Chicago Crime Classification

Evan Nussbaum, MBA

The Problem

. Chicago has one of the highest crime rates substantially higher than the US average

. Crime in Chicago has been tracked for several years

. Goal is to provide better insights into the patterns of

crime in Chicago by classification

Classification Types

. Binary Classification: Determining whether a crime was considered serious or not

. Multi-Classification: Took the 4 most common crimes and differentiated them into a function

Who might care?

Cops

Chicago Residents

Tourists

Factors to Consider

. Hardship Index of neighborhood

. Crime Location

. Distance to police station

. Location Description

. Primary Crime Type

Data Information

. Data has over 6 million rows and 112 columns after cleaning

. First dataset is Chicago crime records from the last decade in CSV format

. Second dataset is socio-economic status for the different Chicago neighborhoods in CSV format

. Third dataset is the police district and where they are located in CSV format

	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	...	Ward	Community Area	FBI Code	X Coordi
60282	11556037	JC103643	01/03/2019 07:20:00 PM	0000X W RWY 27R	2890	PUBLIC PEACE VIOLATION	OTHER VIOLATION	AIRCRAFT	False	False	...	41.0	76.0	26	110037
62200	11626027	JC188126	03/16/2019 05:58:00 PM	001XX N WELLS ST	0460	BATTERY	SIMPLE	STREET	False	False	...	42.0	32.0	08B	117472
			03/12/2019					RESIDENTIAL							

Data Exploration

. Crime Types by Location

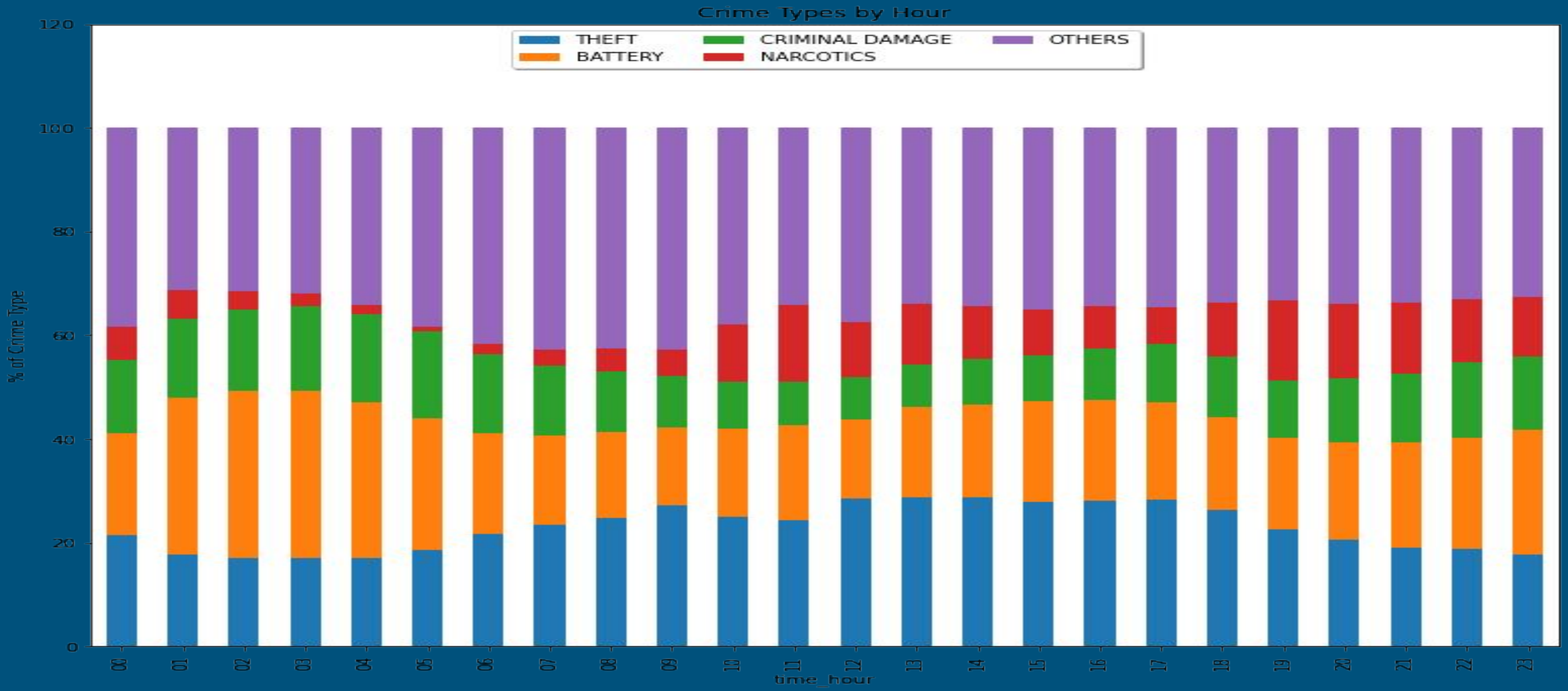
. Crime Types by Hour

. Crime Types by Month

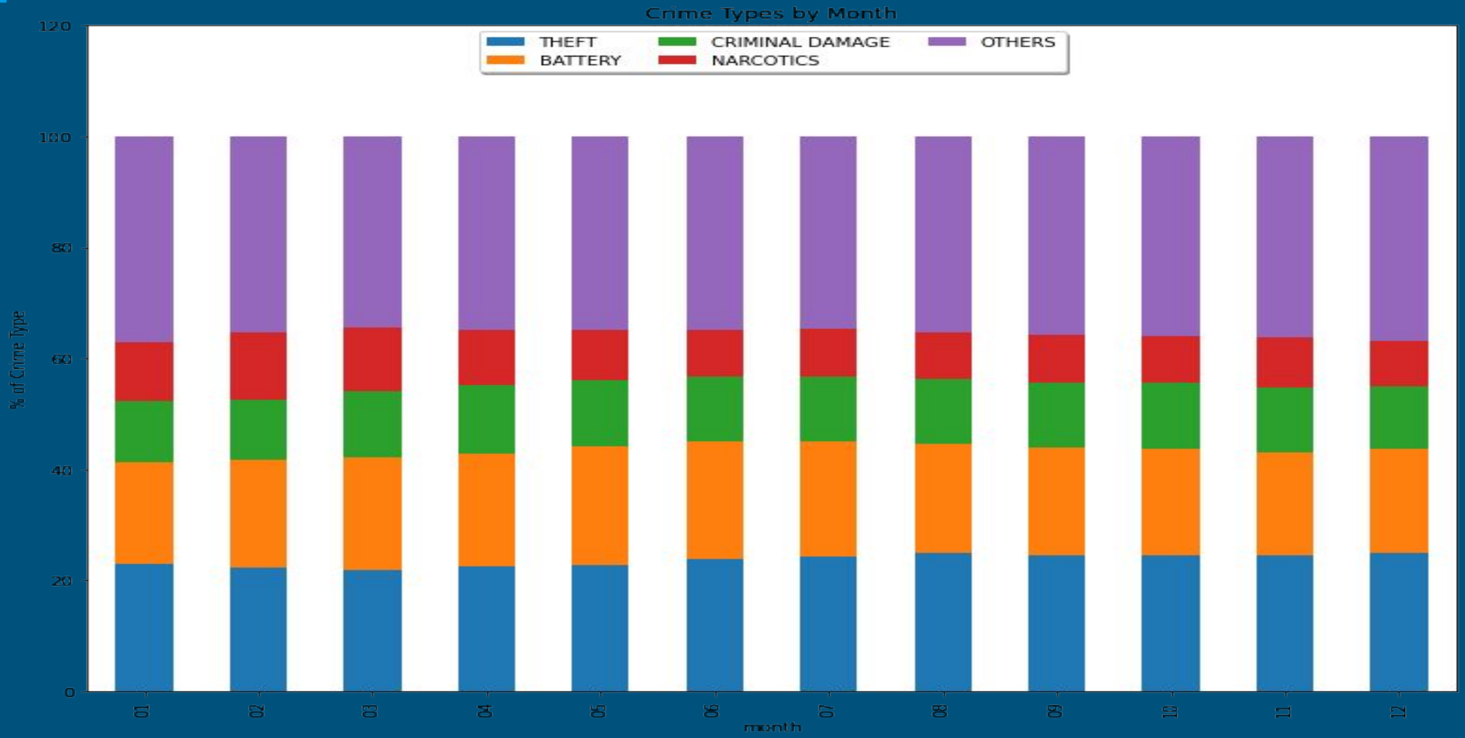
. Socio - Economic Relationships

Crime Types by Location

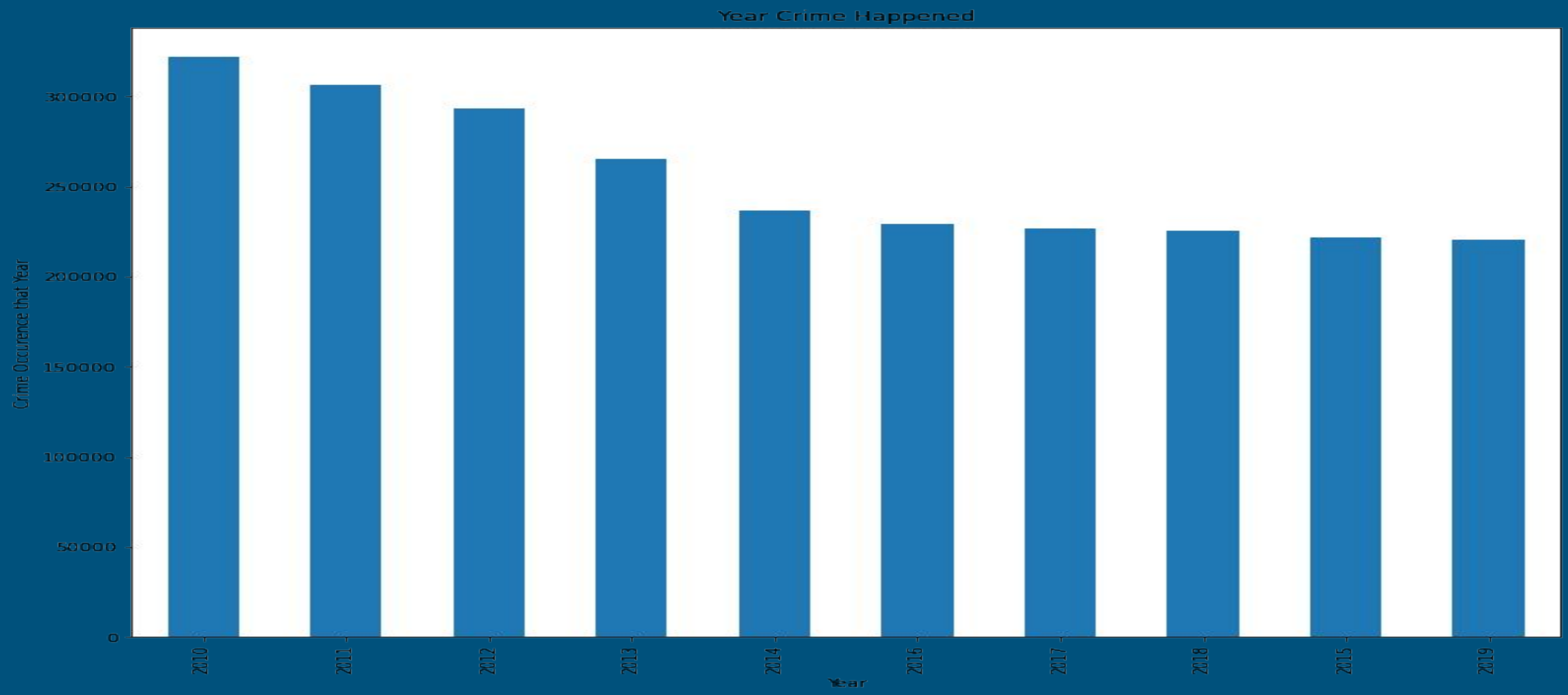
Crime Types by Hour



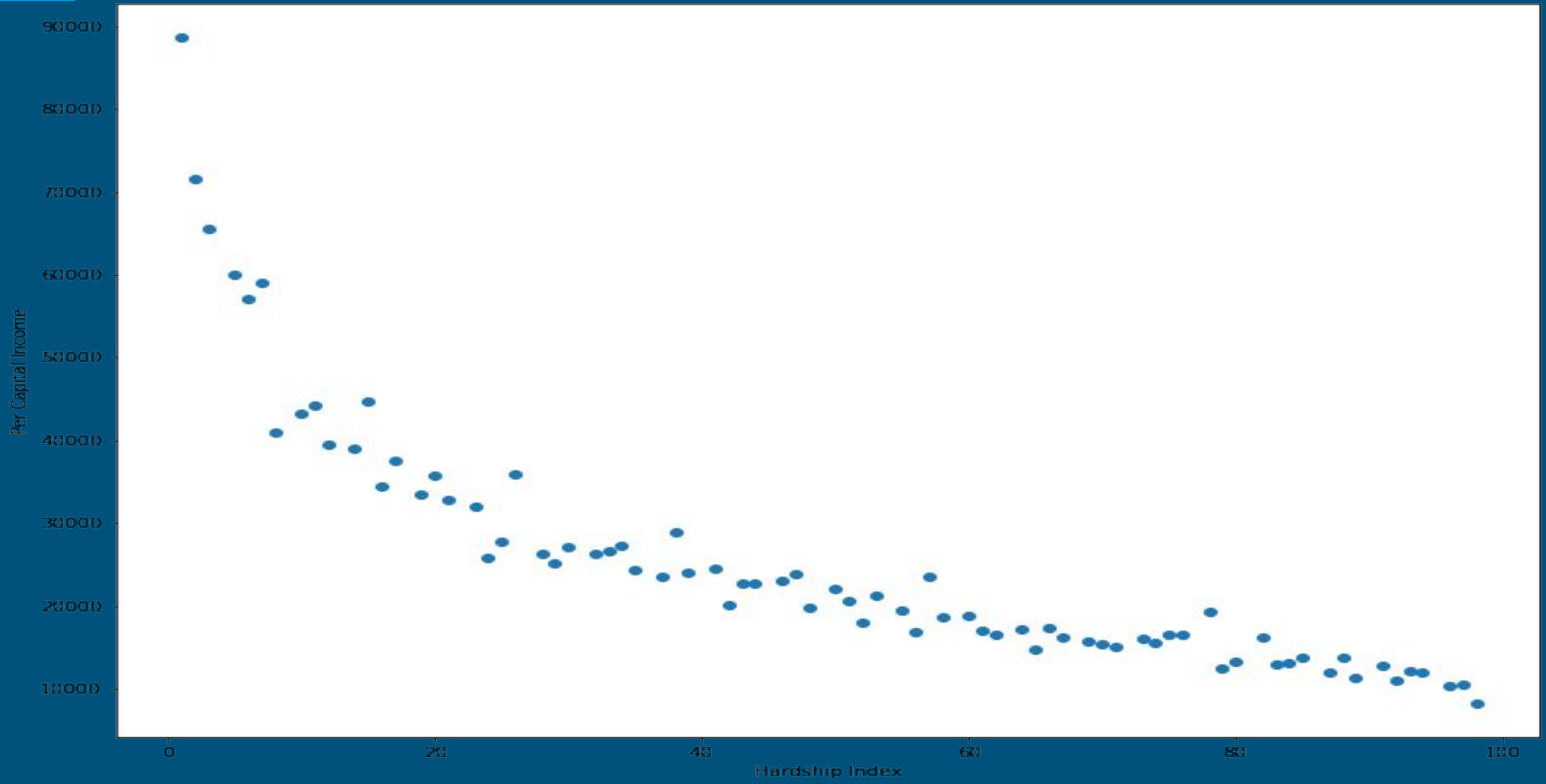
Crime Types by Month



Crime Occurrence by Year



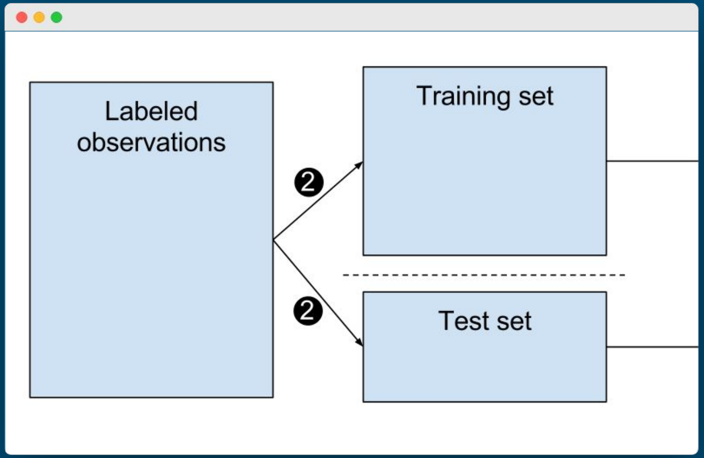
Hardship Index and Per Capita Income



Hardship Index and Percent without high school diploma

Pre-Processing

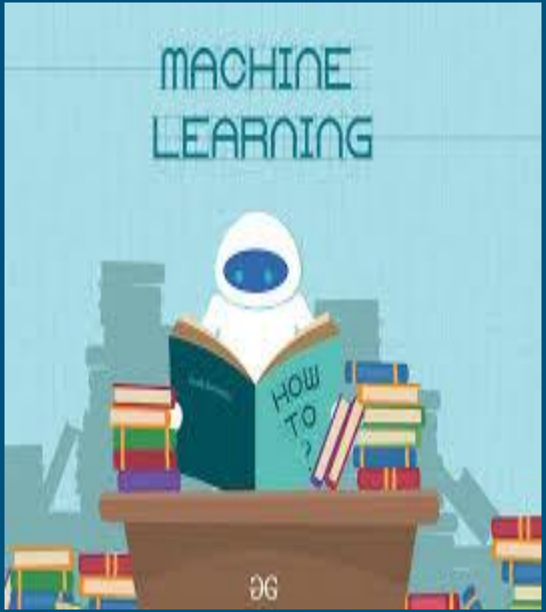
1. Label Encoding
2. Train and Test Split of 70-30
3. Scaling
4. 5 fold cross validation
5. Using scikit learn grid search method



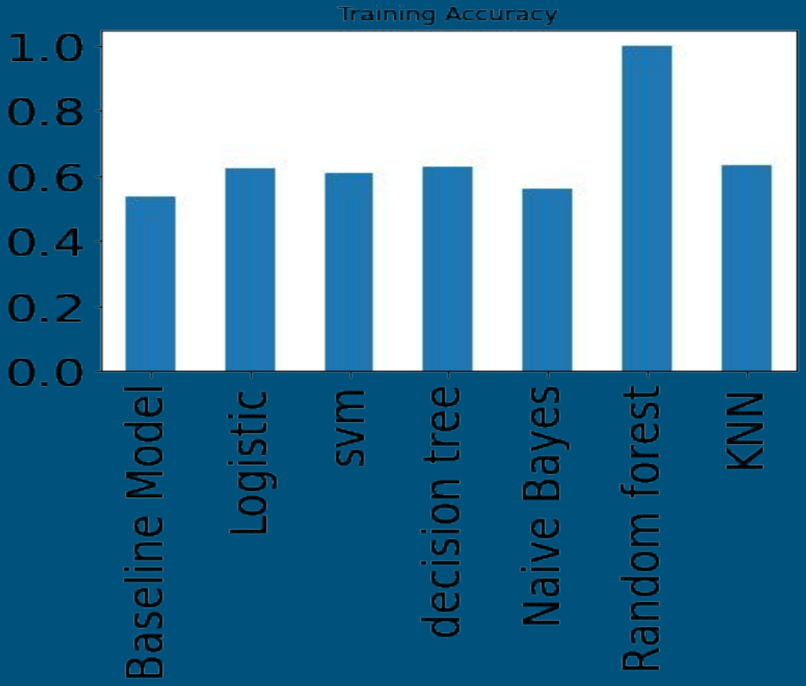
Machine Learning Overview

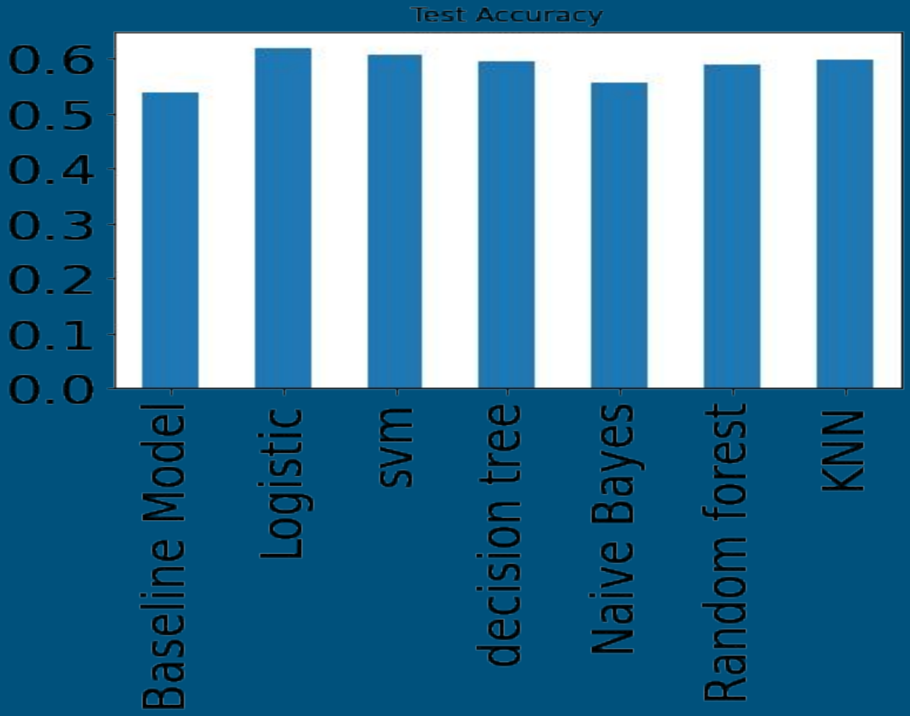
. Supervised Learning Methods

1. Logistic Regression with Lasso
2. SVM
3. Decision Tree
4. Naive Bayes
5. Random Forest
6. KNN

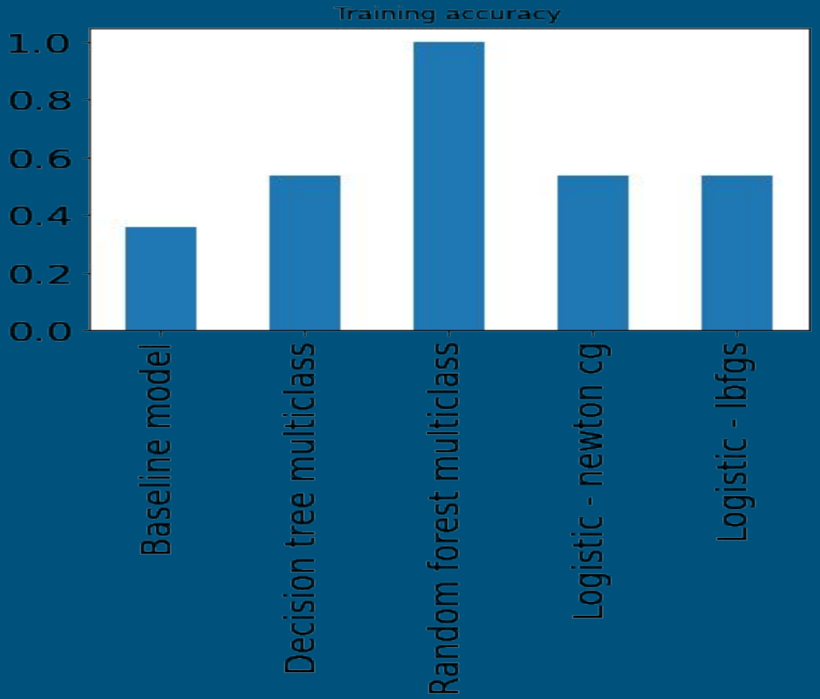


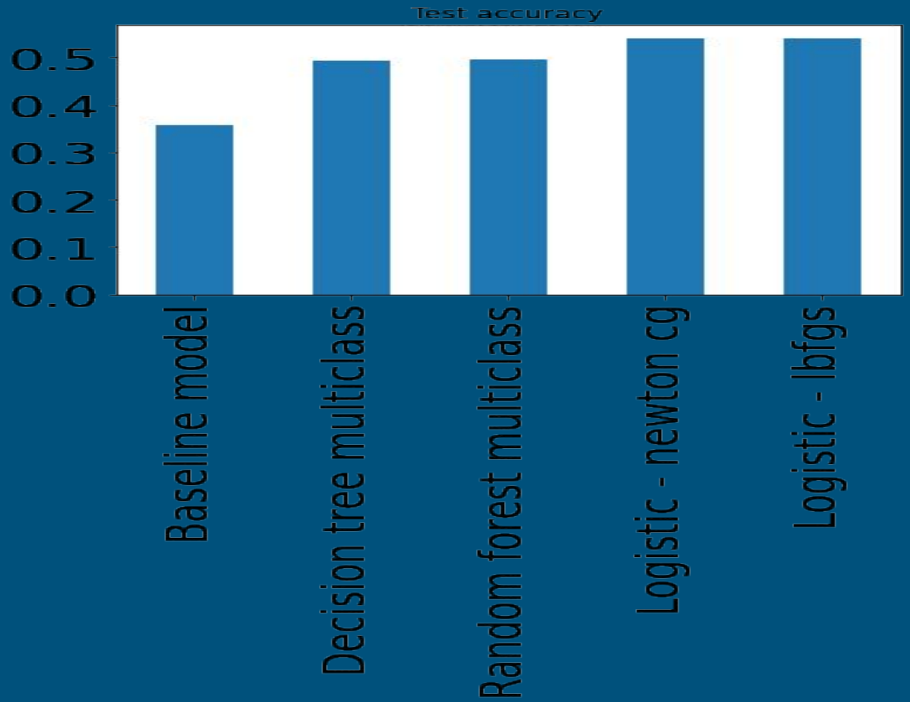
Training and Testing Accuracy Binary Classification





Training and Testing Accuracy Multi-Classification





Most important features Binary Classification

. Department Stores

. Schools

. Grocery Stores

. Apartments

. Crimes that happen from 12-3AM

	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
45	Weekday_Sunday	0.242579	0.242579
12	District_D11.0	-0.206649	0.206649

Most important Features Multi-Classification

. Department Stores

. Grocery Food Stores

. Alleys

. Bars

. Sidewalks

	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
61	Location Description_CTA TRAIN	0.437065	0.437065

Conclusions

. Theft is more likely to happen in Department Stores

. Crime happening late at night can be very violent

. Random forest works very well in training but not as well in testing

. A combination of features is way more influential than just a single feature

Thank You!

Evan Nussbaum, MBA

Email: enussba1@villanova.edu

https://www.linkedin.com/in/evan-nussbaum-2969b2a8/

https://github.com/ejnuss95/Springboard