Identification of Criminal Activity Hotspots using Machine Learning to aid in Effective Utilization of Police Patrolling in Cities with High Crime Rates

Ramshankar Yadhunath, Srivenkata Srikanth, Arvind Sudheer, Suja Palaniswamy

Presented by Ramshankar Yadhunath
AGENDA

RELEVANCE AND NEED FOR THE RESEARCH
What and Why is this research important?

PREDICTIVE MODELLING - HOW DOES IT HELP OUR PROBLEM STATEMENT?
A Machine Learning based approach

THE RESEARCH METHODOLOGY
How did we go about our idea?

RESEARCH OUTCOME AND RESULTS
How did our models perform? What was the novelty in our work?

CONCLUSION
Wrapping it Up!
RELEVANCE AND NEED FOR THE RESEARCH

What and Why is this research important?
AN OVERLOOKED PROBLEM

"There can be no sustainable development without peace and no peace without sustainable development"
- United Nations 2030 Agenda for Sustainable Development

- Crime is a major deterrent to a peaceful world
- Several organizations are taking steps to reduce crime and its effects

BUT, WHAT IF THERE IS A FACTOR HIDDEN TO THESE ORGANIZATIONS?

"THE PROBLEM OF POOR POLICE-POPULATION RATIOS"
WORLD (POLICE:POPULATION) RATIOS

United Nations Recommendation = 222 police personnel per population
The Problem Statement

To facilitate effective distribution of police forces in a city among multiple districts based on the extent to which each district is prone to crime at a given hour, in a given day, for a given month.
PREDICTIVE MODELLING

A Machine Learning Based Approach
PREDICTIVE MODELLING IN A NUTSHELL

Historical Data → ML Model → Predictions!

New Data
USING PREDICTIVE MODELLING FOR OUR WORK

**Historical Data**
- Crime Records' Data of past years

**ML Model**

**New Data**

**Predictions!**
- "Probable" Crime Hotspots
- "Probable" Alarm Rates

CSITSS - 2019
21st December, 2019
How did we go about our idea?
AN OVERVIEW OF THE METHODOLOGY

- OSEMN - Commonly used Data Science Methodology
- OSEM - Our variant
- Linear Process

Obtain -> Scrub -> Explore -> Model
OBTAIN DATA

- City of Chicago Data Portal
- Training Sample: Data from 2015 - 2019 (May)
- Test Sample: Data from 2012 - 2014
- Over 11 lakh entries
**SCRUB (PRE-PROCESS) DATA**

- Removing Missing Values: 98.55% entries retained
- Feature Engineering: Decomposing "Date" feature
- Data Aggregation: Grouping Data to count crimes per time point

Obtain > 11,00,000 rows

Scrub ~ 45,000 rows

Explore

Model

CSITSS - 2019
EXPLORE DATA

- Create a "Target" feature
- Label based on "Normal Distribution" and IQR

TABLE I
FIRST 5 ROWS OF crimeDat

<table>
<thead>
<tr>
<th>Month</th>
<th>Day</th>
<th>Hour</th>
<th>District</th>
<th>CrimeCount</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
<td>9</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>10</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>10</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>16</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>13</td>
<td>31</td>
<td>1</td>
</tr>
</tbody>
</table>
EXPLORE DATA

- Create a "Target" feature
- Label based on "Normal Distribution" and IQR
Supervised Classification Problem
Imbalanced Dataset [28:51:21 ratio]
3 Samples used
7 ML algorithms tested
Gradient Boosting Tree worked best
RESEARCH OUTCOME AND INNOVATION

How did our model perform? What was the novelty in our work?
We have evaluated our models based on the following metrics that are common to most ML problems:

- Accuracy
- Precision
- Recall
- F1 Score
- Unweighted Average Recall

But, we also need a "PROBLEM-SPECIFIC" metric to improve the robustness of our work.
A NOVEL METRIC FOR OUR PURPOSE

Region X

Actual Label: High Alarm, High Alarm, High Alarm
Predicted Label: High Alarm, Medium Alarm, Low Alarm

- We must have a model that "minimizes" the situation
- Problem-specific metric: Percentage of misclassifications of "high alarm" regions as "low alarm" regions
- Let's call this metric "HL-mis" in the further slides
Key Considerations while choosing a Model:

- High Accuracy
- High F1 score
- Low HL-mis

Testing Samples:

- Sample 1: 25% of crimeDat (With class imbalance)
- Sample 2: 25% of crimeDat (Without class imbalance - Achieved by oversampling)
- Sample 3: All crime records from 2012-2014
## Model Comparisons - Traditional Metrics

### Table II
Comparing Model Performances

| Machine Learning Model | Sample 1 | Performance | Sample 2 | | Sample 3 | | | | |
|------------------------|----------|-------------|----------| | | | | | |
|                        | Accuracy | FT Score    | UAR      | Accuracy | FT Score | UAR | Accuracy | FT Score     | UAR | |
| Gradient Boosting Tree | 80.50    | 80          | 78.59    | 81.32    | 81       | 81.47| 67.48    | 64            | 73.16 | |
| Random Forest          | 78.35    | 77          | 76.70    | 88.65    | 88       | 88.65| 66.25    | 62            | 67.39 | |
| Decision Tree          | 71.73    | 71          | 71.20    | 86.40    | 86       | 86.39| 65.39    | 61            | 66.39 | |
| K Nearest Neighbour    | 75.89    | 75          | 73.77    | 80.1     | 79       | 80.10| 69.68    | 64            | 67.85 | |
| Support Vector Machine | 58.32    | 43          | -        | 59.25    | 57       | 59.27| 49.23    | 46            | 56.61 | |
| Naive Bayes            | 58.51    | 47          | 48.93    | 59.84    | 56       | 59.84| 47.01    | 44            | 56.98 | |
| Logistic Regression    | 56.13    | 44          | 46.06    | 58.88    | 57       | 58.89| 49.63    | 47            | 55.78 | |
## MODEL COMPARISONS - OUR METRIC

### TABLE III

**Misclassifying “High Alarm” as “Low Alarm”**

<table>
<thead>
<tr>
<th>Machine Learning Model</th>
<th>Percentage of samples that are 2, but wrongly classified as 0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sample 1</td>
</tr>
<tr>
<td>Gradient Boosting Tree</td>
<td>0.04</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.04</td>
</tr>
<tr>
<td>K Nearest Neighbour</td>
<td>0.56</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>6.21</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>6.08</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>5.91</td>
</tr>
</tbody>
</table>
CONCLUSION

Wrapping it Up!
OUR CONTRIBUTIONS

- Our work looks at predictive policing from the angle of "Optimizing low police force" to control crime even in those cities with very high crime rates.
- This work can also be incorporated on a state-level or county-level basis and can be the foundation to more complex police force allocation mechanisms.
- Our paper is based along the notion of using "Data science as a means of promoting social good".
- The new problem specific metric is an effective way to evaluate the robustness of a model that can predict the alarm rate of a region.


THANK YOU