IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Analysis and prediction of crimes in Boston

Fathima Sania

B.E COMPUTER SCIENCE, P D A COLLEGE OF ENGINEERING, KALABURAGI

ABSTRACT

Crime is a major issue every country in the world is tackling in their own way. Crime affects the residents mental state as well as it also effects financially on a larger scale. Here we will be trying to find an optimal solution that will decrease the number of crimes in Boston by trying to build a model that will predict the number of crimes daily and also by analyzing the data and finding insights and information and the root cause. Our main goal here is to somehow decrease the crime rate.

Introduction

In urban analytics field, crimes are a significant phenomenon which need to be concerned about and researched, thus many researches have been done on it and gave us some references. Harries et al. (2006) pointed that most type of crime tend to increase in levels of occurrence with increasing population density, but this relationship could be moderated by socioeconomic status; Narayan et al. (2007) said that fraud, homicide and motor vehicle theft are cointegrated with male youth unemployment rate and real male average weekly earnings; Lochner et al. (2000) found age and education were more negatively correlated with crimes requiring little skills; Lafree et al.(1992) found white crime rates declined as family income and educational attainment increased, which is quite the opposite for African-Americans. As for these theories above, I want to analyze and predict the crimes for fighting the crime rate and play a role in creating a better and safer community.

Dataset:

The dataset used here is the Boston crime data of 8 years i.e., from 2015-01-01 to till date. The data is available here at dataset.boston.gov . The data columns are:

(1) **INCIDENT_NUMBER:** this is the unique ID of every emergency reported.
(2) **OFFENSE_CODE:** This contains the unique code associated with each type of emergency.

(3) **OFFENSE_CODE_GROUP:** This contains the OFFENSE_CODEs associated with particular emergencies as a group.

(4) **OFFENSE_DESCRIPTION:** This is the description of the reported emergency.

(5) **DISTRICT:** This column contains the 12 unique district codes in which Boston is divided.

(6) **REPORTING_AREA:** This shows the area code from where the emergency is reported.

(7) **SHOOTING:** This shows if there is a shooting reported or not on the location of reported emergency.

(8) **OCCURRED_ON_DATE:** The date and time of the occurrence of emergency.

(9) **YEAR:** Year of reported emergency.

(10) **MONTH:** Month of reported emergency.

(11) **DAY_OF_WEEK:** The day of the week on which the emergency is reported.

(12) **HOUR:** The hour of day of reported emergency.

(13) **UCR_PART:** It contains the part of Uniform Crime Reports i.e., Part One (heinous crimes), Part Two (intermediate level crimes), Part Three (low level crimes)

(14) **STREET:** Name of street where the emergencies occurred.

(15) **Lat:** Latitude of occurred emergency.

(16) **Long:** Longitude of occurred emergency.

(17) **Location:** Contains tuple of Lat and Long i.e. (Lat, Long).

The dataset size is vast, which contains ~620,000 rows and 17 columns: 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

Workflow:

In this project, we followed the standard data science workflow:

1. Exploratory data analysis (EDA)

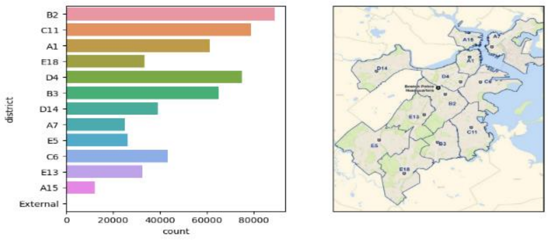
1. Data preprocessing and feature engineering

2. Model training, experiments, and evaluation

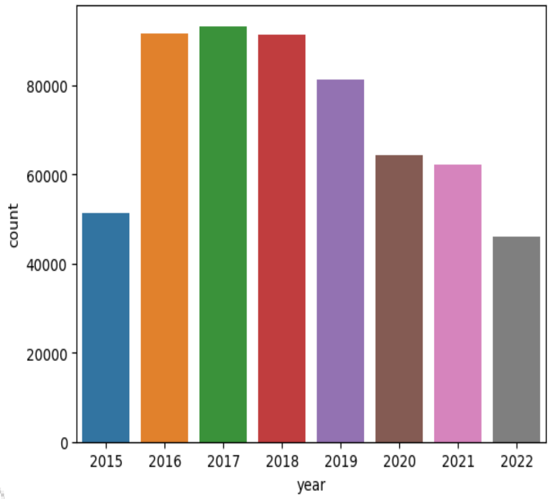
1. Exploratory Data Analysis

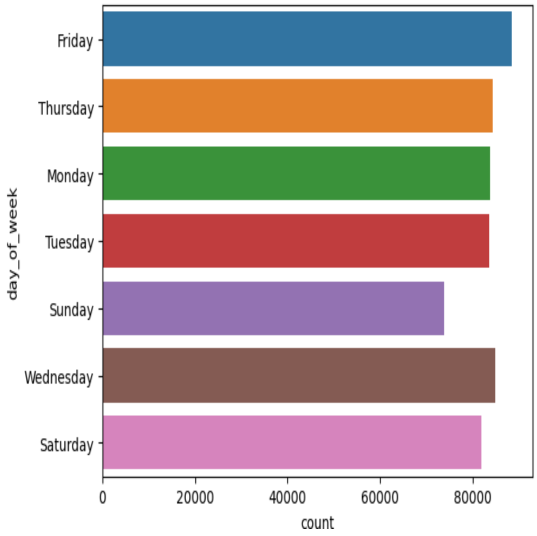
Since the business values and goals are clear, the first step is to clean the data because data is the king always in data science, so we need to keep it clean in order to get the best insights and information out of it and generate the best solution.

After cleaning the data, we understand the dataset by generating visualization from the data and see if we can find some insights. Look at the graph below



District B2 has high crime rate and A15 has lowest

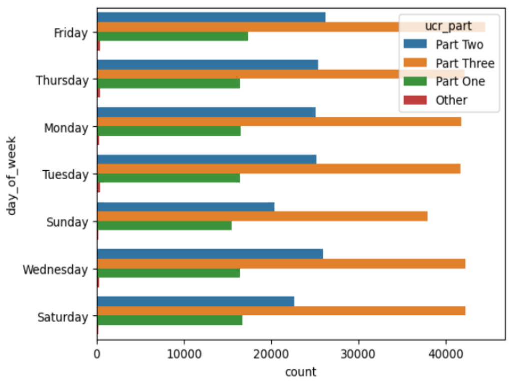




On the left above we see that the crime rate is increasing from 2015 to 2017 and after 2017 the crime rate is decreasing every year.

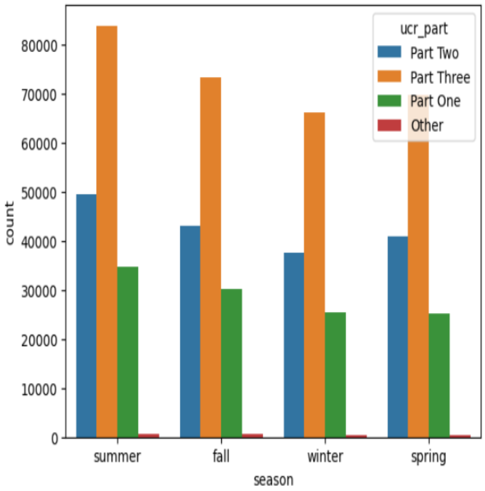
On the right we see Friday has the highest crime rate followed by Wednesday and Thursday because the next day is weekend and maybe people enjoy going out that makes them easy target for criminals and other type of emergencies takes place like minor accidents.

```
<AxesSubplot:xlabel='count', ylabel='day_of_week'>
```



Here we can see that every day of the week the most reported emergencies is UCR part three which are non-lethal emergencies/crime, and we assume that this crime happens out of necessity to survive.

```
<AxesSubplot:xlabel='season', ylabel='count'>
```



From the above graph we can see that summer has the highest emergencies reported, maybe due to people are mostly out of house that makes easy target for theft and robbery and winter has lowest reported emergencies because there's snowfall that makes it hard for the offenders to commit crimes and because people might not be roaming outside in the open due to freezing cold temperature.

We found some more insights but unfortunately not every insight can be included on a webpage. Nevertheless, you can look here.

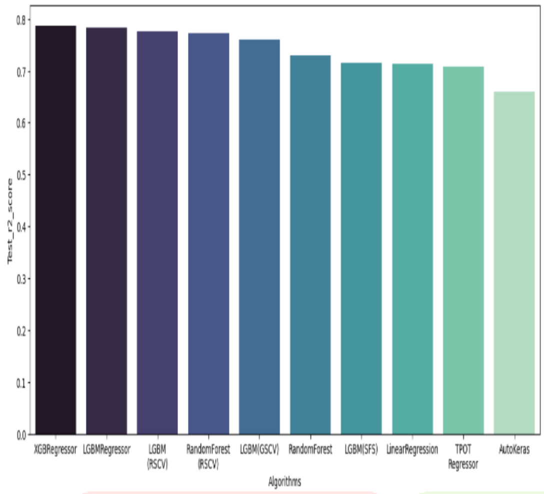
2. Data preprocessing and feature engineering

Now when the EDA is done, we need to curate the model for the modeling. We tried creating more features from the existing data that give more strong patterns i.e., we created the *SEASON* from the *MONTH* column and further created more columns like snowfall(inches), precipitation and some factors from external sources that contribute to the crimes and emergencies like unemployment rate. As for the categorical data like *UCR_PART*, *SEASON*, etc. We performed one-hot encoding and that changes our number of features from 17 to 42 and for datetime data type of column *OCCURRED_ON_DATE* we converted it into ordinal number since we know most of ML algorithms only want numbers to work.

Until now we didn't have any feature with number of emergencies that occurred on each day that we will be predicting. So, we needed to create our target variable. For that we grouped the data according to dates, weekdays, district and count number of emergencies reported each day in each district with the help of the existing feature named *INCIDENT_NUMBER* and left with ~32,000 datapoints.

3. Model Training

After the feature engineering, our training data was ready for model training. We treat this as a regression problem because we are predicting the number of crimes. We further split the training data into train/test with 70/30 ratio. We built model with many algorithms like *SVM, XGBoost, AdaBoost, LinearRegression, LGBMRegressor, RandomForestRegressor, etc.* and also did hyperparameter tuning using *GridSearchCV* and *RandomizedSearchCV* for xgboost model as it was giving the best r2_score among all tried regression algorithms to predict number of crimes that may happen on a particular day in a particular district on a particular day of the week and also we tried autoML techniques like *TPOTRegressor* and *AutoKeras* and did some feature selection. Below is the graphical representation of the applied algorithms performance.



With this proof-of-concept model we built, the best performance score is around 78.8%. Then we proceeded to deploy this model as a cloud endpoint, which can be used to make predictions in client's business application.

Conclusion

As we have seen already that the performance of our best model is around 78.80% which is not considered a very good model and we all know that not every data is for model building but, every data has information hidden that help in targeting the problem. Don't forget our main aim is still to decrease the number of crimes.

So, our model help in fighting the crime rate and play a role in creating a better and safer community.