CrimeIntel

A Data-Mining Approach to Public Safety

Max Smith | Aaron Shyuu | Thomas Dolan

Data Prep

- ETL Pipeline organized our preprocessing.
- **Specifics:**
  - Queried remote database
  - Transformed to DataFrame
  - Cleaned Data
  - Transform data types
  - Integrated supplementary data

```python
def join_with_econ(df):
    """
    Joins crime data with revised_econ data
    """
    econ_df = pd.read_excel("../data/cfnai/cfnai-realtime-revised.xlsx")
    econ_df['month'] = econ_df.Month.dt.month
    econ_df['month'] = pd.to_numeric(econ_df.month, downcast="unsigned")
    econ_df['year'] = econ_df.Month.dt.year
    econ_df.drop("Month", inplace=True, axis=1)
    return pd.merge(df, econ_df) # Can't do this inplace :(

def join_with_crimetype(df):
    """
    Joins crime data with revised_econ data
    """
    ctype_df = pd.read_excel("../data/fbi_crime_types/crimetype_SRS.xlsx")
    return pd.merge(df, ctype_df)


def etl_pipeline(df):
    """
    This is the whole pipeline.
    Add function calls to mutate the inputted dataframe into something that we can work with.
    """
    remove_null_objects(df)
    add_dt_attributes(df)
    remove_irrelevant_attributes(df)
    convert_coords_to_geometry(df)
    df = join_with_econ(df)
    df = join_with_crimetype(df)
    return df
```

Tools Used

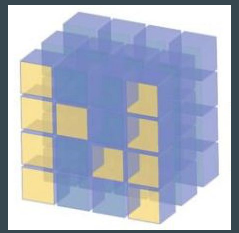
Google BigQuery

Jupyter Notebooks

SQL

Python

Pandas / Numpy / Matplotlib / Sklearn / Seaborn /GeoPandas



Question 1

Crime steadily declined between 2002-2015, then plateaued in 2016-2019. What factors contributed to this plateau?

Data mining technique: Linear Regression

Knowledge gained: 70% of the plateau was caused by 3 crimes: theft, narcotics, and battery

How it can be applied: Law enforcement might seek additional resources and allocate those resources toward combating theft, narcotics, and battery to get back to pre-2016 trends.

Q1 Visuals

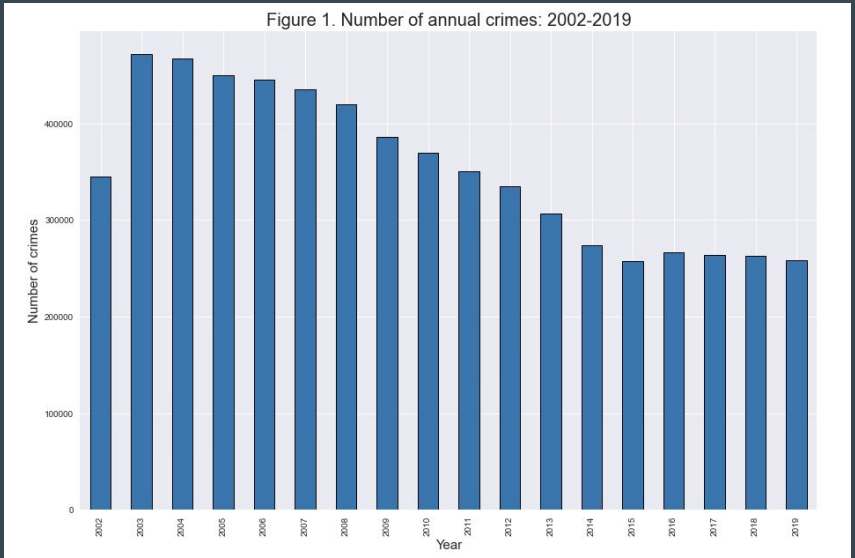


Figure 1. Number of annual crimes: 2002-2019
Number of crimes
0
100000
200000
300000
400000
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Year

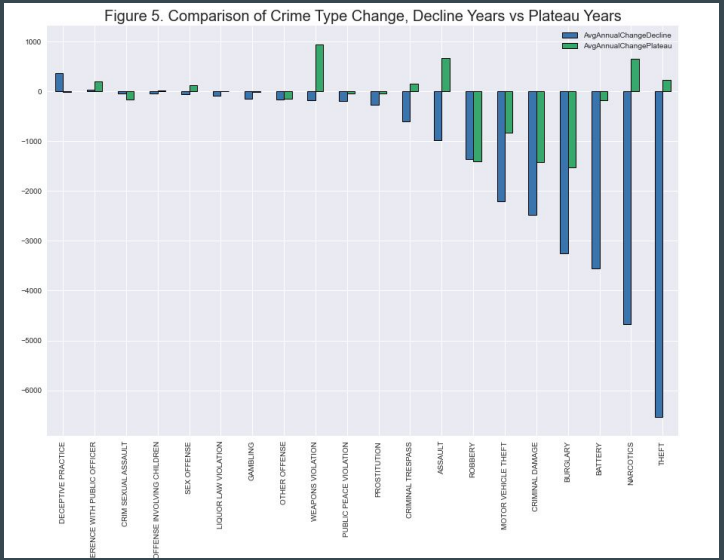


Figure 5. Comparison of Crime Type Change, Decline Years vs Plateau Years
AvgAnnualChangeDecline
AvgAnnualChangePlateau
1000
0
−1000
−2000
−3000
−4000
−5000
−6000
DECEPTIVE PRACTICE
INTERFERENCE WITH PUBLIC OFFICER
CRIM SEXUAL ASSAULT
OFFENSE INVOLVING CHILDREN
SEX OFFENSE
LIQUOR LAW VIOLATION
GAMBLING
OTHER OFFENSE
WEAPONS VIOLATION
PUBLIC PEACE VIOLATION
PROSTITUTION
CRIMINAL TRESPASS
ASSAULT
ROBBERY
MOTOR VEHICLE THEFT
CRIMINAL DAMAGE
BURGLARY
BATTERY
NARCOTICS
THEFT

Question 2

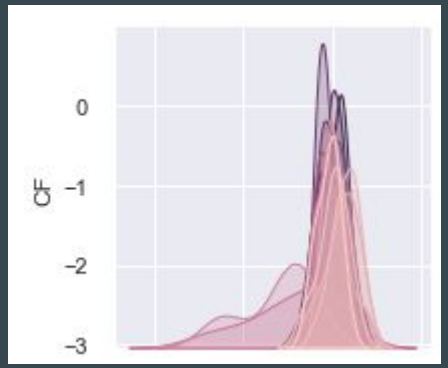
How does economic health correlate to crime?

Data mining technique: Outlier Analysis

Knowledge gained: The economic dropoff caused by COVID-19 restriction are correlated with a major decrease in crimes involving narcotics, theft, assault and battery. Additionally, domestic crimes, as a share of total crime, increased by 2 standard deviations.

How it can be applied: The drop in crime is likely due to COVID due to the increase in domestic crime ratio. Therefore the change is only likely to be temporary and major decisions on policing should not be made based on this decline.

Q2 Visuals


CF
0
−1
−2
−3

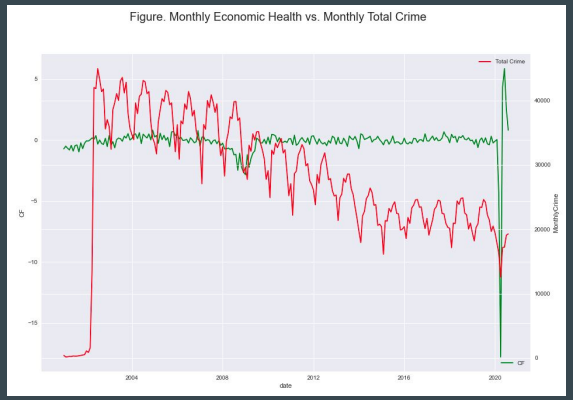
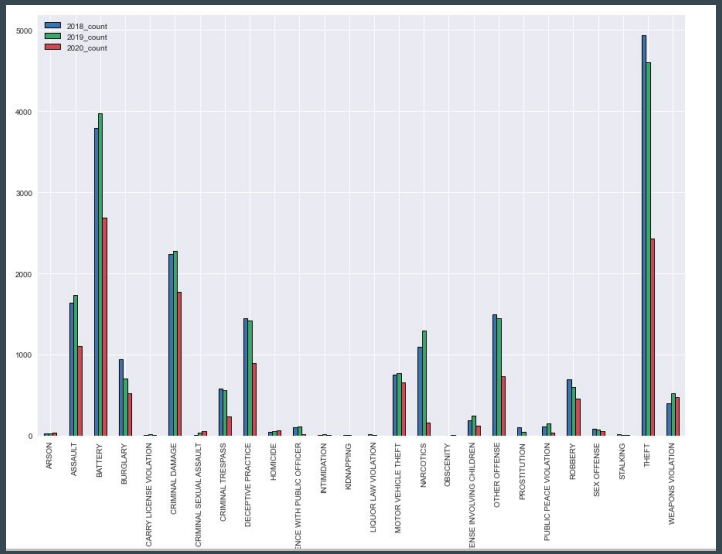


Figure. Monthly Economic Health vs. Monthly Total Crime
Total Crime
CF
MonthlyCrime
date
2004
2008
2012
2016
2020

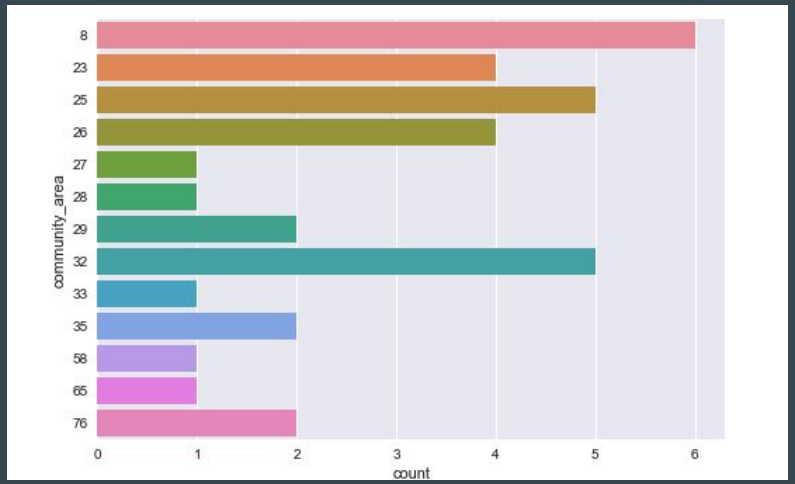

2018_count
2019_count
2020_count
ARSON
ASSAULT
BATTERY
BURGLARY
CARRY LICENSE VIOLATION
CRIMINAL DAMAGE
CRIMINAL SEXUAL ASSAULT
CRIMINAL TRESPASS
DECEPTIVE PRACTICE
HOMICIDE
INTIMIDATION
KIDNAPPING
LIQUOR LAW VIOLATION
MOTOR VEHICLE THEFT
NARCOTICS
OBSCENITY
OTHER OFFENSE
PROSTITUTION
PUBLIC PEACE VIOLATION
ROBBERY
SEX OFFENSE
STALKING
THEFT
WEAPONS VIOLATION

Question 3

What can we learn about where and when specific kinds of crime are happening by analyzing associations between crime types and their location and time data? Do additional crime type categorizations improve these results?

- Data mining technique: Frequent Itemset Mining
- Knowledge gained: We identified three major crime hotspots and gained key insights into the nature of the crime in these locations. We also identified biases in the data as well as gained several new entry points for additional analysis. We proved that additional categorization types can improve the overall analysis and lead to new insights.
- How it can be applied: These insights improve our understanding of the crime occurring in certain times and places, allowing organizations such as law enforcement and community support programs to more effectively allocate resources. We have shown that updated uniform crime reporting programs are effective in improving this understanding and should be adopted.

Question 3 Vis



	rule	lift	confidence	support	crime_code_category	location_description	community_area
6	['RESIDENCE-GARAGE'] ==> Burglary	8.856286	0.505816	0.009495	Burglary	RESIDENCE-GARAGE	NaN
7	[25, 'SIDEWALK'] ==> Drug Abuse	4.930013	0.468830	0.005493	Drug Abuse	SIDEWALK	25.0
5	['DEPARTMENT STORE'] ==> Larceny	3.866943	0.819444	0.010368	Larceny	DEPARTMENT STORE	NaN
3	['Prostitution'] ==> STREET	3.197930	0.819378	0.007665	Prostitution	STREET	NaN
0	['SIDEWALK'] ==> Drug Abuse	3.182014	0.302601	0.030520	Drug Abuse	SIDEWALK	NaN
1	['SIDEWALK'] ==> Robbery	3.149042	0.118870	0.011989	Robbery	SIDEWALK	NaN
2	['STREET'] ==> Motor Vehicle Theft	3.074678	0.139473	0.035736	Motor Vehicle Theft	STREET	NaN
4	['GROCERY FOOD STORE'] ==> Larceny	3.025541	0.641142	0.008058	Larceny	GROCERY FOOD STORE	NaN

Question 4

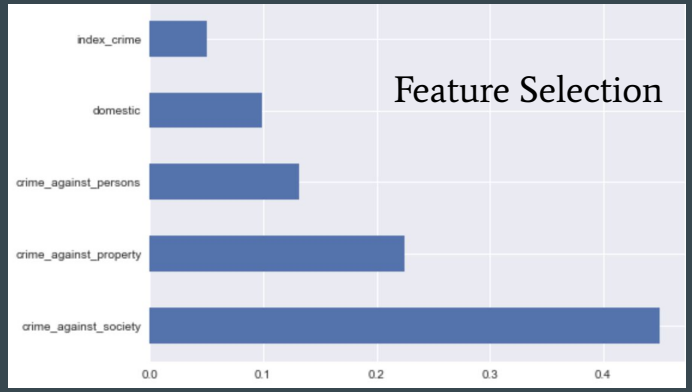
Can we predict whether an Arrest will be made for a commited crime?

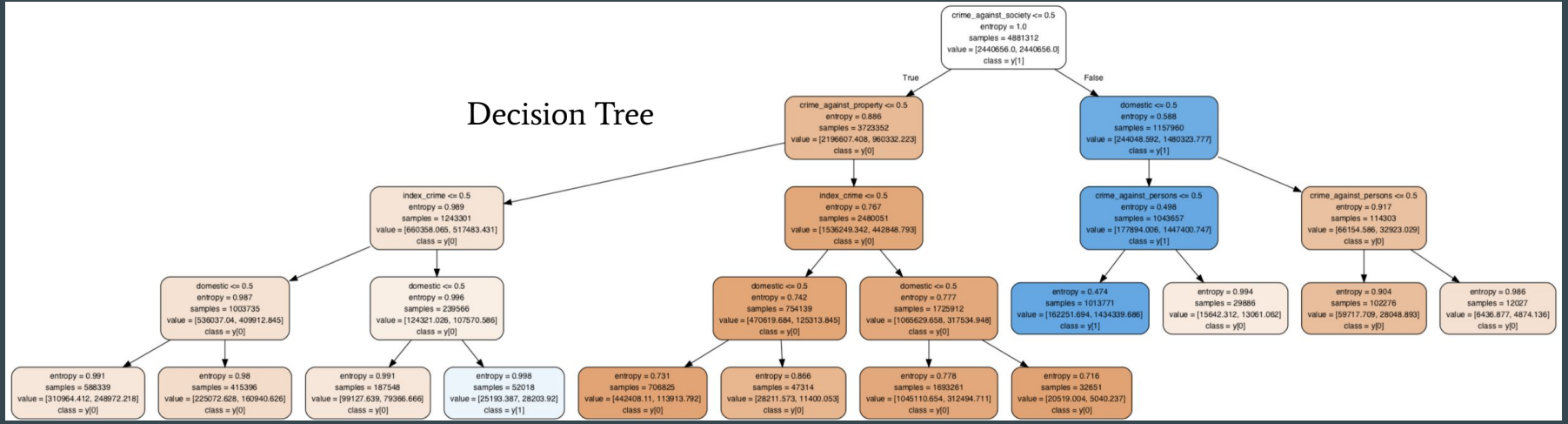
Data mining technique: Classification & Feature Selection (Decision Tree, Naive Bayes, Logistic Regression, K-Nearest Neighbor)

Knowledge gained: Domestic crime, crime against society, crime against property and index crimes are good predictors for the arrest (or non-arrest) of a criminal. Given these predictors, Decision tree, Naive Bayes and Logistic Regression models predict arrest better than KNN based on evaluation metrics.

How it can be applied: A good predictive model with feature selection can help law enforcement allocate resources by understanding the likelihood of arrests and how certain properties of crimes could make arrests more difficult. It also helps with managing prison population.

Q4 Visuals


Feature Selection
index_crime
domestic
crime_against_persons
crime_against_property
crime_against_society
0.0
0.1
0.2
0.3
0.4


Decision Tree
crime_against_society <= 0.5
entropy = 1.0
samples = 4881312
value = [2440656.0, 2440656.0]
class = y[1]
True
False
crime_against_property <= 0.5
entropy = 0.886
samples = 3723352
value = [2196607.408, 960332.223]
class = y[0]
domestic <= 0.5
entropy = 0.588
samples = 1157960
value = [244048.592, 1480323.777]
class = y[1]
index_crime <= 0.5
entropy = 0.989
samples = 1243301
value = [660358.065, 517483.431]
class = y[0]
index_crime <= 0.5
entropy = 0.767
samples = 2480051
value = [1536249.342, 442848.793]
class = y[0]
crime_against_persons <= 0.5
entropy = 0.498
samples = 1043657
value = [177894.006, 1447400.747]
class = y[1]
crime_against_persons <= 0.5
entropy = 0.917
samples = 114303
value = [66154.586, 32923.029]
class = y[0]
domestic <= 0.5
entropy = 0.987
samples = 1003735
value = [536037.04, 409912.845]
class = y[0]
domestic <= 0.5
entropy = 0.996
samples = 239566
value = [124321.026, 107570.586]
class = y[0]
domestic <= 0.5
entropy = 0.742
samples = 754139
value = [470619.684, 125313.845]
class = y[0]
domestic <= 0.5
entropy = 0.777
samples = 1725912
value = [1065629.658, 317534.948]
class = y[0]
entropy = 0.474
samples = 1013771
value = [162251.694, 1434339.686]
class = y[1]
entropy = 0.994
samples = 29886
value = [15642.312, 13061.062]
class = y[0]
entropy = 0.904
samples = 102276
value = [59717.709, 28048.893]
class = y[0]
entropy = 0.986
samples = 12027
value = [6436.877, 4874.136]
class = y[0]
entropy = 0.991
samples = 588339
value = [310964.412, 248972.218]
class = y[0]
entropy = 0.98
samples = 415396
value = [225072.628, 160940.626]
class = y[0]
entropy = 0.991
samples = 187548
value = [99127.639, 79366.666]
class = y[0]
entropy = 0.998
samples = 52018
value = [25193.387, 28203.92]
class = y[1]
entropy = 0.731
samples = 706825
value = [442408.11, 113913.792]
class = y[0]
entropy = 0.866
samples = 47314
value = [28211.573, 11400.053]
class = y[0]
entropy = 0.778
samples = 1693261
value = [1045110.654, 312494.711]
class = y[0]
entropy = 0.716
samples = 32651
value = [20519.004, 5040.237]
class = y[0]