PREDICTION AND ANALYSIS OF CRIMES IN CITIES

-by VINIT NIRMAL AND SAHIL DESHPANDE

INTRODUCTION

Increased digitization of data is translating into interesting real-world problems.

Crime is one of the most common factors affecting quality of city life.

Utilizing data and finding useful patterns to predict criminal activities can help effective management of law enforcement resources.

Compare two different cities to see if any correlation exists.

DATASET

To get reliable data, used online resources made public by each city [1]. Our dataset consists of two cities, namely Boston and Los Angeles:

- Police district of the crime location
- Time of the day
- Day of the week
- Type of crime

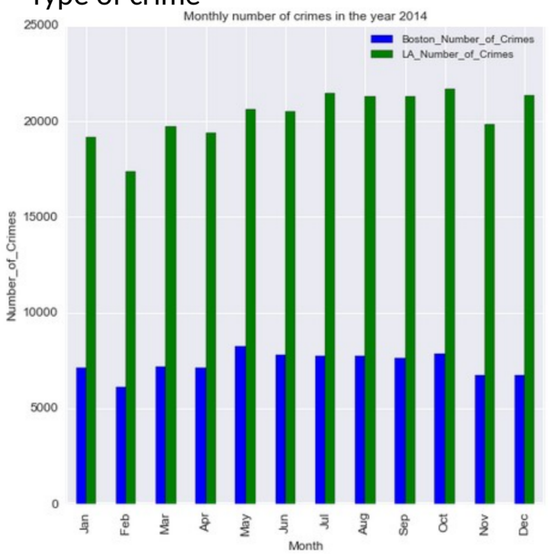

Fig 1. Boston vs Los Angeles comparison

INITIAL ANALYSIS

To find some meaningful pattern in crimes in Boston, we clustered the data in multiple ways and used Linear and Logistic Regression to find various coefficients:

1) CLUSTERING:

Use clustering to determine the similar areas in the city bases on the type of crime that occurred, the time when it occurred, etc.

- K-Means++ Algorithm [2]
- Minimum error for k = 15

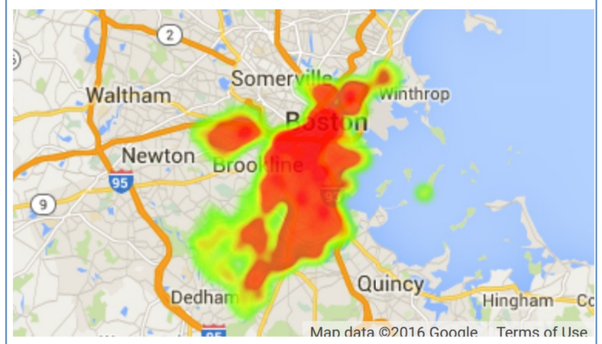

Fig 2. Heatmap of Larceny and Assault Crimes

2) REGRESSION TECHNIQUES:

Find the most important factors for the total number of crimes occurring in an area.

OLS Regression Results

Dep. Variable:	TotalCrimes	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	5.158e+09
Date:	Mon, 25 Apr 2016	Prob (F-statistic):	1.09e-05
Time:	01:05:21	Log-Likelihood:	0.94924
No. Observations:	13	AIC:	22.10
Df Residuals:	1	BIC:	28.88
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Arrest	4.2771	0.086	49.758	0.013	3.185 5.369
Assault	2.5170	0.015	169.775	0.004	2.329 2.705
Burglary	2.6332	0.005	555.342	0.001	2.573 2.693
DrugAlcohol	2.9349	0.008	381.251	0.002	2.837 3.033
Harassment	-7.6406	0.049	-156.432	0.004	-8.261 -7.020
Larcency	3.0397	0.001	4190.941	0.000	3.031 3.049
PublicDisorder	1.3134	0.027	48.773	0.013	0.971 1.656
Sex	2.8042	0.016	174.472	0.004	2.600 3.008
Theft	3.2933	0.005	615.779	0.001	3.225 3.361
Vandalism	4.2427	0.014	313.097	0.002	4.071 4.415
WhiteCollar	3.9988	0.008	471.484	0.001	3.891 4.107
Other	2.9259	0.001	2574.802	0.000	2.911 2.940

Omnibus:	35.803	Durbin-Watson:	1.065
Prob(Omnibus):	0.000	Jarque-Bera (JB):	56.295
Skew:	3.156	Prob(JB):	5.96e-13
Kurtosis:	11.006	Cond. No.	3.67e+03

Fig 3. Linear Regression coefficients

PREDICTION METHODS

Divide data into training, validation and testing data:

- Total 250,000 crime entries for Boston.
- Around 200,000 to model and train a model.
- Remaining 50,000 as prediction pairs
- Random division over entire dataset.
- Predict type of crime for a given location, day and time using two techniques [2].

1) Naïve Bayes Classifier:

- Statistical model that predicts class membership probabilities
- Based on Bayes' theorem.

2) Decision Tree Classifier:

- Supervised learning algorithm to predict class label values.
- Learns decision rules implied by the training data

EVALUATING RESULTS

- Naïve Bayesian Classifier:
  - An accuracy of 51% in Boston crime prediction
  - Reaches 54% for Los Angeles crime prediction.
- Decision Tree Classifier:
  - Reports less prediction accuracy with 42 % for Boston
  - And 43% for Los Angeles.

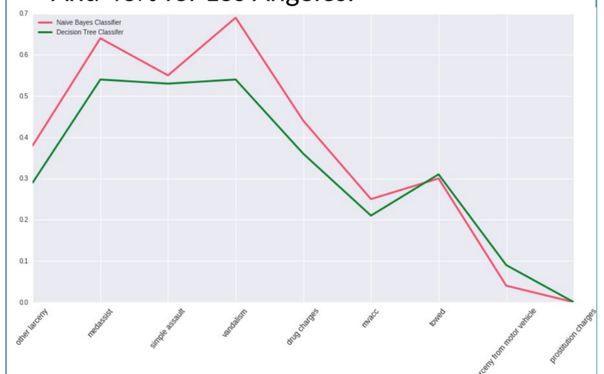

Fig 4.Accuracy per type for both models

ANALYZING TRENDS

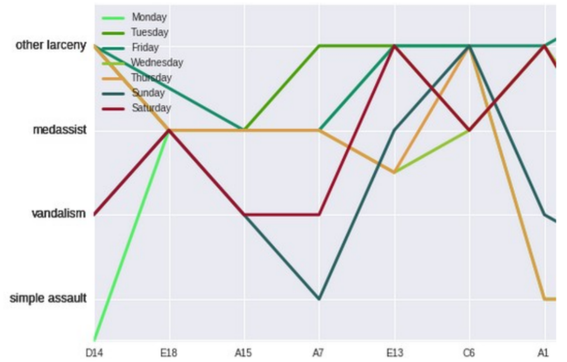

Fig 5. Majority types of crimes per day per district.

CONCLUSION

- Overall, Boston is a safer city than LA.
- Within Boston, assault and larceny crimes form majority of the dataset.
- For prediction model, Naïve Bayesian Classifier performs better and faster compared to Decision Tree Classifiers.

FUTURE WORK

Further analysis of patterns and improved prediction models using additional data on average income, age, gender distribution, locations of police stations, access to pubs and bars and many other social aspects of a city

REFERENCES

[1] US City Census – Crime Datasets

[2] Python scikit and sklearn manuals

ACKNOWLEDGEMENTS

- Prof. George Kollios and Katherine Zhao for explaining concepts and data interpretation.
- City of Boston and Los Angeles data repository