UP SCHOOL DATA SCIENCE

Getir Data Case: Crimes in Boston

Content

1. Know How (15 min)
2. Know Self (5 min)
3. Q&A (5 min)

1. Know How

- ➔ Problem & Goal
- ➔ Data
- ➔ EDA

- ➔ 1.Baseline Model
- ➔ 2.Model: Predict for each day for Ucr Part 3
- ➔ 3.Model: Predict district & location
- ➔ All Models
- ➔ Next Steps

- ➔ What excited you about your capstone project?
- ➔ What was the most challenging issue for you in the project? How did you get over it?

Problem & Goal

For me it is important that crime is predictable. Thus, Boston Police Forces can focus on specific districts on certain days. They can use existing human and other resources more efficiently.

In this study, I focused on 2 main problems.

A. How many crimes can be committed in a day?

B. In which region can the crimes be committed?

Data

Crimes in Boston data from Kaggle were used in the study. Data contains information about the crime such as date, location, crime group, crime code. It contains 319073 rows and 17 columns. The city's tabular data lists crimes that took place from 2015 to 2018.

```
data.head()
```

	INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE	YEAR	MONTH	DAY_OF_WEEK	HOUR	UCR_PART	STREET	Lat	Long	Location
0	I182070945	619	Larceny	LARCENY ALL OTHERS	D14	808	NaN	2018-09-02 13:00:00	2018	9	Sunday	13	Part One	LINCOLN ST	42.357791	-71.139371	(42.35779134, -71.13937053)
1	I182070943	1402	Vandalism	VANDALISM	C11	347	NaN	2018-08-21 00:00:00	2018	8	Tuesday	0	Part Two	HECLA ST	42.306821	-71.060300	(42.30682138, -71.06030035)
2	I182070941	3410	Towed	TOWED MOTOR VEHICLE	D4	151	NaN	2018-09-03 19:27:00	2018	9	Monday	19	Part Three	CAZENOVE ST	42.346589	-71.072429	(42.34658879, -71.07242943)
3	I182070940	3114	Investigate Property	INVESTIGATE PROPERTY	D4	272	NaN	2018-09-03 21:16:00	2018	9	Monday	21	Part Three	NEWCOMB ST	42.334182	-71.078664	(42.33418175, -71.07866441)
4	I182070938	3114	Investigate Property	INVESTIGATE PROPERTY	B3	421	NaN	2018-09-03 21:05:00	2018	9	Monday	21	Part Three	DELHI ST	42.275365	-71.090361	(42.27536542, -71.09036101)

Data

Descriptive Columns	Time Related Columns	Location Related Columns
Incident_Number	Occured_On_Date	District
Offense_Code	Year	Reporting_Area
Offense _Code_Group	Month	Street
Offense_Description	Day_Of_Week	Location
Shooting	Hour	Lat
Ucr_Part		Long

EDA

Q1: How has crime changed over the years?

Q2: Is it possible to predict where or when a crime will be committed?

Q3: What can you say about the distribution of different offenses over the city?

https://www.kaggle.com/sengulkrdrl/boston-crime-analysis

A.Baseline Model

Our first goal is to predict the number of crimes that may occur in the future. Why is this important? Because if we know the number of crimes that can be committed, we can take action.

- If more crimes are to be committed, a more intense shift may be applied that day.
- The Police Department can focus on the day.
- Equipment can be supplied according to the number of crimes.

Why linear regression?

After preparing our data, we will try to estimate the number of crimes that will occur per day. Here we will start by using linear regression as it is simple, easy to understand, easy to implement for baseline model.

Why didn't I want to apply time series?

Because my goal is to predict multiple variables together, it can be difficult for timeseries models. The forecast is determined only by the past behavior of the variable in timeseries. ARIMA is a univariate model (working with one variable only) and hence cannot exploit the leading indicators or explanatory variables.ARIMA requires a lot of time series observations in this dataset.But if we want to handle single variable we can use ARIMA. Our data has a time dimension so we can apply time series.

Why did I choose R-squared as the metric?

Because R-squared tells us how much of variance can be explained by the linear model. I want to develop model to *explain* the variation in Y.

R-squared is conveniently scaled between 0 and 1 and it can be compared with accuracy.

Input:

Features: OccuredDate, Districts & DayofWeek

Target: CaseCount

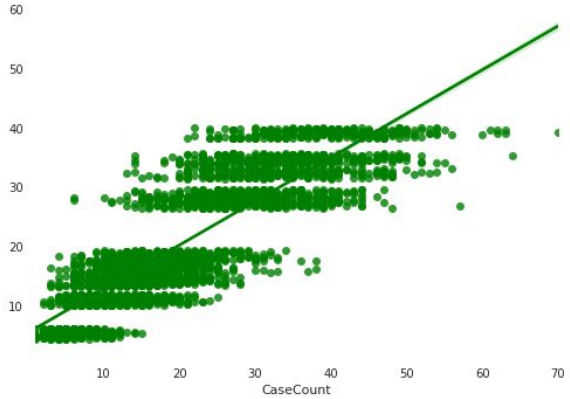
	OccuredDate	CaseCount	DayOfWeek	D_A1	D_A15	D_A7	D_B2	D_B3	D_C11	D_C6	D_D14	D_D4	D_E13	D_E18	D_E5
0	735764	23	1	1	0	0	0	0	0	0	0	0	0	0	0
1	735764	5	1	0	1	0	0	0	0	0	0	0	0	0	0
2	735764	10	1	0	0	1	0	0	0	0	0	0	0	0	0
3	735764	34	1	0	0	0	1	0	0	0	0	0	0	0	0
4	735764	26	1	0	0	0	0	1	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
14104	736940	9	1	0	0	0	0	0	0	0	1	0	0	0	0
14105	736940	20	1	0	0	0	0	0	0	0	0	1	0	0	0
14106	736940	4	1	0	0	0	0	0	0	0	0	0	1	0	0
14107	736940	10	1	0	0	0	0	0	0	0	0	0	0	1	0
14108	736940	3	1	0	0	0	0	0	0	0	0	0	0	0	1

14109 rows × 15 columns

Output:

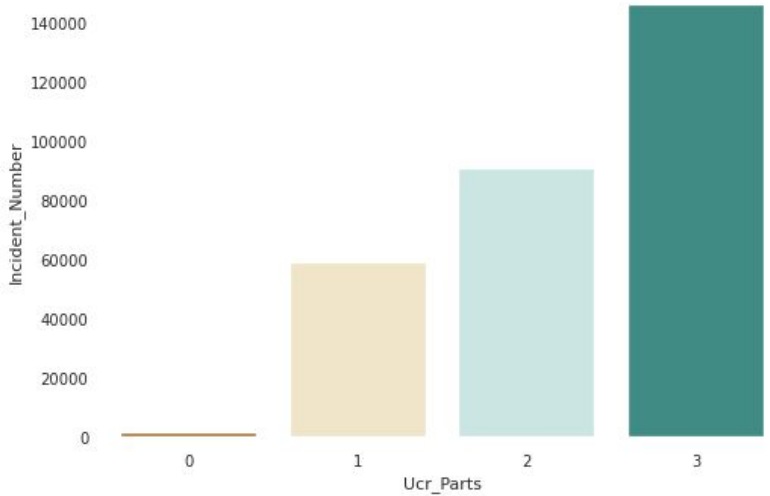
R2 Score: 0.741

(Linear Regression)



B.Model: Predict for each day for Ucr Part 3

We are now customizing the model. Ucr_Part 3 is the Ucr part type with the highest number of crimes. Therefore, we will try to estimate how many crimes can be committed from Ucr_Part 3 in the future.



Input:

Features: DateD14, DayofMonth, Month, Weekday, countD14,...

Target: countUCR

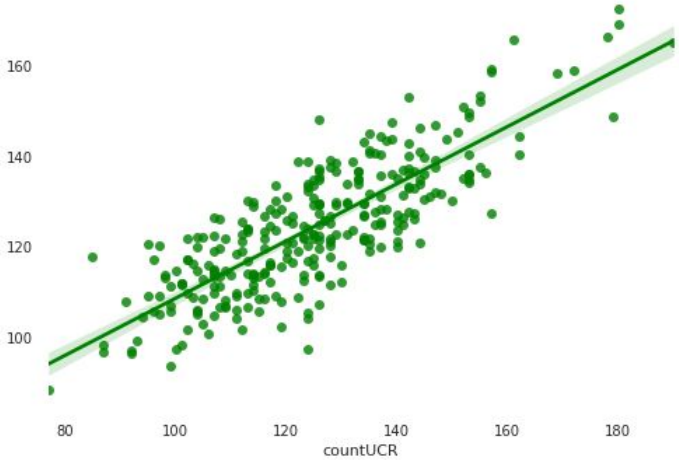
	countUCR	DateD14	countD14	countC11	countD4	countB3	countB2	countC6	countA1	countE5	countA7	countE13	countE18	countA15	DayofMonth	Month	Weekday
0	116	2015-06-15	16	31	40	26	34	16	23	18	10	8	12	5.00000	15	6	0
1	108	2015-06-16	10	41	32	24	39	20	25	7	13	5	21	5.00000	16	6	1
2	90	2015-06-17	13	32	36	25	26	12	24	13	14	7	21	2.00000	17	6	2
3	141	2015-06-18	11	34	30	34	70	16	29	9	8	20	15	8.00000	18	6	3
4	116	2015-06-19	20	36	31	39	58	14	33	11	6	8	17	3.00000	19	6	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1172	126	2018-08-30	24	37	29	31	27	25	23	8	8	10	15	5.30895	30	8	3
1173	138	2018-08-31	27	34	22	21	43	25	16	15	11	21	14	5.30895	31	8	4
1174	142	2018-09-01	30	33	22	27	35	18	25	13	5	15	20	5.30895	1	9	5
1175	111	2018-09-02	17	44	16	21	29	16	17	9	7	10	16	5.30895	2	9	6
1176	89	2018-09-03	9	23	20	22	34	9	20	3	3	4	10	5.30895	3	9	0

1177 rows × 17 columns

Output:

R2 Score: 0.665

(Linear Regression)



C.Model: Predict District

Now, we want to predict the region. If we can predict in which region the crime will occur, the police can work by focusing on the regions. It can help us plan and guide patrol cars. Patrol cars can route in busy places during peak hours.

In other words, it is important to predict in which district a crime may occur in order to prevent crime.

Why XGB Classifier?

We will be using XGB Classifier for its ease of use and predictive power.

There are too many rows in my dataset.

Why accuracy metric?

I'm just curious about the correct classifications.

Input:

Features: Year, Seasons, Hour, Shooting, DayOfWeek, Ucr_Parts,...

Target: District

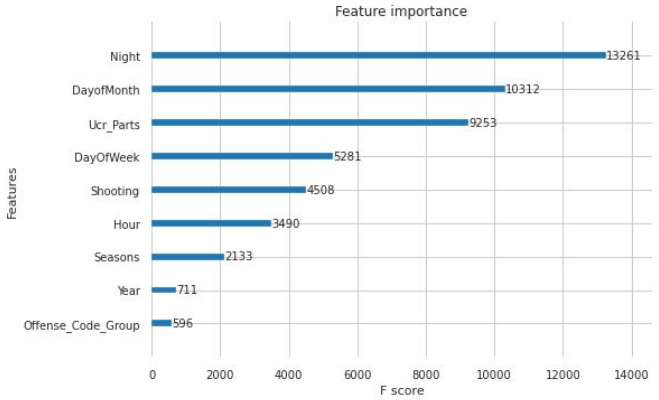
	Offense_Code_Group	District	Year	Seasons	Hour	Shooting	DayOfWeek	Ucr_Parts	DayofMonth	Night
149	36	B2	2018	0	9	0	1	3	3	0
137	58	E18	2018	0	10	0	1	3	3	0
127	24	C11	2018	0	11	0	1	2	3	0
128	60	C11	2018	0	11	0	1	3	3	0
129	27	A1	2018	0	11	0	1	3	3	0
...	...	...	...	...	...	...	...	...	...	...
318627	58	D4	2015	2	10	0	1	3	15	0
318626	31	D4	2015	2	0	0	1	1	15	1
318625	60	D4	2015	2	12	0	1	3	15	0
318624	28	B2	2015	2	12	0	1	3	15	0
318560	27	B3	2015	2	16	0	1	3	15	0

296573 rows × 10 columns

Output:

Accuracy: 0.20

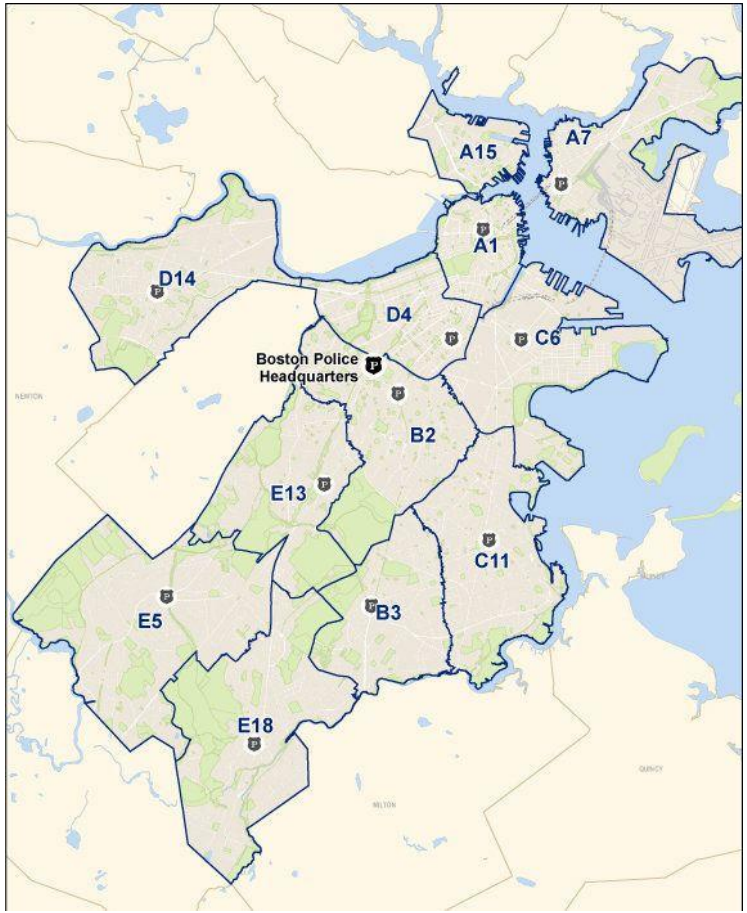
(XGB Classifier)



While predicting the districts, our model was not successful enough. The reasons for this may be the following.

- Districts are close to each other and have no clear boundaries.
- We have a large number of districts.
- We don't have enough data.

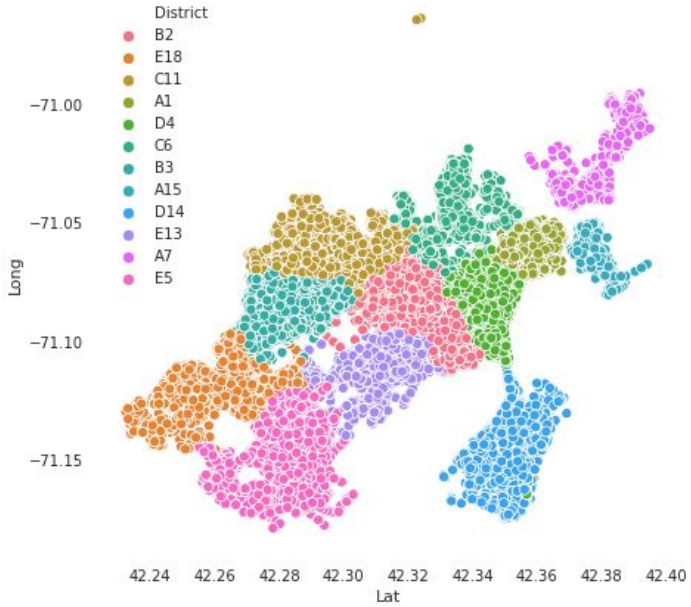
For this reason, we will try to model the districts by grouping them among themselves. So I aim for better prediction.



Model 3 with Grouped Districts

We divided the districts into 3 groups according to their distance.

We will try to classify the district groups.



Input:

Features: Year, Seasons, Hour, Shooting, DayOfWeek, Ucr_Parts

Target: District Group

	Offense_Code_Group	District	Year	Seasons	Hour	Shooting	DayOfWeek	Ucr_Parts	DayofMonth	Night
149	36	2	2018	0	9	0	1	3	3	0
137	58	1	2018	0	10	0	1	3	3	0
127	24	2	2018	0	11	0	1	2	3	0
128	60	2	2018	0	11	0	1	3	3	0
129	27	3	2018	0	11	0	1	3	3	0
...	...	...	...	...	...	...	...	...	...	...
318627	58	2	2015	2	10	0	1	3	15	0
318626	31	2	2015	2	0	0	1	1	15	1
318625	60	2	2015	2	12	0	1	3	15	0
318624	28	2	2015	2	12	0	1	3	15	0
318560	27	1	2015	2	16	0	1	3	15	0

296573 rows × 10 columns

Output:

Accuracy: 0.49

(XGB Classifier)

	precision	recall	f1-score	support
1	0.58	0.00	0.01	16014
2	0.49	0.99	0.66	29119
3	0.53	0.02	0.04	14182
accuracy			0.49	59315
macro avg	0.54	0.34	0.23	59315
weighted avg	0.53	0.49	0.33	59315

All Models

	Model 1:	Model 2:	Model 3:
Predict:	The number of crimes that will occur per day	The number of crimes that will occur per day for Ucr_Part 3	In which district group the crime can be committed
Model:	Linear Regression	Linear Regression	XGB Classifier
Metric:	R-squared: 0.74	R-squared: 0.66	`Accuracy: 0.49`

Next Steps

- The dataset as such is not suitable for regression and classification problems.If we want to make a similar prediction or classification, a stronger preprocessing is required.
- Multivariate Time Series can be applied to estimate the number of crimes that can occur in a given time period.
- With the clustering method, we can identify cluster centers and insert them into the model as features. Also, when a crime occurs, the patrol car in the nearest center can quickly reach the crime position.
- We can build a model for each district. Was this crime committed in D4 district?

What was the most challenging issue for you in the project? How did you get over it?

Asking the right questions to the data.I was worried that I was going to **determine the problem**. I have tried many things with the data.

Initially, I didn't know how to **approach** data. I overcame this by *trying and working planned*. I think this work has contributed a lot to me in **understanding data**.

What excited you about your capstone project?

- To be looking for a real solution to a **real problem**. I think the answers I am looking for in this problem are similar in the industry. I wanted to create a really usable product.
- Working with more complex data was challenging and at the same time I felt I was learning new things.

2. Know Self

- ➔ In what areas do you think you have improved during your experience at UP School?

- ➔ In which area will you continue to improve yourself from now on?

In what areas do you think you have improved during your experience at UP School?

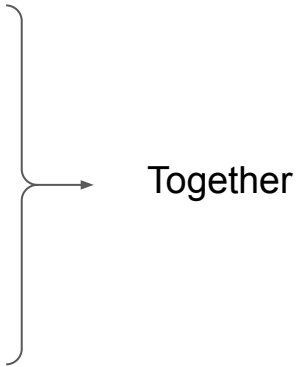
Clarity of Purpose

Solving Challenging Problems

Time Management

Lifelong Learning

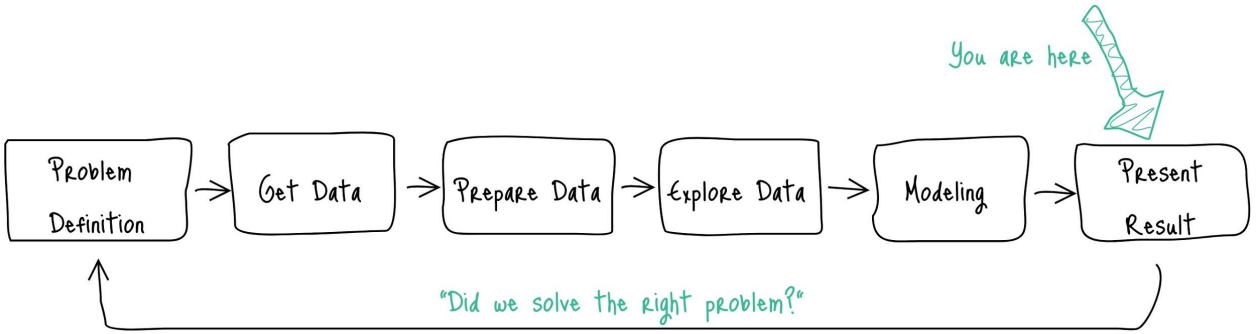
Feed Forward



In which area will you continue to improve yourself from now on?

I want to work as a data scientist, apply what I have learned here and solve different problems. In the second phase, I plan to do a master's degree in order to have deeper knowledge in research. I think the industry and the academy feed each other. Therefore, there are things to learn from both.

3. Q&A


You are here
Problem Definition
Get Data
Prepare Data
Explore Data
Modeling
Present Result
"Did we solve the right problem?"

Thank you for listening!

Şengül Karaderili

sengulkaraderili@gmail.com