Chicago Crime Classification

Intro

Chicago has one of the largest homicide rates in the country when it comes to major U.S. cities. Even though the crime rate in recent years has been at historic lows, Chicago was responsible for nearly half the homicides in 2016 according the Chicago police department Bureau of records. While it is unclear why the numbers remain so high, the police department has been tracking crime statistics for several years to better understand insights into the pattern of crimes in Chicago. I have developed binary classification models to predict if the crime was serious or not. In order to increase my reliability, I have compared the accuracy rate among various different classifiers.

1. Data

Kaggle is the world's largest data science community with powerful tools and resources to help achieve your data science goals. It simply a gathering place for data scientists and machine learning practitioners. The dataset reflects incidents of crime that occurred in the city of Chicago from 2001 to present. The data is extracted from the Chicago police Department. Some of the variables include arrest made, longitude, latitude, crime type, police district etc. Other I

used included police district location and socio-economic data to get a better idea of the status of each neighborhood.

2. Methods

Binary Classification: This is the idea of classifying the elements of a given set into two groups. In this case, I classified crimes to be whether they were “serious” or “not serious”.

Multiclass Classification: This is the idea of classifying instances into one of three or more classes. In this case, I differentiated the top 4 crimes that happened in Chicago.

Here are the following methods I used for classification:

1. Logistic Regression: used to find the probability of an event success and event failure

   Advantages: It is easier to implement, interpret, and very efficient to train.

   Disadvantages: If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.

2. Support Vector Machines: supervised learning models with associated learning algorithms that analyze data used for classification

   Advantages: works relatively well when there is a clear margin of separation between classes

   Disadvantages: It is not suitable for large datasets

3. Decision Tree: It used to go from observations about an item to conclusions about the item’s target value

   Advantages: Requires less effort for data preparation during pre-processing

   Disadvantages: A small change in the data can cause a large change in the structure of the decision tree causing instability.

4. Naïve Bayes: based on applying Bayes theorem with strong independence assumptions between the features

   Advantages: Affords fast, highly scalable model building and scoring

Disadvantages: It assumes the attributes are independent and treats all attributes equally

5. Random Forest: A classification algorithm that consists of many different decision trees

   Advantages: handles very large datasets with higher dimensionality.

   Disadvantages: Large number of trees can make the algorithm too slow and ineffective
6. KNN: a simple algorithm that stores all available cases and classifies new cases based on a similarity measure

   Advantages: No training period and data can be added seamlessly

   Disadvantages: Does not work well with large dataset and high dimensions

I will dive into deeper later in the report on which machine learning algorithm had the best training and testing accuracy.

3. Data Wrangling

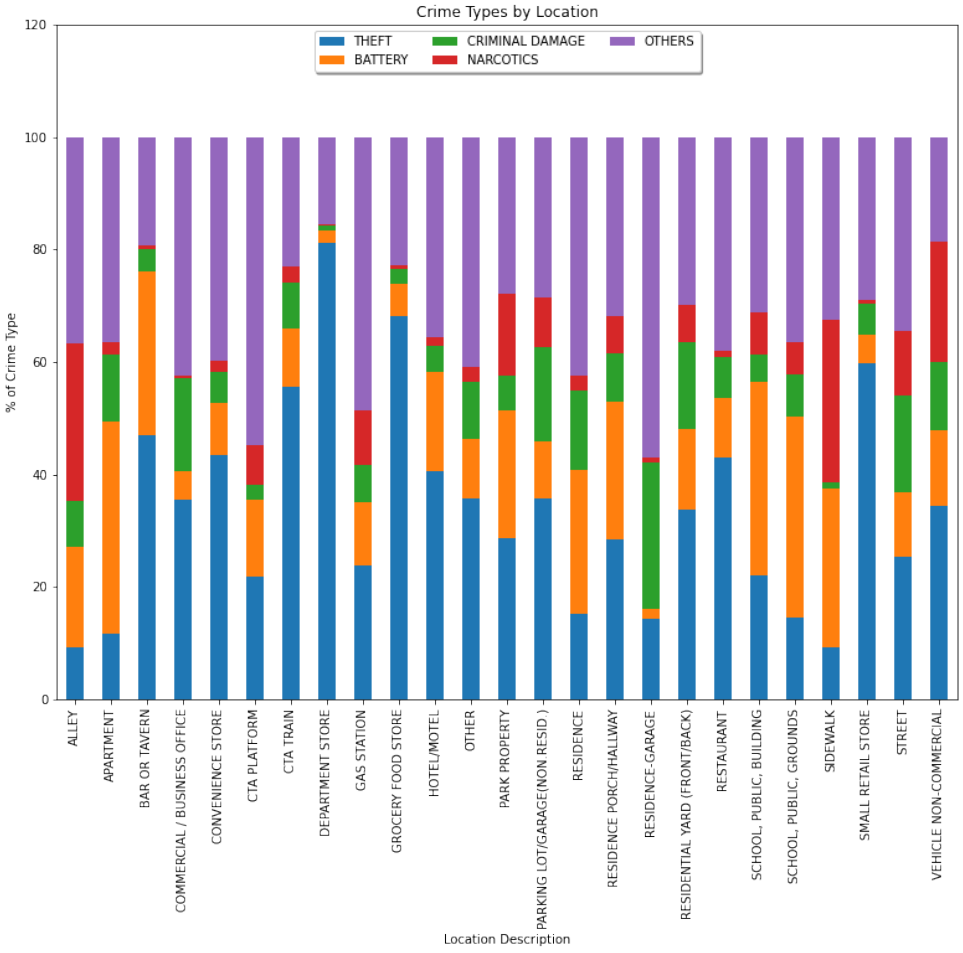
First thing I did was filter the data from 2010 – 2019. I only analyze the data from the past decade. I then grouped the primary type of crimes where we found out the most common crimes were theft, battery, criminal damage, and narcotics. These top 4 were later used in our multi-classification problem. I then grouped crimes that I considered serious and not serious. While this subjective, I grouped the serious crimes as arson, assault, battery, sexual assault, criminal damage, criminal trespass, homicide, and robbery. I then grouped by the most common locations a crime happened. The most common ones were streets, residences, apartments, and sidewalks. I then added columns for what police district they happened in and what time of the day these crimes happened. I then created a formula for to add a column of where the closest police station was when the crime occurred. I also added the socio-economic data to see the hardship index for each Chicago data as well as various other important variables. This data was then considered cleaned and imported for Exploratory Data Analysis.

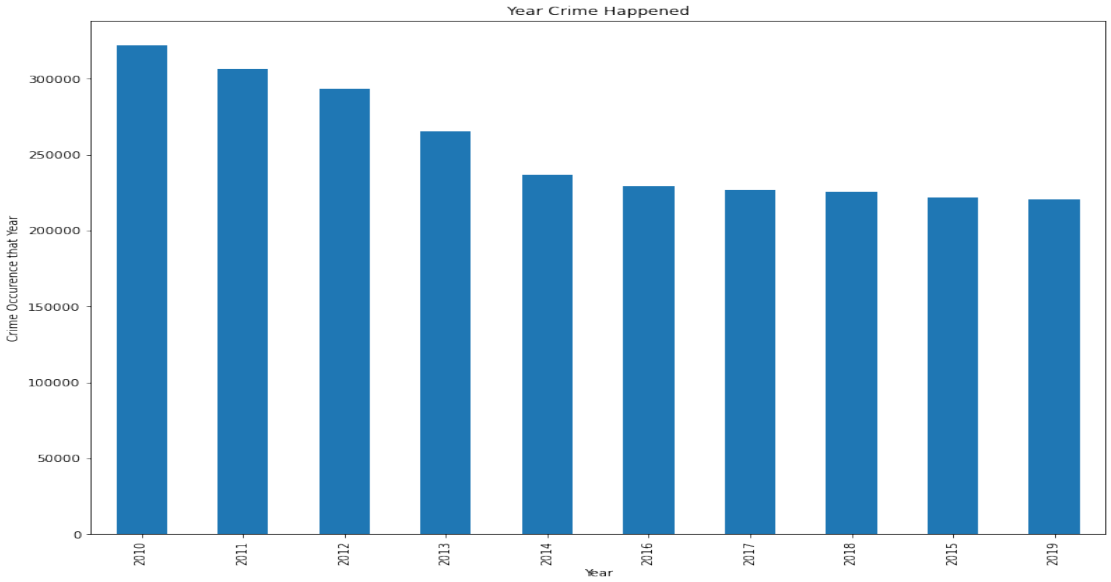
4. EDA


Hour Crime Happened
Hour Crime Rate (%)
0
1
2
3
4
5
6
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
time_hour

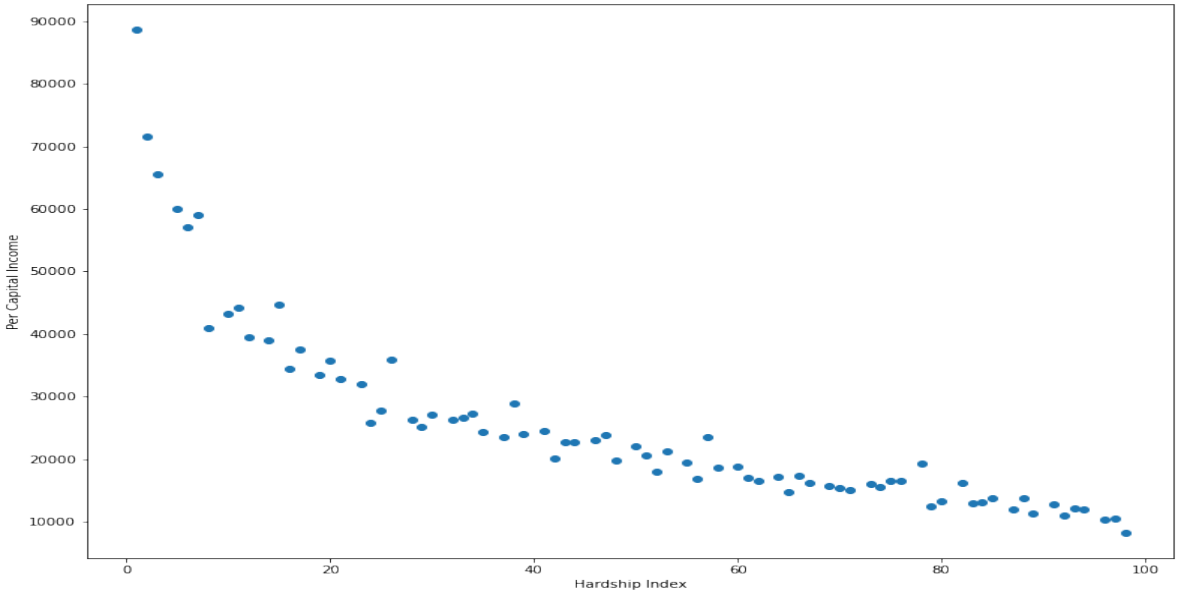

Day Crime Happened
Crime Weekday Rate (%)
0
2
4
6
8
10
12
14
Weekday_Monday
Weekday_Tuesday
Weekday_Wednesday
Weekday_Thursday
Weekday_Friday
Weekday_Saturday
Weekday_Sunday


Month Crime Happened
Crime Month Rate (%)
0
2
4
6
8
01
02
03
04
05
06
07
08
09
10
11
12
month

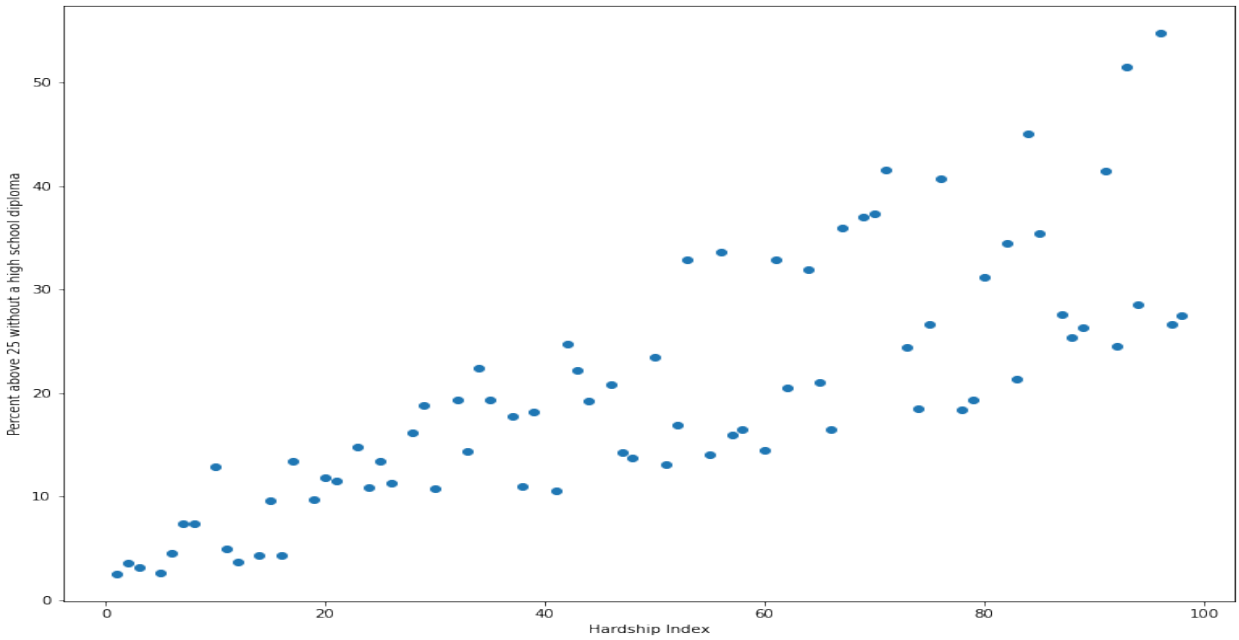

Crime Types by Location
THEFT
BATTERY
CRIMINAL DAMAGE
NARCOTICS
OTHERS
% of Crime Type
Location Description
ALLEY
APARTMENT
BAR OR TAVERN
COMMERCIAL / BUSINESS OFFICE
CONVENIENCE STORE
CTA PLATFORM
CTA TRAIN
DEPARTMENT STORE
GAS STATION
GROCERY FOOD STORE
HOTEL/MOTEL
OTHER
PARK PROPERTY
PARKING LOT/GARAGE(NON RESID.)
RESIDENCE
RESIDENCE PORCH/HALLWAY
RESIDENCE-GARAGE
RESIDENTIAL YARD (FRONT/BACK)
RESTAURANT
SCHOOL, PUBLIC, BUILDING
SCHOOL, PUBLIC, GROUNDS
SIDEWALK
SMALL RETAIL STORE
STREET
VEHICLE NON-COMMERCIAL


Year Crime Happened
Crime Occurence that Year
Year
2010
2011
2012
2013
2014
2016
2017
2018
2015
2019

	n	r	CI95%	r2	adj_r2	p-val	BF10	power
pearson	77	-0.849167	[-0.9, -0.77]	0.721085	0.713547	1.725781e-22	2.584e+19	1.0

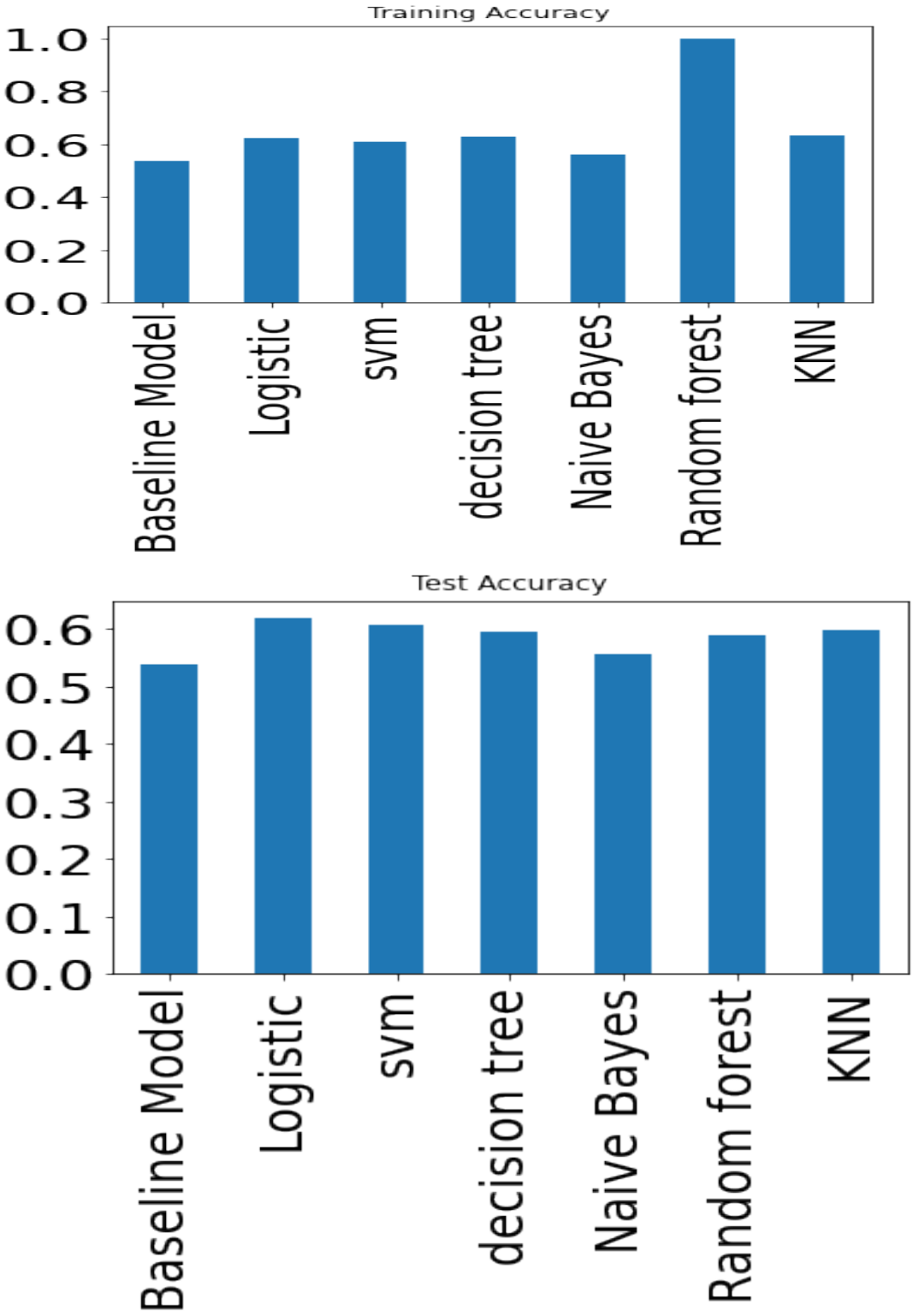


	n	r	CI95%	r2	adj_r2	p-val	BF10	power
pearson	77	0.802538	[0.71, 0.87]	0.644068	0.634448	1.704910e-18	3.507e+15	1.0

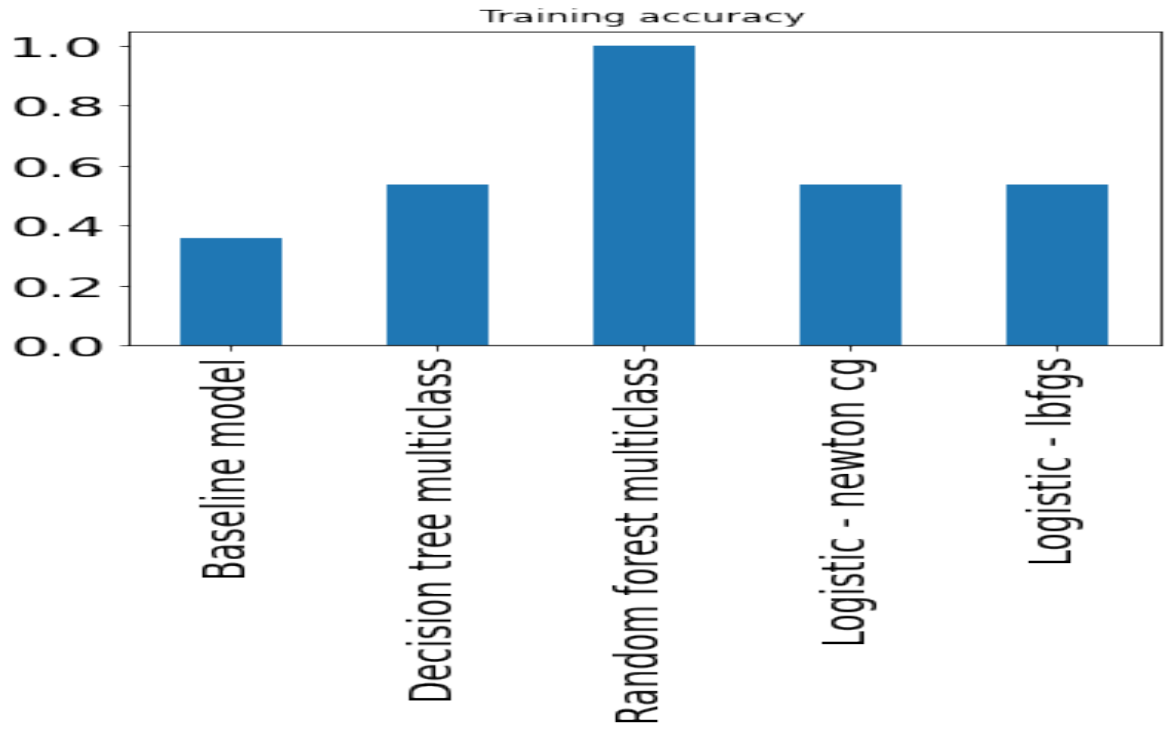


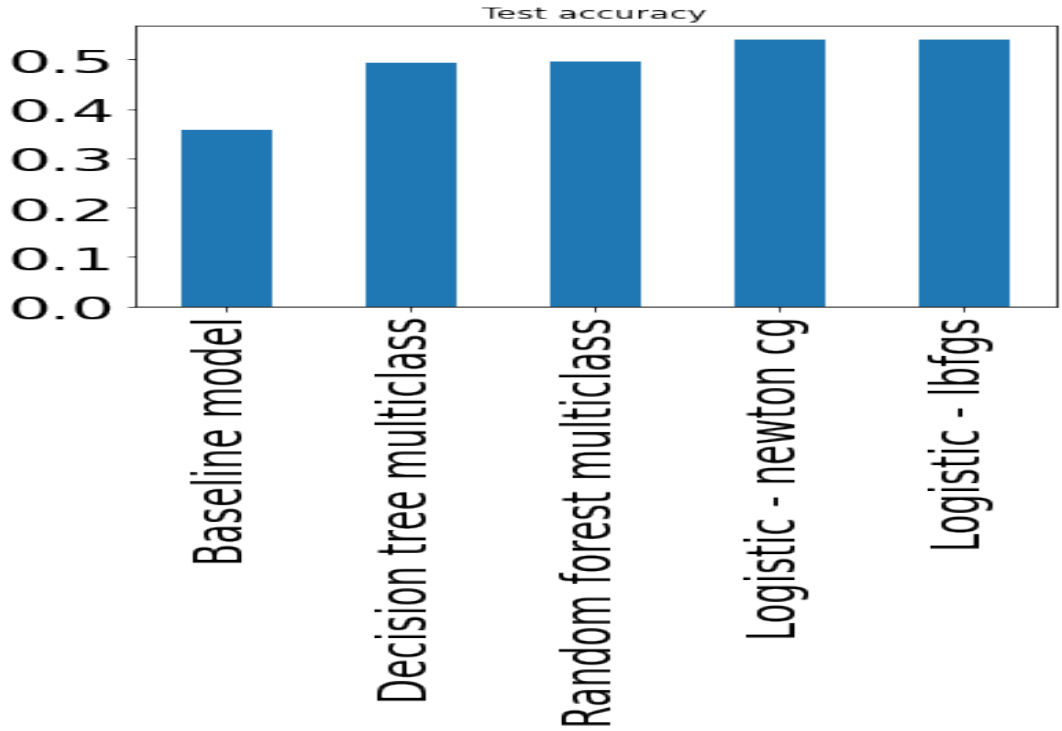
5. Algorithms & Machine Learning

Binary Classification



Multi-Classification





In our training, random forest for both binary and multi-class is the most accurate by far. However, testing is much closer with logistic regression have a slight edge in both classes.

6. Predictions, Conclusions, and Future Work

Binary Features by Coefficients

	feature	coef	abscoef
62	Location Description_DEPARTMENT STORE	-1.640735	1.640735
74	Location Description_SCHOOL, PUBLIC, GROUNDS	1.085269	1.085269
73	Location Description_SCHOOL, PUBLIC, BUILDING	0.902477	0.902477
64	Location Description_GROCERY FOOD STORE	-0.832584	0.832584
56	Location Description_APARTMENT	0.741321	0.741321
39	Timeblock_3	0.659701	0.659701
76	Location Description_SMALL RETAIL STORE	-0.630873	0.630873
69	Location Description_RESIDENCE PORCH/HALLWAY	0.624302	0.624302
70	Location Description_RESIDENCE-GARAGE	-0.420711	0.420711
41	Timeblock_9	-0.392885	0.392885
75	Location Description_SIDEWALK	0.354176	0.354176
34	Timeblock_0	0.346644	0.346644
18	District_D18.0	-0.338996	0.338996
35	Timeblock_12	-0.315350	0.315350
68	Location Description_RESIDENCE	0.295452	0.295452

Multi-class Features by Coefficients

	feature	coef	abscoef
62	Location Description_DEPARTMENT STORE	1.955381	1.955381
64	Location Description_GROCERY FOOD STORE	1.469881	1.469881
55	Location Description_ALLEY	-1.346190	1.346190
57	Location Description_BAR OR TAVERN	1.317098	1.317098
75	Location Description_SIDEWALK	-1.200522	1.200522
58	Location Description_COMMERCIAL / BUSINESS OFFICE	1.086883	1.086883
76	Location Description_SMALL RETAIL STORE	1.062653	1.062653
74	Location Description_SCHOOL, PUBLIC, GROUNDS	-0.946189	0.946189
65	Location Description_HOTEL/MOTEL	0.846309	0.846309
56	Location Description_APARTMENT	-0.834617	0.834617
60	Location Description_CTA PLATFORM	-0.791962	0.791962
66	Location Description_PARK PROPERTY	-0.778364	0.778364
73	Location Description_SCHOOL, PUBLIC, BUILDING	-0.771870	0.771870
59	Location Description_CONVENIENCE STORE	0.694312	0.694312
68	Location Description_RESIDENCE	-0.551704	0.551704
77	Location Description_STREET	-0.520202	0.520202

We can conclude that crimes happening in departments stores are most likely to thefts Crimes happening late at night can tend to be very violent. Crimes are more likely to happen in the summer as Chicago has sub-freezing temperatures in the months of January and February. A combination of features is way more influential than using one single variable to determine a crime. For future, there is certainly other variables that can be added to get even better accuracy. One example was to ask the police their feedback on what crime they are normally looking for. Certain cops may have certain tendencies to look for a specific crime in an area. I would also love to incorporate other cities and see if Chicago has similar crime tendencies to other major U.S. cities. Overall, we have a very powerful start to see whether a specific area is considered a crime hotspot or not.