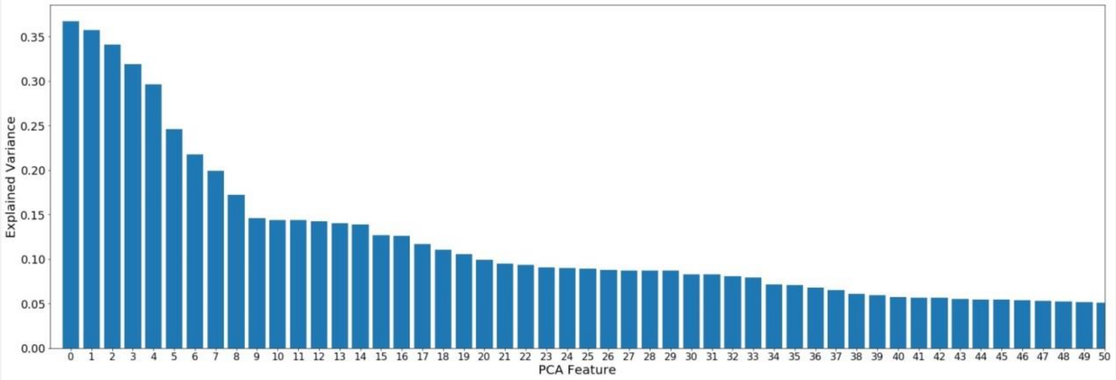
SGD Classifier	class_weight='balanced',
	loss='log'
		SGDc

Multinomial NB		MNB
Decision Tree Classifier		DTCa
Decision Tree Classifier	class_weight='balanced'	DTCb
Extra Tree Classifier		ETCa
Extra Tree Classifier	class_weight='balanced'	ETCb
Random Forest Classifier		RFCa
Random Forest Classifier	class_weight='balanced'	RFCb
Gaussian NB		GNB

- Evaluated the models with the following:
  - Original data
  - Original data with dimensionality reduced to 2 features using scikit-learn's PCA



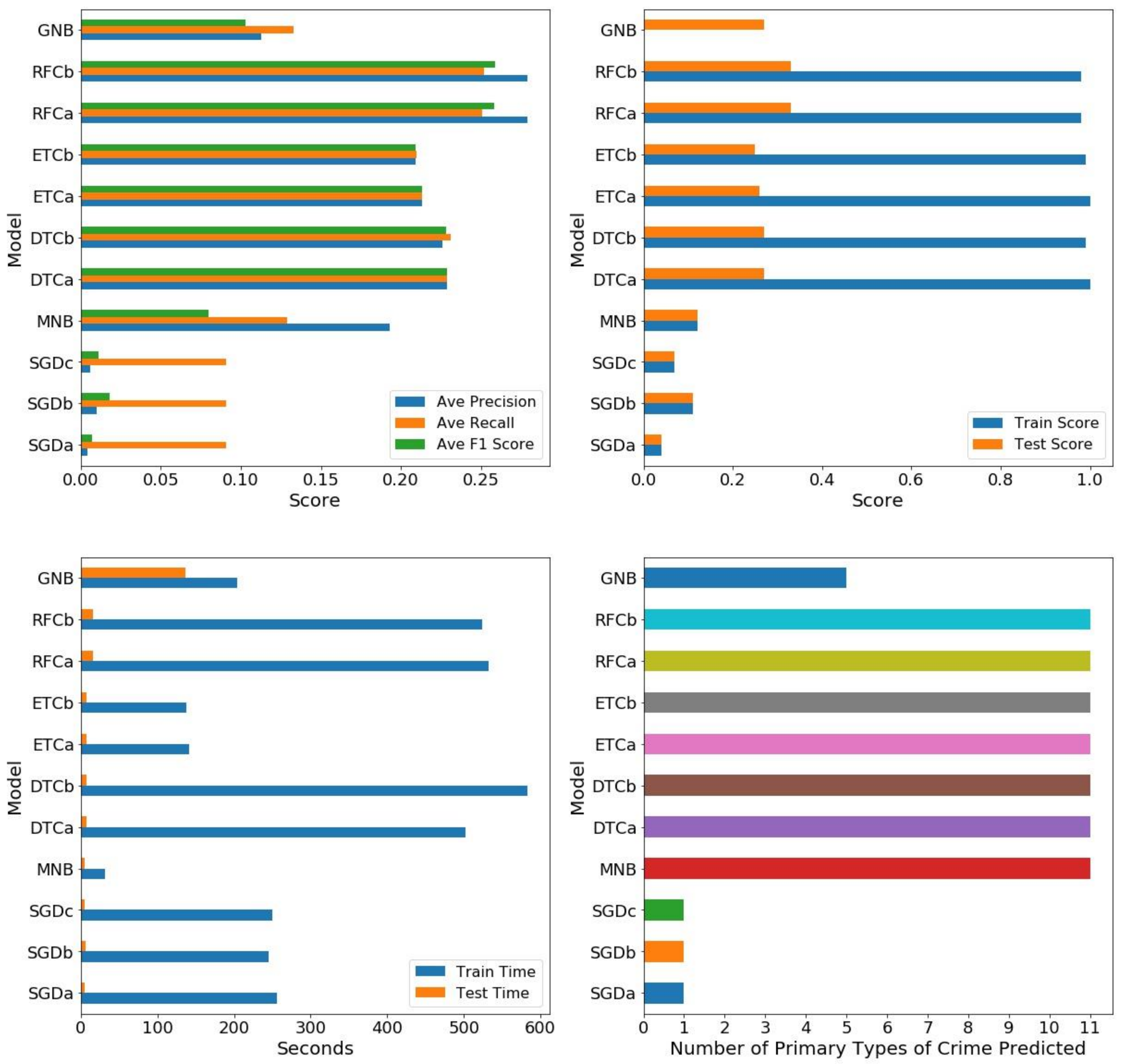
  - Scaled data using scikit-learn's MinMaxScaler
  - Scaled data with dimensionality reduced to 9 features



MODEL EVALUATION:
ORIGINAL DATA

Model Name	Adjusted Model Parameters	Shortened Model Name
SGD Classifier		SGDa
SGD Classifier	class_weight='balanced'	SGDb
SGD Classifier	class_weight='balanced', loss='log'	SGDc
Multinomial NB		MNB
Decision Tree Classifier		DTCa
Decision Tree Classifier	class_weight='balanced'	DTCb
Extra Tree Classifier		ETCa
Extra Tree Classifier	class_weight='balanced'	ETCb
Random Forest Classifier		RFCa
Random Forest Classifier	class_weight='balanced'	RFCb
Gaussian NB		GNB

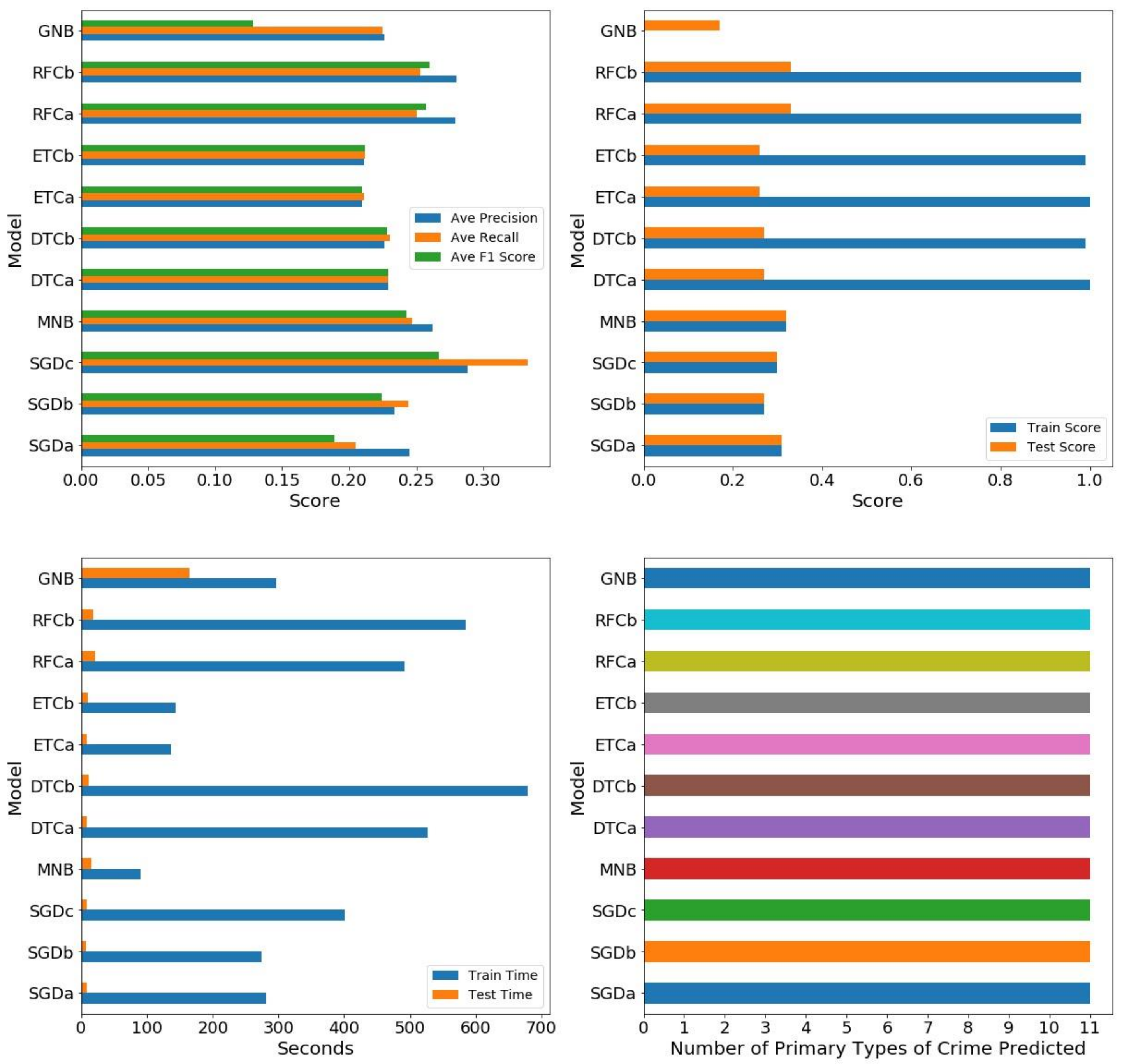
- SGD Classifiers and Gaussian NB did not predict all 11 types of crime
- Random Forest Classifiers had the best scores, but overfit the training data



MODEL EVALUATION:
SCALED DATA

Model Name	Adjusted Model Parameters	Shortened Model Name
SGD Classifier		SGDa
SGD Classifier	class_weight='balanced'	SGDb
SGD Classifier	class_weight='balanced', loss='log'	SGDc
Multinomial NB		MNB
Decision Tree Classifier		DTCa
Decision Tree Classifier	class_weight='balanced'	DTCb
Extra Tree Classifier		ETCa
Extra Tree Classifier	class_weight='balanced'	ETCb
Random Forest Classifier		RFCa
Random Forest Classifier	class_weight='balanced'	RFCb
Gaussian NB		GNB

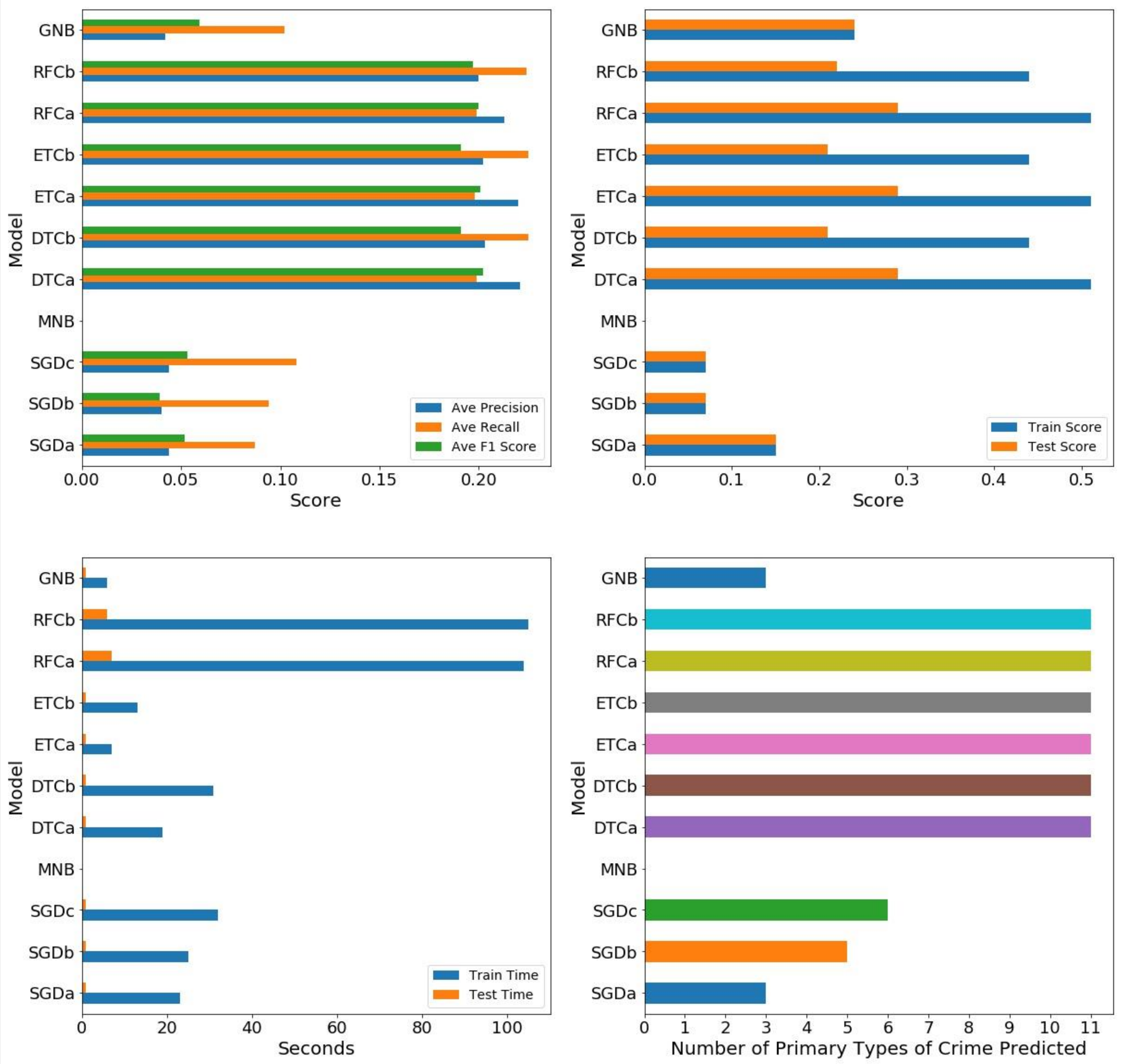
- All models predicted all 11 types of crime
- SGD Classifier with balanced class weight and log loss performed the best



MODEL EVALUATION:
DATA DIMENSIONALITY REDUCED TO 2 FEATURES

Model Name	Adjusted Model Parameters	Shortened Model Name
SGD Classifier		SGDa
SGD Classifier	class_weight='balanced'	SGDb
SGD Classifier	class_weight='balanced', loss='log'	SGDc
Multinomial NB		MNB
Decision Tree Classifier		DTCa
Decision Tree Classifier	class_weight='balanced'	DTCb
Extra Tree Classifier		ETCa
Extra Tree Classifier	class_weight='balanced'	ETCb
Random Forest Classifier		RFCa
Random Forest Classifier	class_weight='balanced'	RFCb
Gaussian NB		GNB

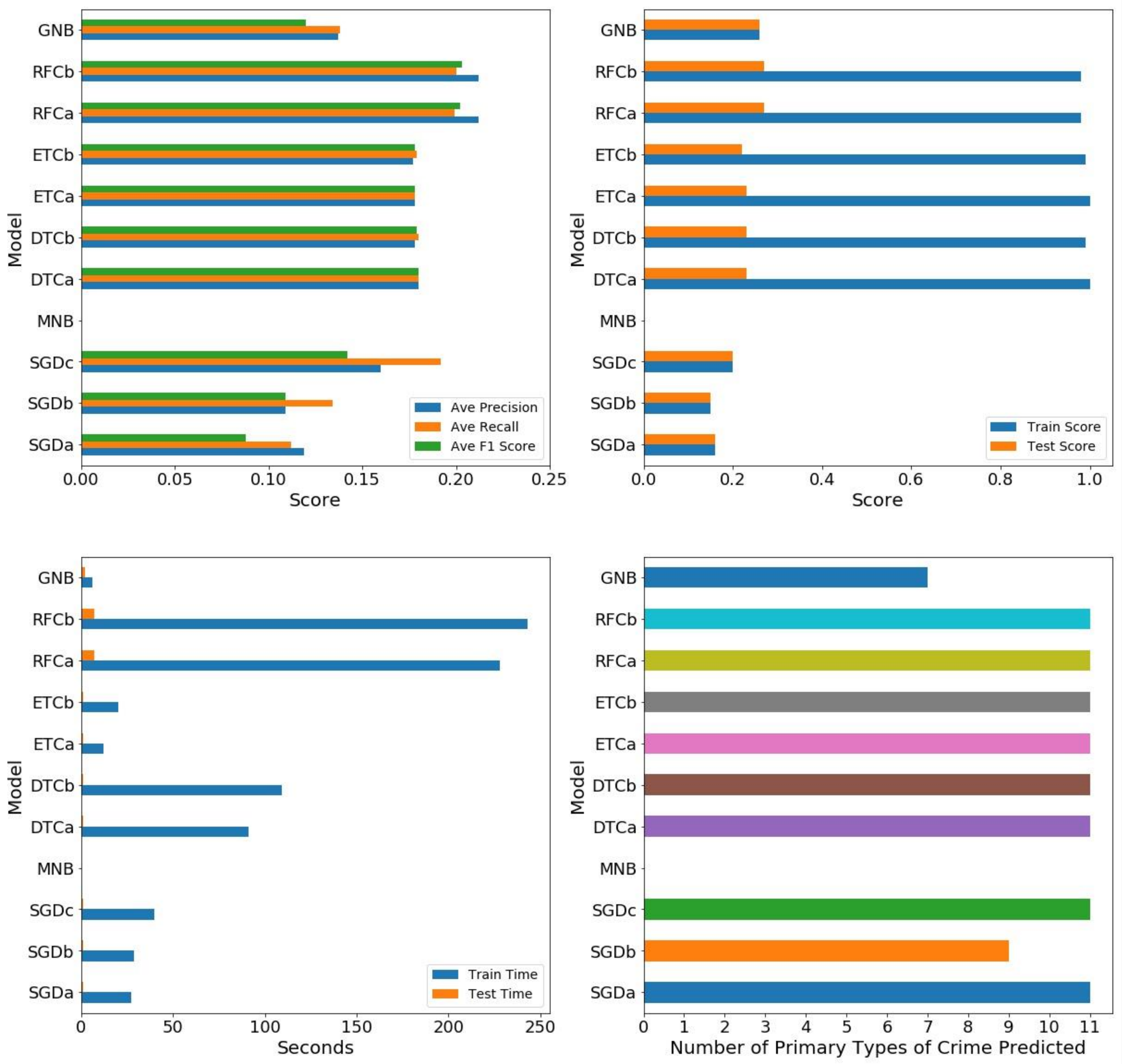
- Train time significantly reduced, but accuracy suffered
- SGD Classifiers and Gaussian NB did not predict all 11 types of crime
- Tree Classifiers performed the best and did not overfit the training data
  - Scores generally lower than for scaled data



MODEL EVALUATION:
SCALED DATA DIMENSIONALITY REDUCED TO 9 FEATURES

Model Name	Adjusted Model Parameters	Shortened Model Name
SGD Classifier		SGDa
SGD Classifier	class_weight='balanced'	SGDb
SGD Classifier	class_weight='balanced', loss='log'	SGDc
Multinomial NB		MNB
Decision Tree Classifier		DTCa
Decision Tree Classifier	class_weight='balanced'	DTCb
Extra Tree Classifier		ETCa
Extra Tree Classifier	class_weight='balanced'	ETCb
Random Forest Classifier		RFCa
Random Forest Classifier	class_weight='balanced'	RFCb
Gaussian NB		GNB

- Train time reduced, but accuracy suffered
- SGD Classifier with balanced class weight and Gaussian NB did not predict all 11 types of crime
- Random Tree Classifiers performed the best but overfit the training data

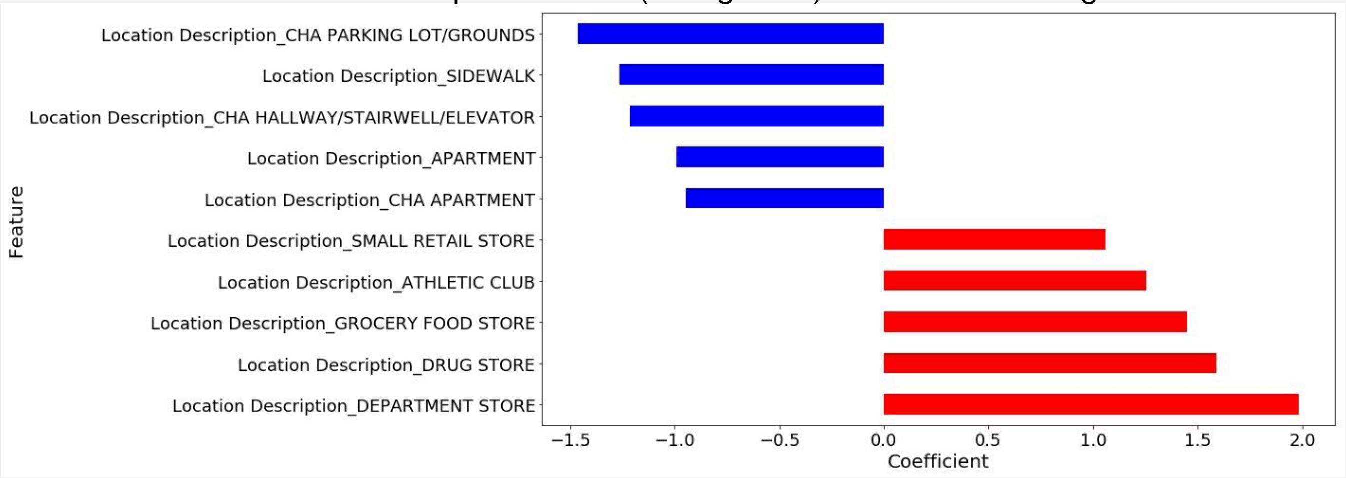


FEATURES OF IMPORTANCE

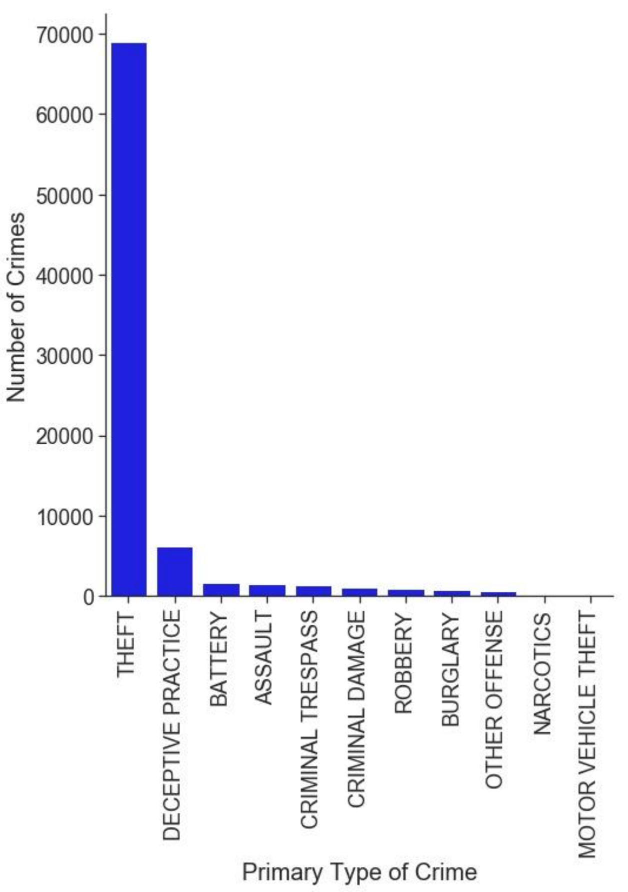
- Used scaled training data on the SGD Classifier with balanced class weight and log loss
- Extracted the features associated with the top 10 coefficients (in magnitude)
- Explored the feature with the largest, positive coefficient for the top 4 most frequent crimes:
  - Theft
  - Battery
  - Criminal damage
  - Narcotics

THEFT

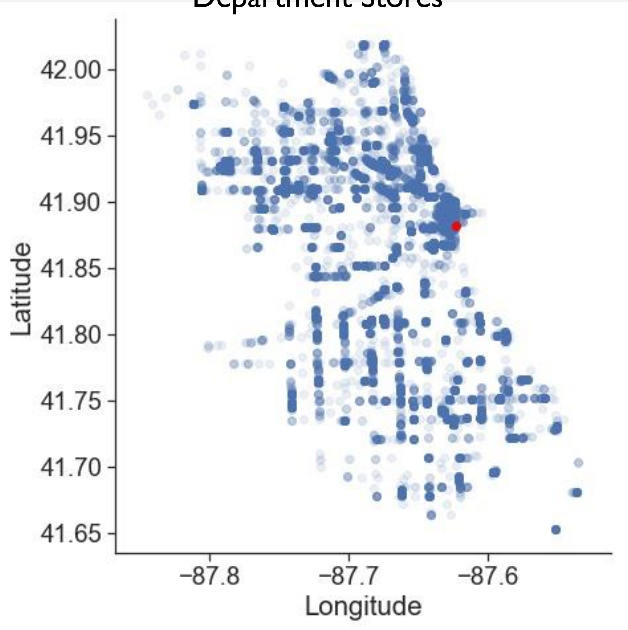
Coefficients of the Top 10 Features (in Magnitude) for Crimes Involving Theft



Breakdown of Crimes Reported in Department Stores

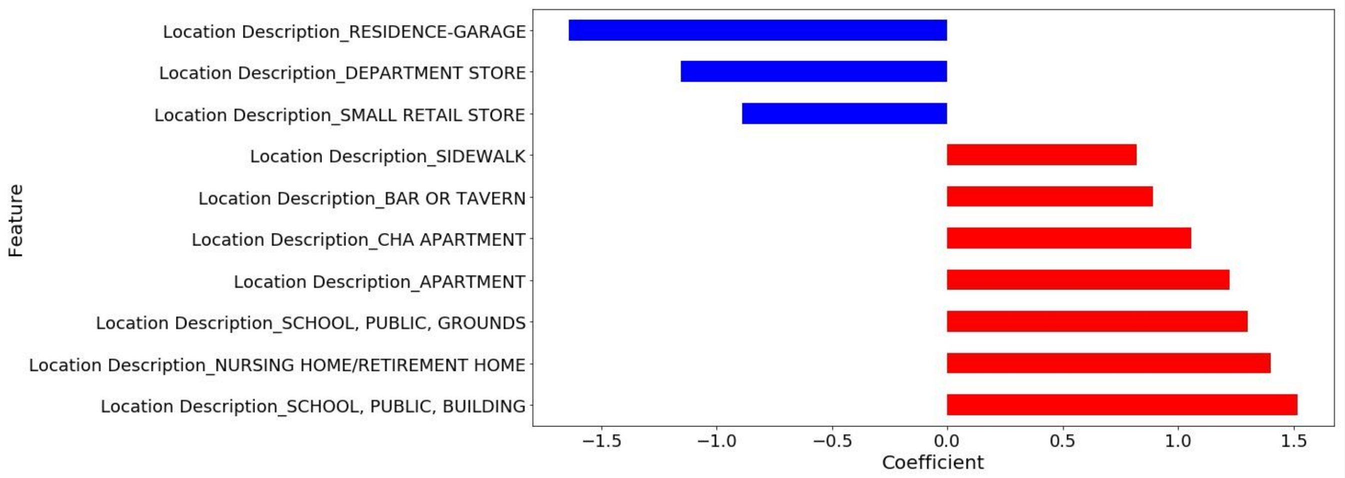


Spatial Distribution of Crimes Involving Theft in Department Stores

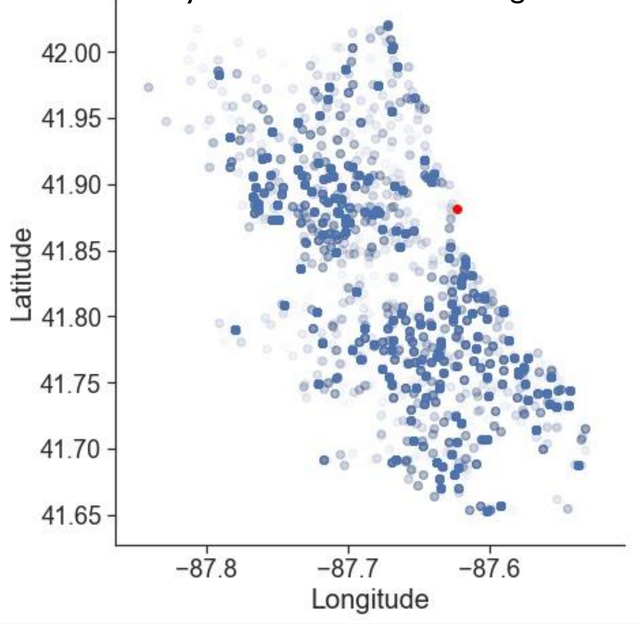


BATTERY

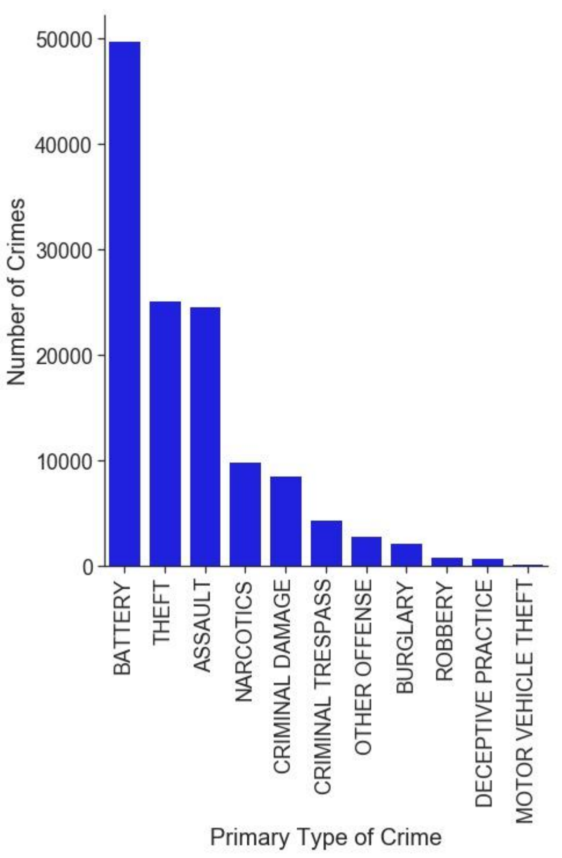
Coefficients of the Top 10 Features (in Magnitude) for Crimes Involving Battery



Spatial Distribution of Crimes Involving Battery in Public School Buildings

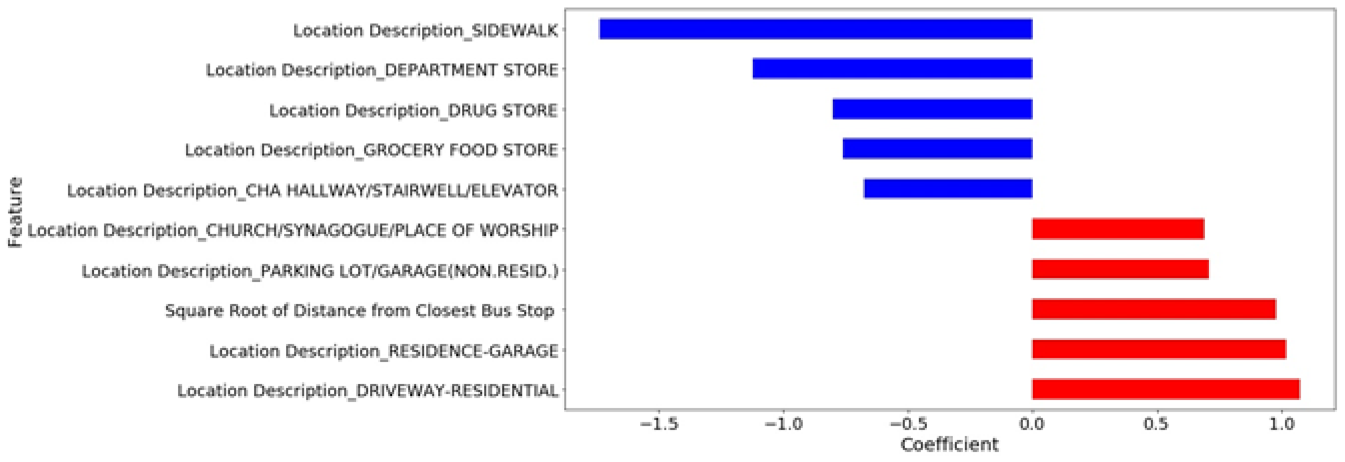


Breakdown of Crimes Reported in Public School Buildings

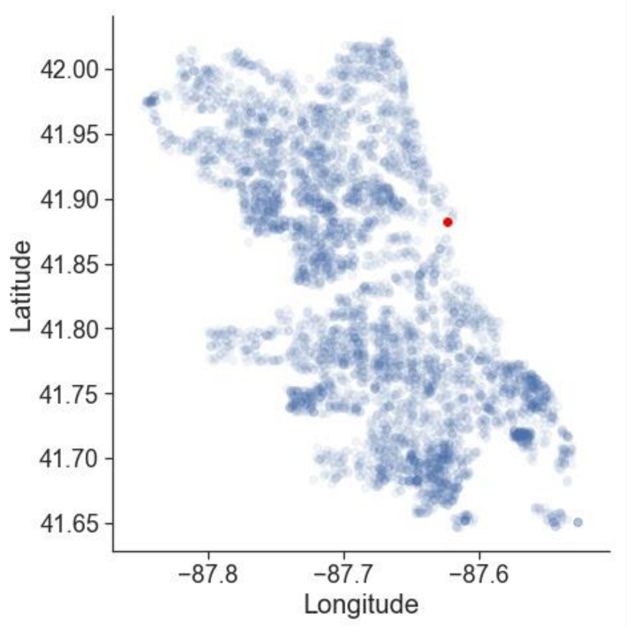


CRIMINAL DAMAGE

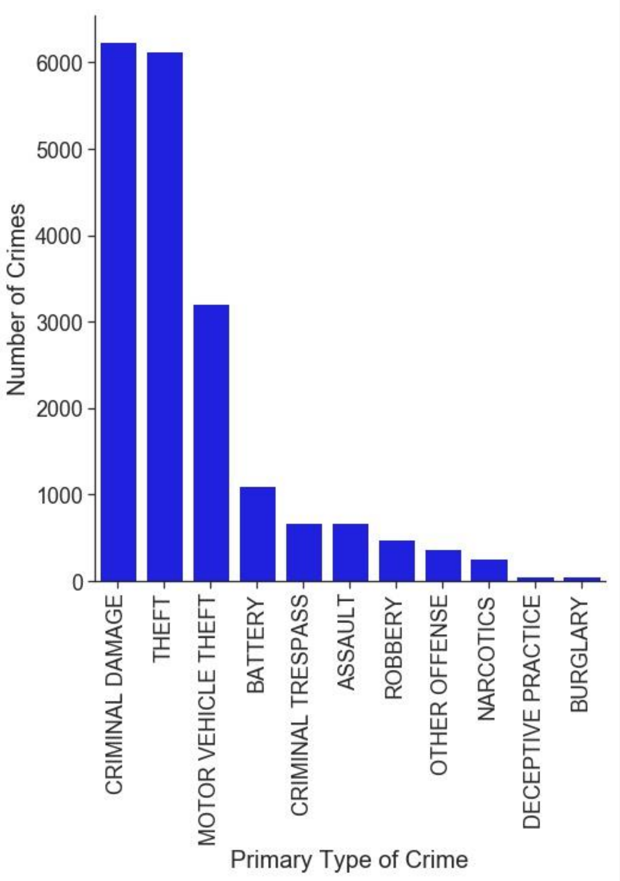
Coefficients of the Top 10 Features (in Magnitude) for Crimes Involving Criminal Damage



Spatial Distribution of Crimes Involving Criminal Damage on Residential Driveways

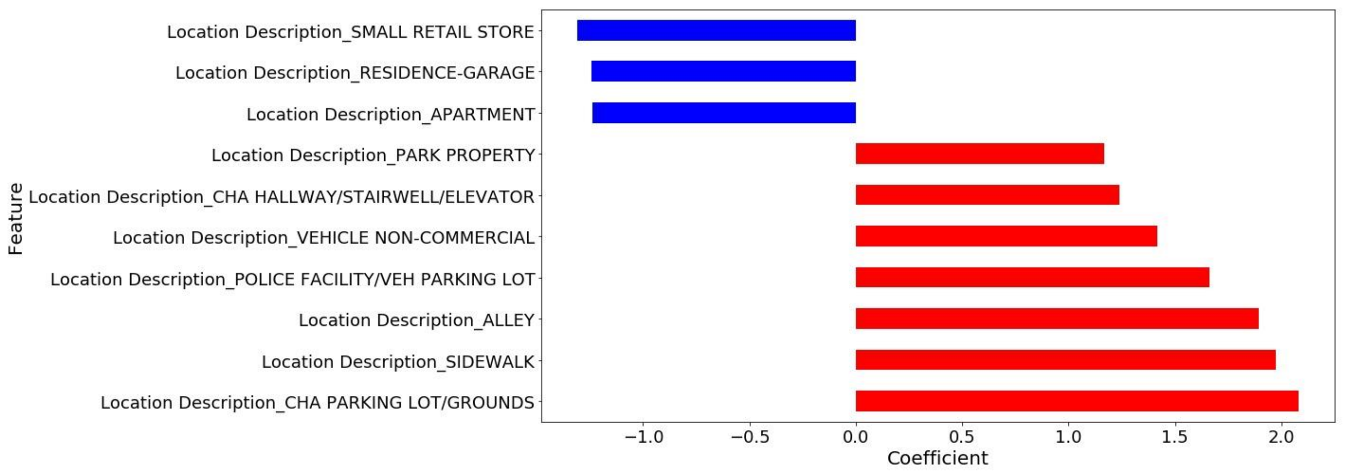


Breakdown of Crimes Reported on Residential Driveways

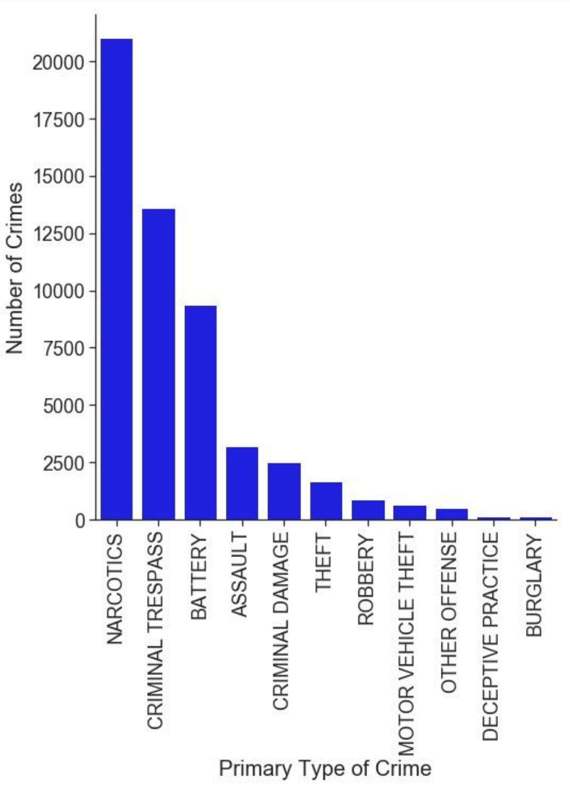


NARCOTICS

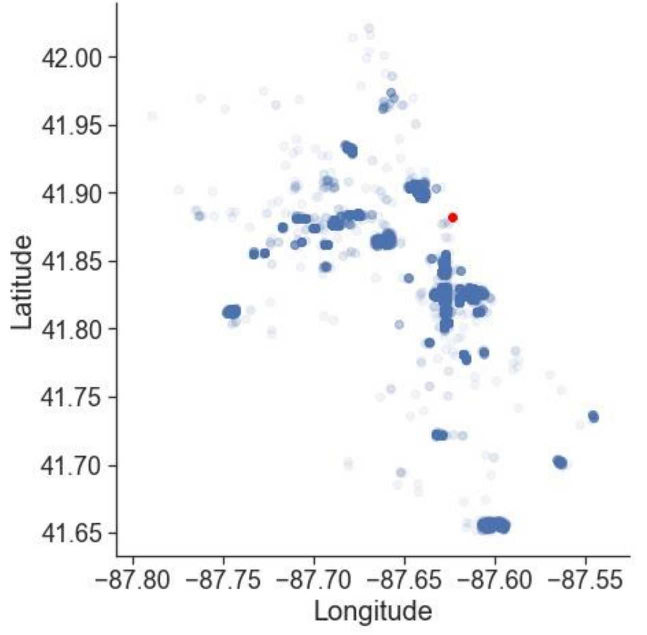
Coefficients of the Top 10 Features (in Magnitude) for Crimes Involving Narcotics



Breakdown of Crimes Reported on CHA Grounds



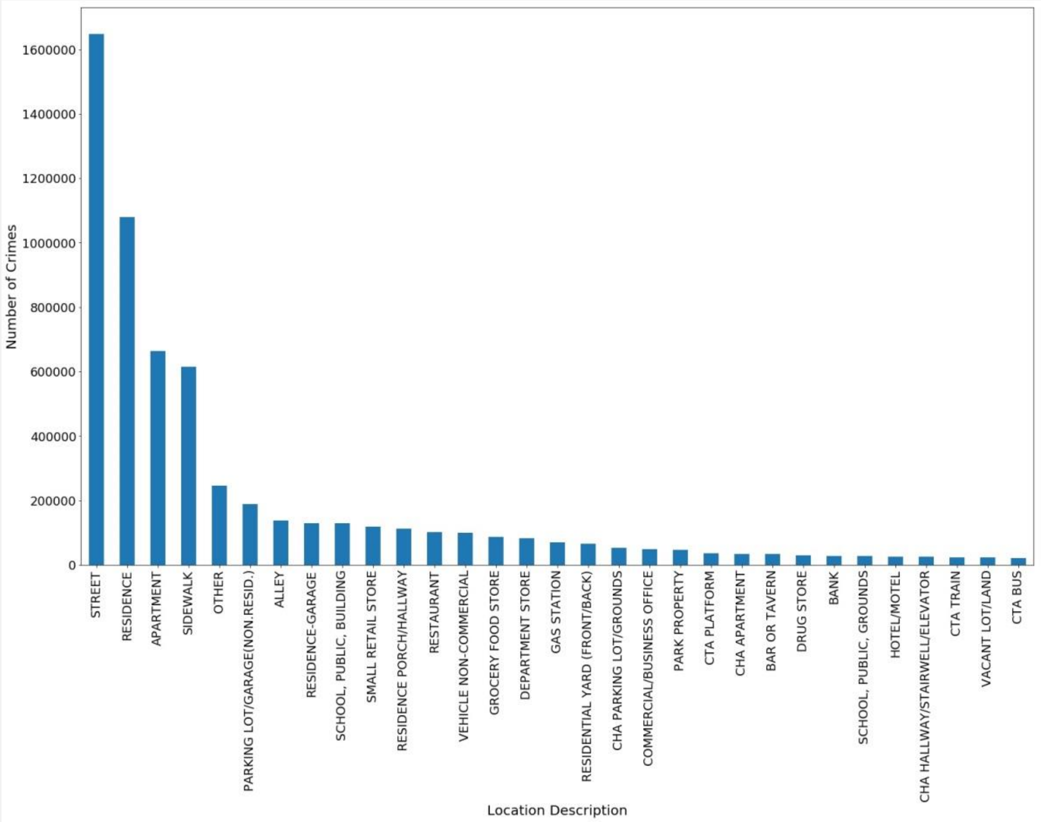
Spatial Distribution of Crimes Involving Narcotics on CHA Grounds



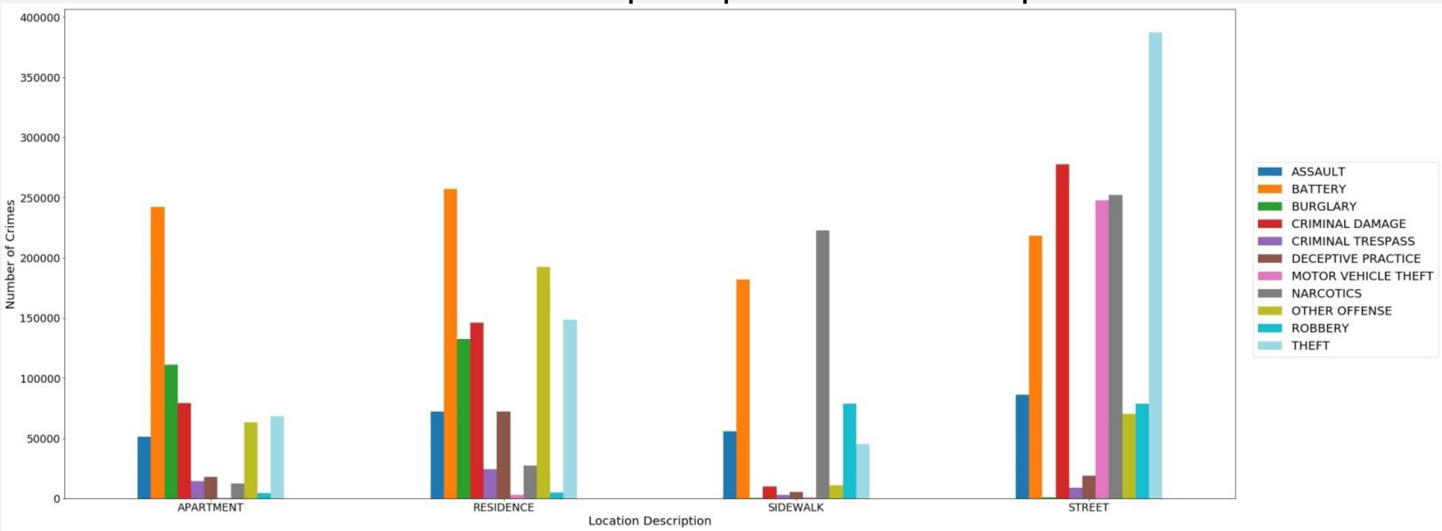
LOCATION DESCRIPTION

- Location description the most important indicator for the reported crime type

Number of Crimes per Location Description

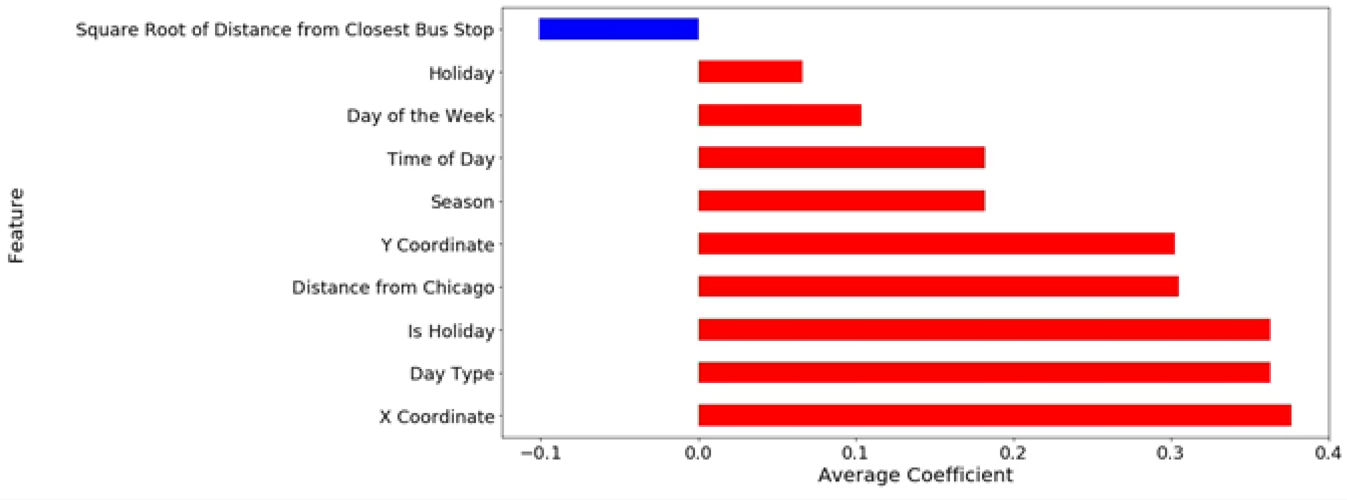


Number of Crimes per Top 4 Location Descriptions



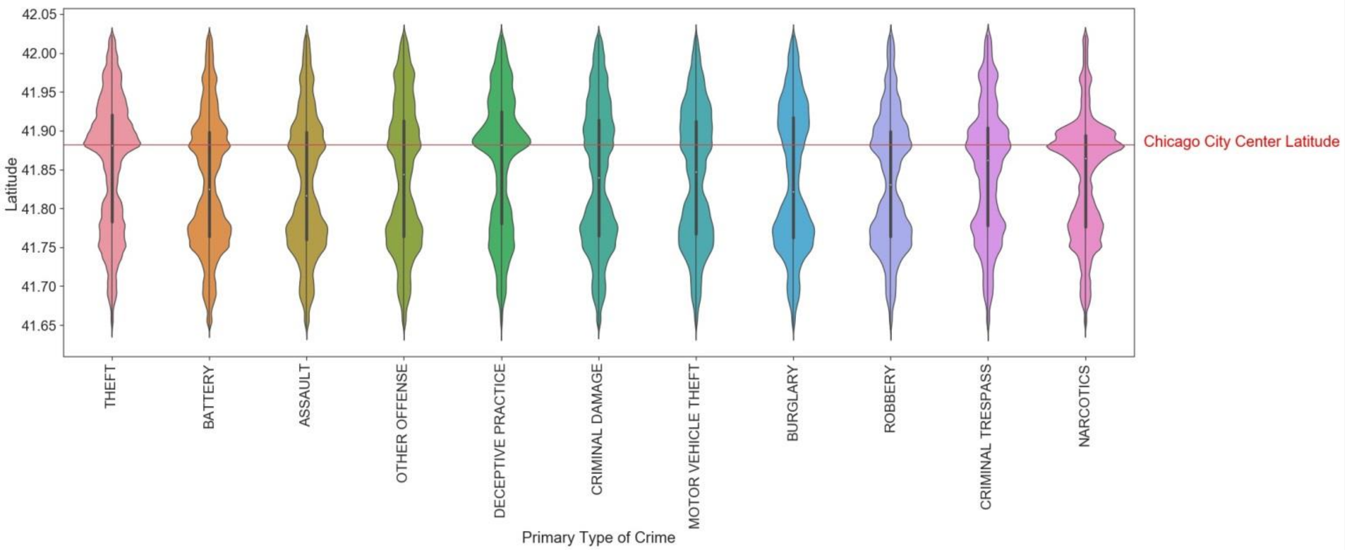
FEATURES OF IMPORTANCE

- Averaged the coefficients for each feature across all 11 primary types of crimes and then averaged the coefficients for all of the dummy variables of a feature
- Resulting top 10 features (in magnitude):

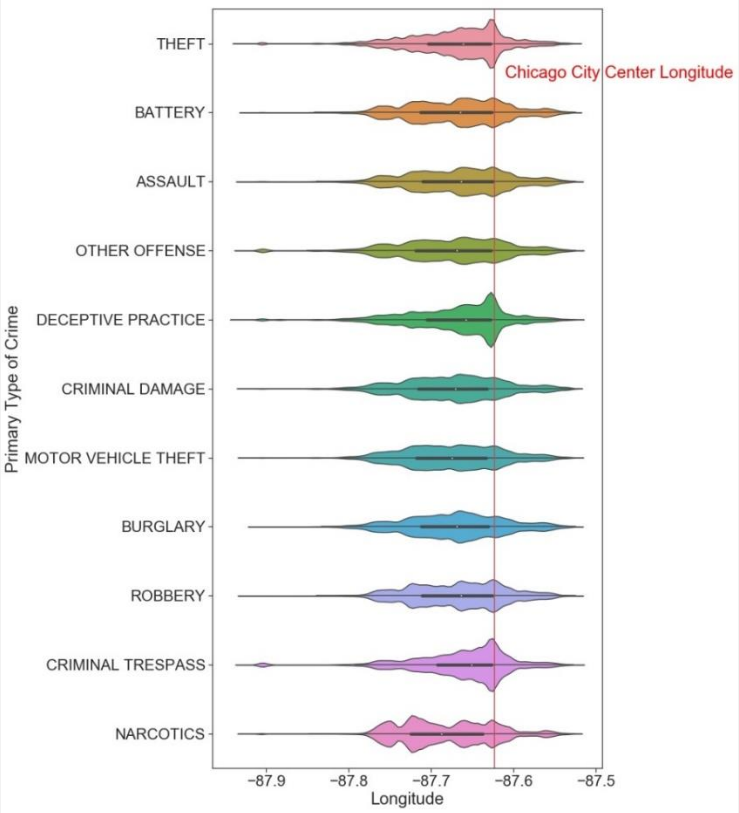


LATITUDE AND LONGITUDE

Distribution of Latitudes for each Primary Type of Crime



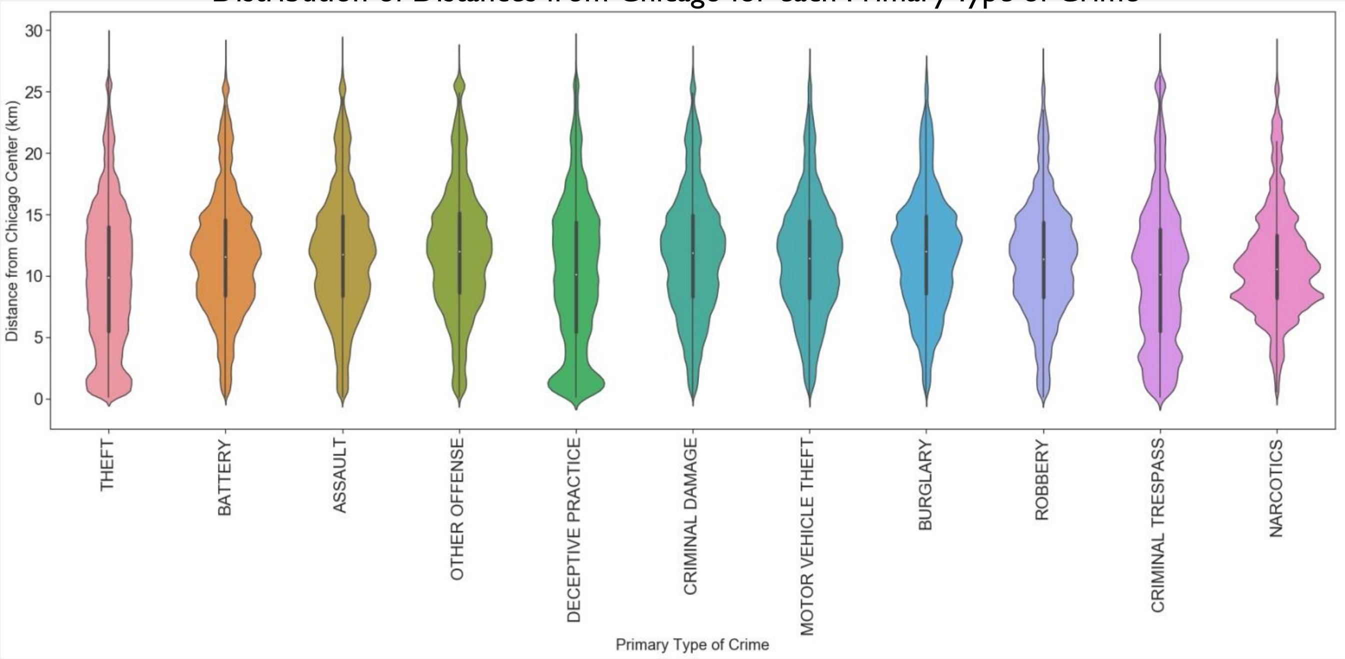
Distribution of Longitudes for each Primary Type of Crime



- Theft, deceptive practice, criminal trespassing, and narcotics have higher number of crimes towards the northern part of Chicago
- Pronounced increase in the number of crimes:
  - For theft, deceptive practice, and narcotics near the latitude of the city center
  - For theft, deceptive practice, and criminal trespassing near the longitude of the city center

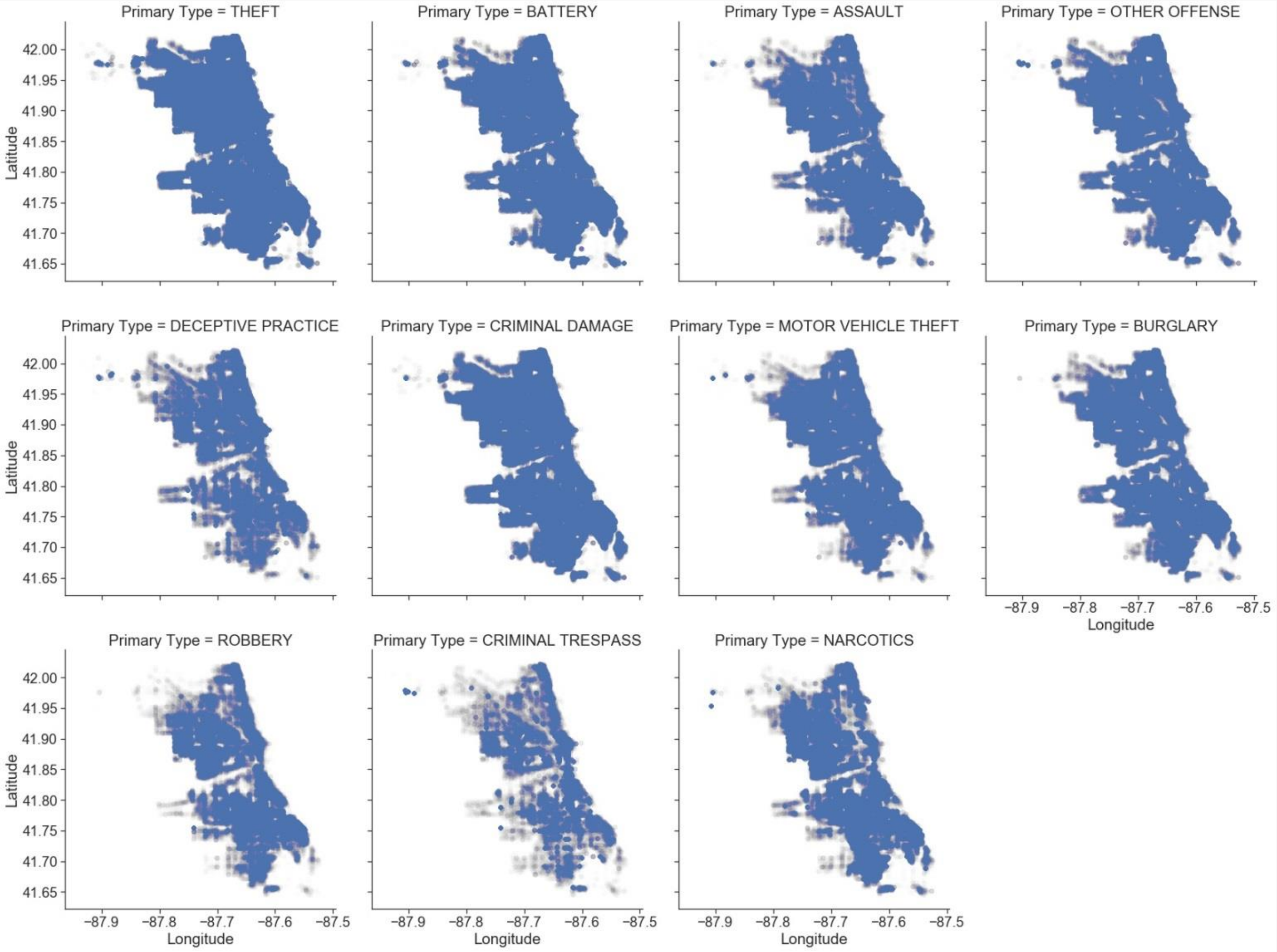
DISTANCE FROM CHICAGO CITY CENTER

Distribution of Distances from Chicago for each Primary Type of Crime



- Increase in the number of crimes involving theft, deceptive practice, and criminal trespassing closer to the city center
- Increase in the number of crimes involving narcotics along the same latitude as city center but farther west.

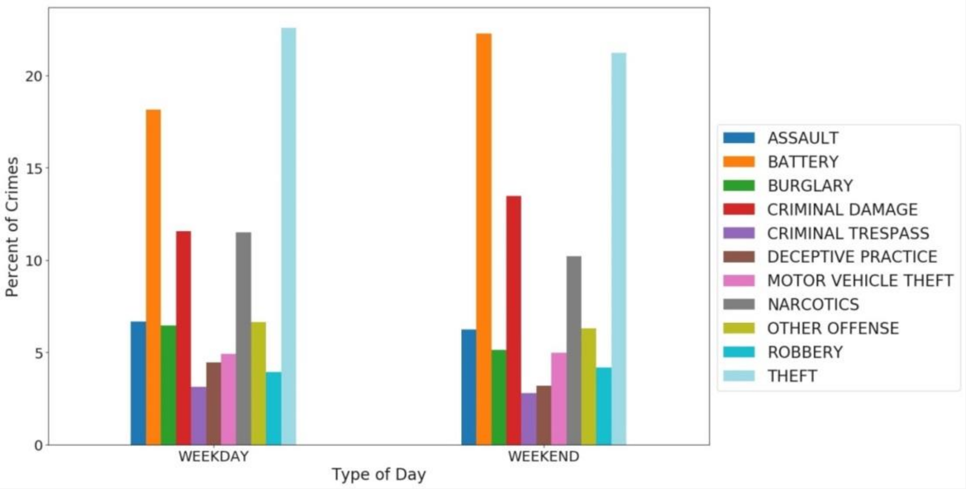
SPATIAL DISTRIBUTION OF CRIMES



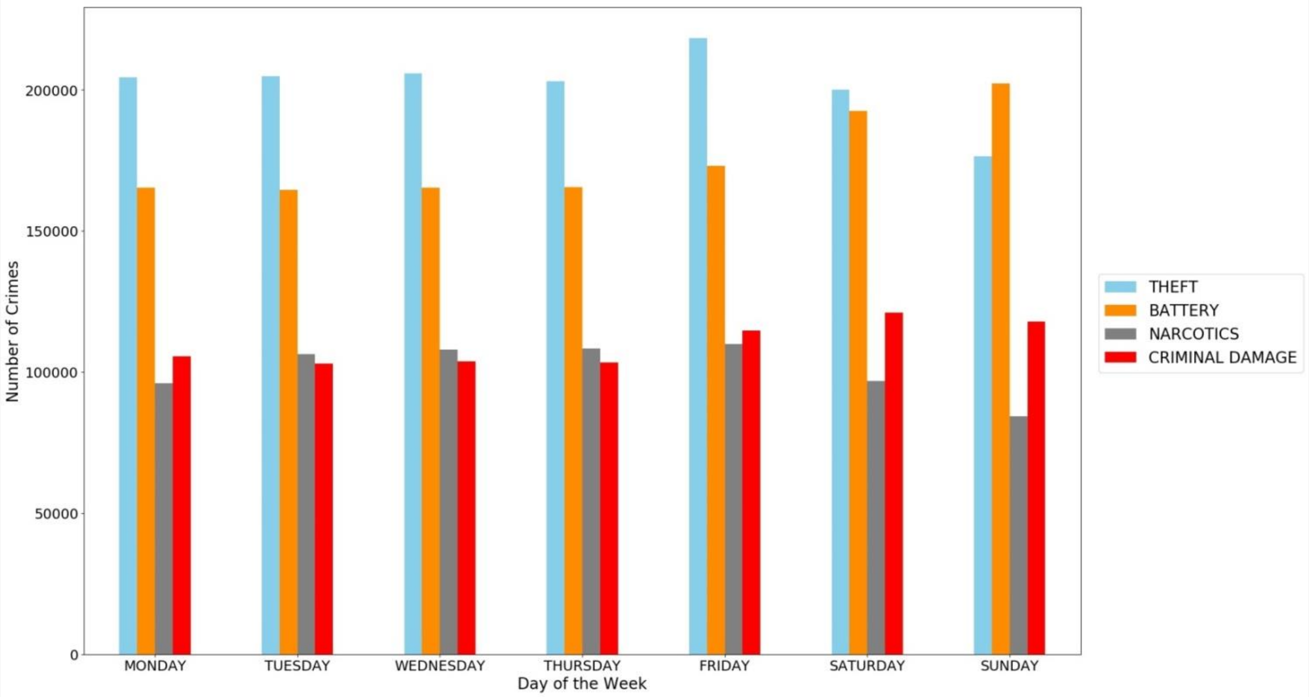
- High concentration of crimes along Lake Michigan (eastern edge of the map)
- Crimes involving narcotics:
  - Some gaps along the lake
  - Larger area of reduced concentration farther inland on the north side
- Crimes involving theft, battery, criminal damage, and other offenses have higher concentrations of crimes in the northwest area of Chicago

DAY OF THE WEEK

Percentage of Each Primary Type of Crime per Type of Day



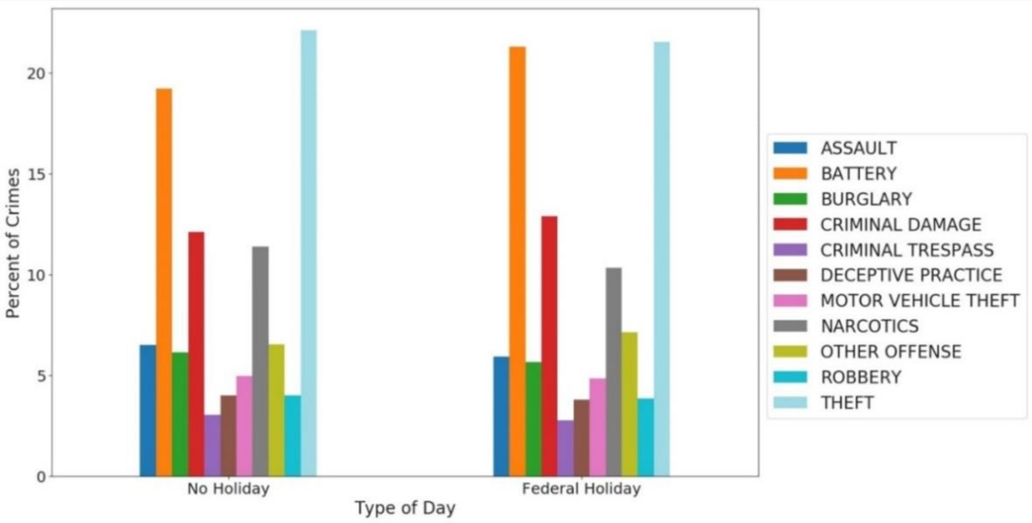
Number of Top 4 Primary Types of Crime per Day of the Week



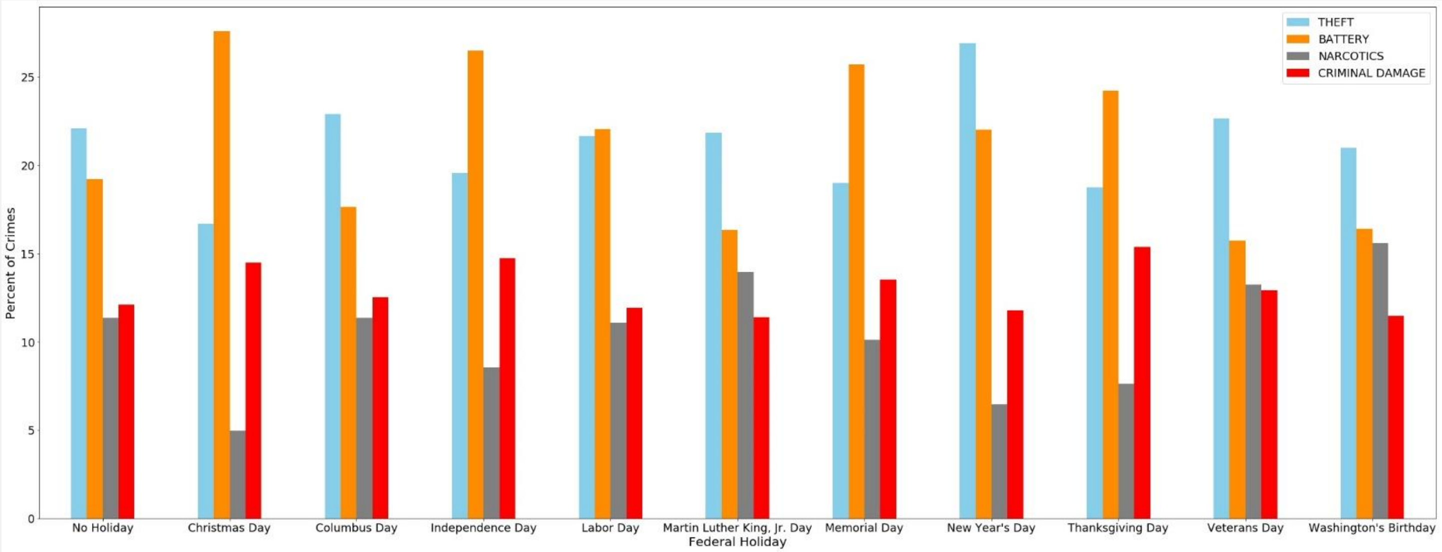
- Chance of crimes involving battery and criminal damage increases over the weekend
- Chance of crimes involving theft an narcotics decreases slightly over the weekend
- Number of crimes involving battery peaks on Sunday
- Number of crimes involving theft and narcotics reaches a minimum on Sunday

FEDERAL HOLIDAYS

Percentage of Each Primary Type of Crime per Type of Day



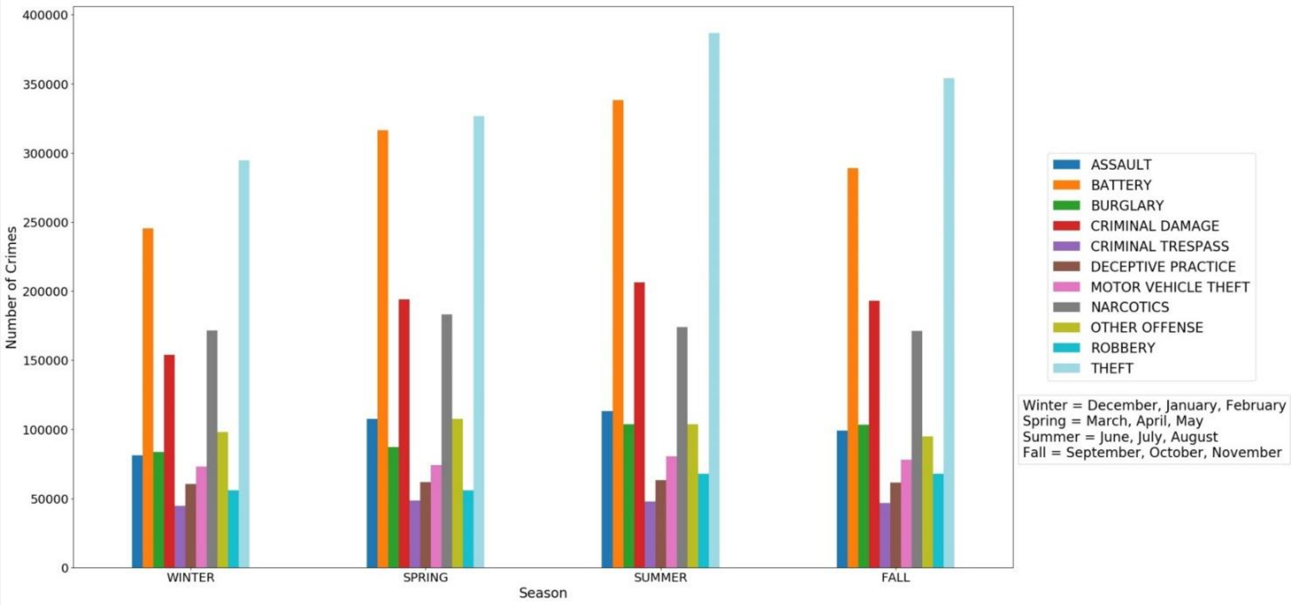
Percentage of Top 4 Primary Types of Crime per Federal Holiday



- On federal holidays, greater chance for crimes involving battery and slightly lower chance for crimes involving narcotics
- Higher proportions of crimes involving battery on Christmas, Independence Day, Labor Day, Memorial Day, New Year's, and Thanksgiving
- Higher proportion of crimes involving theft on New Year's
- Lower proportions of crimes involving narcotics on Christmas, Independence Day, New Year's, and Thanksgiving

SEASON

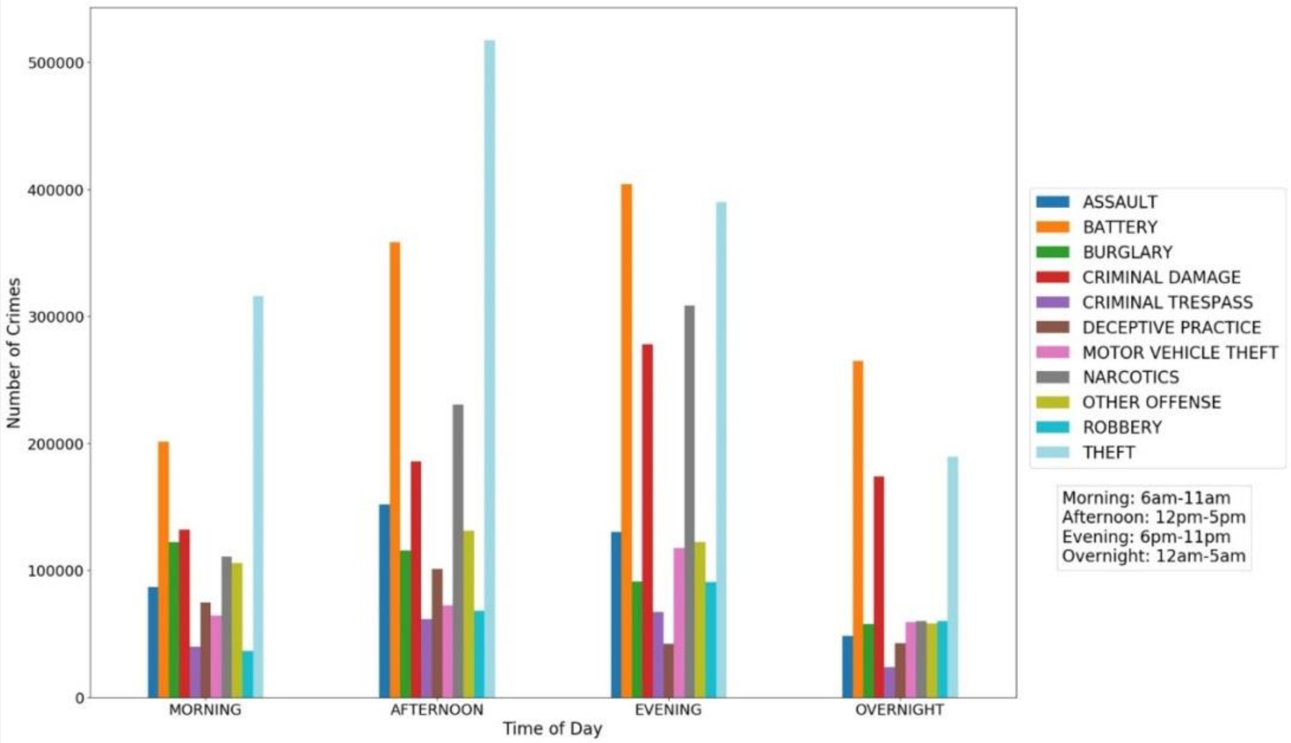
Number of Each Primary Type of Crime per Season



- Number of crimes involving theft peaks in summer
- Crimes involving battery more frequent during spring and summer
- Crimes involving criminal damage more frequent during spring summer and fall

TIME OF DAY

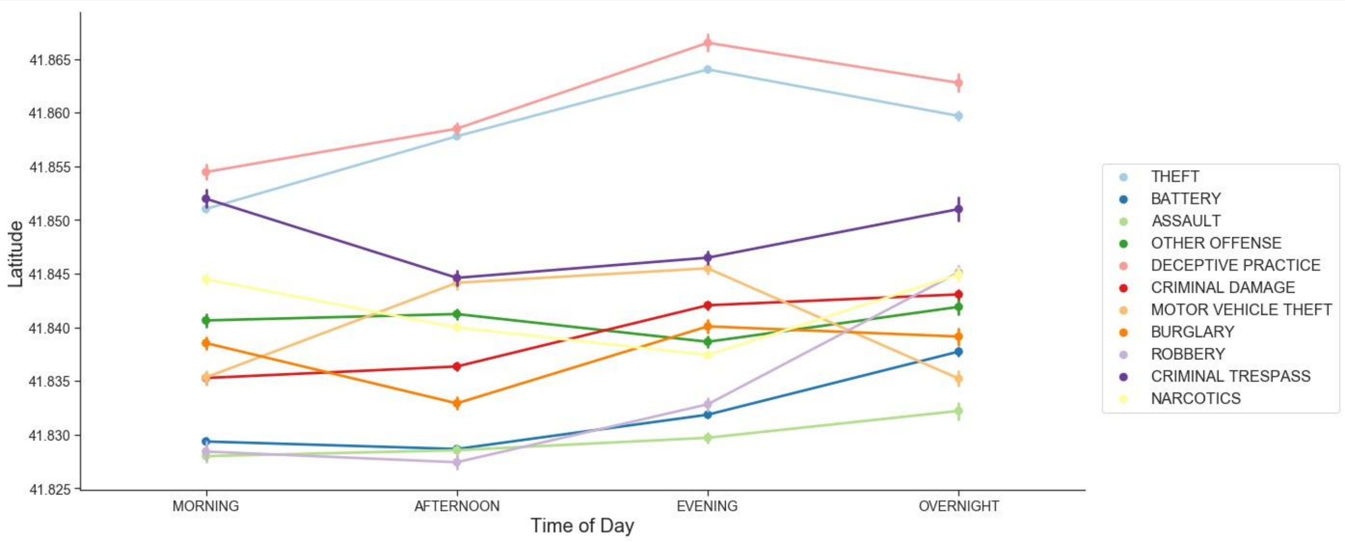
Number of Each Primary Type of Crime per Time of Day



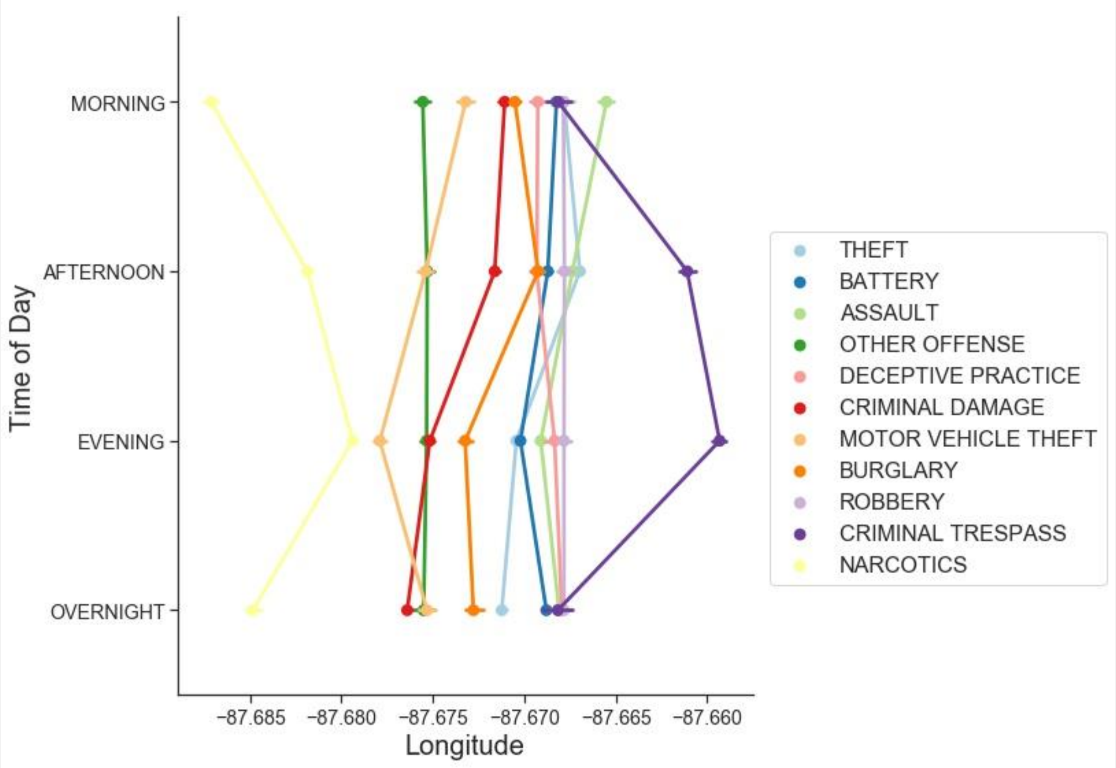
- Number of crimes involving theft peaks during the afternoon
- Number of crimes involving battery peaks in the evening and overnight, it is the highest of all crimes
- Crimes involving criminal damage and narcotics peak in the evening

TIME OF DAY AND LATITUDE/LONGITUDE

Average Latitude per Time of Day for each Primary Type of Crime



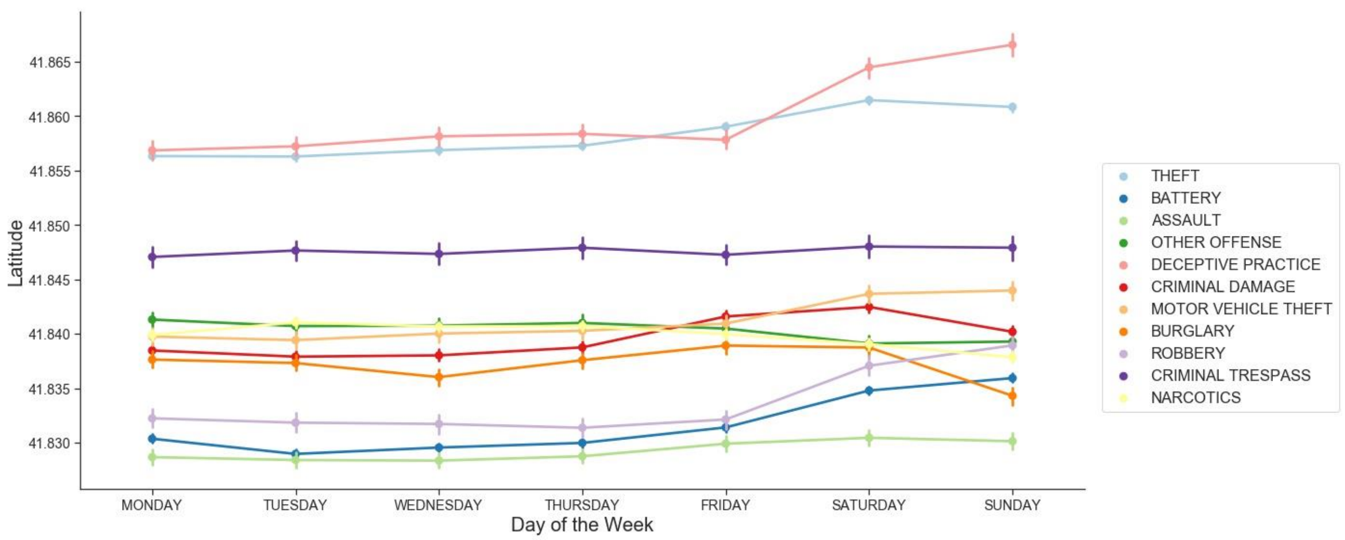
Average Longitude per Time of Day for each Primary Type of Crime



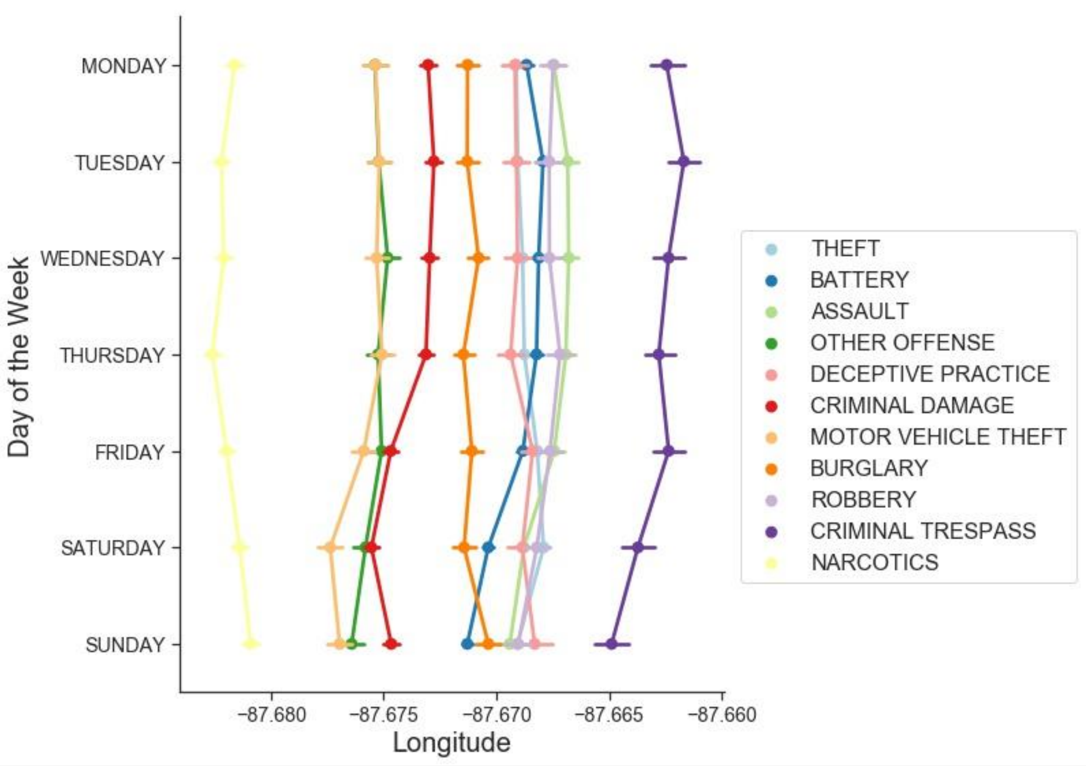
- Apparent shift north for deceptive practice, theft, criminal damage, burglary, motor vehicle theft, and robbery
- Shift south for criminal trespassing
- Shift east for narcotics and criminal trespassing

DAY OF THE WEEK AND LATITUDE/LONGITUDE

Average Latitude per Day of the Week for each Primary Type of Crime



Average Longitude per Day of the Week for each Primary Type of Crime



- Apparent shift north for deceptive practice, robbery, and battery during the weekend
- Shift south for burglary
- Subtle shift west for motor vehicle theft, criminal damage, battery, and criminal trespassing into the weekend

RECOMMENDATIONS

- Increase police officer/security guard presence in public schools to help prevent the occurrence of battery.
  - Start with schools within community 25 (western part of north side of Chicago) to see if it makes a difference.
- In areas surrounding the Chicago city center, especially to the south, police officers should be on the lookout for individuals using/dealing narcotics.
  - More officers should patrol these areas in an effort to thwart narcotic related crimes.
- Increase police officer presence on the eastern edge of the city along Lake Michigan.
  - Even criminal trespassing, which occurred the least frequently of the 11 studied primary types of crime had high concentrations of crime in this area.

RECOMMENDATIONS

- Police officers need to be prepared to deal with and be on the lookout for more instances of battery on weekends and holidays.
- Increase the number of police officers patrolling near the city center to thwart theft within department stores.
  - More police presence may also reduce deceptive practice and criminal trespassing near the city center.
- Police should help residents set up neighborhood watches.
  - Crimes involving criminal damage on residential driveways are quite widespread across Chicago and there are no significant hot spots for police officers to concentrate on.

FUTURE WORK

- In order to more accurately obtain the missing latitudes/longitudes, the blocks could be geocoded.
- The ward boundaries change after every federal census (Knox, 2005). The latitude/longitude could be used to redo all of the reported wards so that only one ward boundary system is used.
- Using all available crime reports for the top 11 most frequent types of crime, cross validation could be performed on the examined models to see if the SGD Classifier is indeed the most optimal model.
  - After choosing the best model, it could be tweaked to improve accuracy by:
    - adjusting the model parameters
    - removing unneeded features
    - adding new features
- The location description was the most important indicator for the reported crime type when looking at feature coefficients individually. There were approximately 100 unique location descriptions in my dataset. Natural language processing could possibly be used to group similar locations together.
  - The data could then be used to train/test the SGD Classifier to see if there is any improvement in the accuracy or if when looking at the most influential coefficients, any features other than location description show up.

FUTURE WORK

- The top 4 location descriptions (street, residence, apartment, and sidewalk) could be analyzed to see if the number of certain crimes for each location changes spatially.
  - e.g. looking at how the concentration of crimes involving theft on streets changes across Chicago
- For crimes involving narcotics, there was a large area with a lower concentration of crime slightly inland from the lake on the north side of Chicago. A study should be done to see what it is about this area that makes it less prone to crimes involving narcotics; or if for some reason, they occur here but aren't reported.
  - Demographics, income, business density, and officer density are a few features that could be examined.
- As some variation in the frequency of certain crimes (especially theft and battery) occurred with season, the weather conditions preceding and during the crimes could be examined to see if it would be a good predictor for the type of crime.
- As several of the crime types varied throughout the day, foot traffic data could be studied to see if there really is a significant relationship between the number of people outside and the occurrence of certain crimes (e.g. battery).
- Variation in average latitude/longitude throughout the day and week was seen for several types of crime. In order to see where the biggest shifts in crime are occurring, the spatial distribution of each crime for each time of day and day of the week could be examined.