Report: Real-Time Crime Hotspot Forecasting

Introduction

The objective of this project is to forecast crime hotspots in Portland City, Oregon using the CFS dataset provided by the Portland police department. The aim is to develop a model that identifies crime hotspots - high-risk areas for crime in Portland so that there is efficient allocation of resources by law enforcement.

The primary evaluation metrics for this task are the **Prediction Accuracy Index (PAI)** and **Prediction Efficiency Index (PEI)**. These metrics assess the accuracy and efficiency of predicted hotspots for specific crime categories over a two-week timeframe.

Data Exploration and Preprocessing

Dataset Overview

The CFS dataset contains records of police calls for service, including geographic coordinates, timestamps, and crime categories. The dataset spans multiple years, with a focus on the March-May 2017 period for evaluation.

Data Exploration

The CSF data included information about the *CSF Category* , the *CSF Group* under each category, the *location* of the crime in EPSG:2913 Coordinate Reference System, the *census tract,* and the *occurrence date* of the CSF.
The data is read with the help of the pandas library in Python. Initially, it was checked for any missing values. Missing values were only present in the census_tract column.There were 3041 missing values in the census_tract column. Since I didn't use this column for analysis or joining other external data sets due to lack of time, I dropped that entire column instead of the row consisting of missing values for that column.

Exploratory Data Analysis was done to discover key insights into the data and to visualize any pattern present.

- The distributions of the categorical columns like CSF Category and CSF Group as shown in figures in the “Exploratory Data Analysis.ipynb” file indicated that most of the CFS fall under the Street Crime Category and within that under the Disorder group.

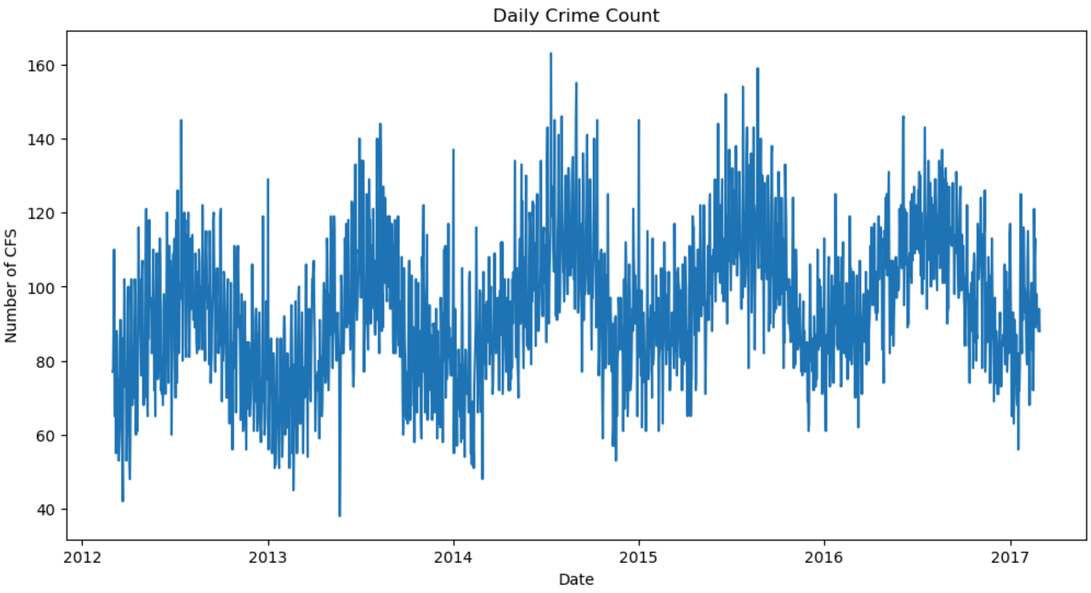
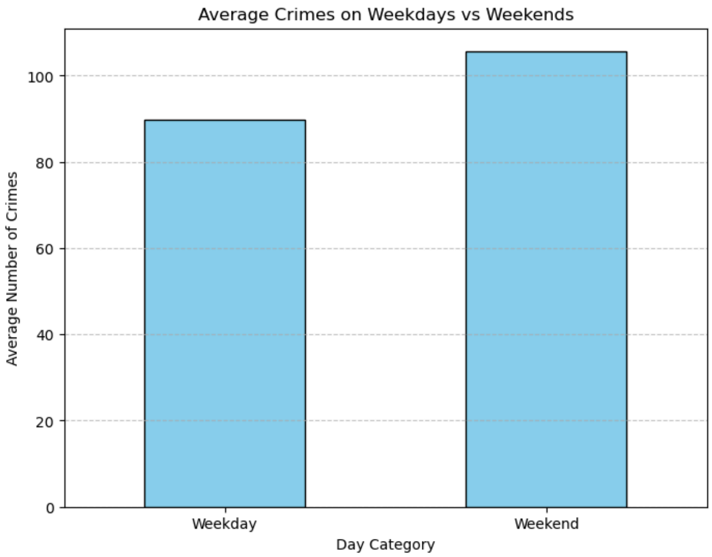


Figure (1)

- From the above Figure(1) (plot of aggregated daily CSF over the 5-year period), it can be seen that there is a clear repetitive pattern that follows a seasonal trend. So, a month can also be a good predictor of crime.
- The plot of the number of crimes on each day of the week indicates that crimes on weekends are comparatively higher than on weekdays. It can be further illustrated with the following Figure (2) depicting average crimes on weekdays vs Weekends.



Figure(2)

- If we were to predict crime daily, it can be said that whether a day is a weekend or a weekday would play a vital role in that.

- Seeing this, I was wondering whether more crimes occur on holidays. So inspecting average crimes on national holidays vs other days revealed that average crime on national holidays (95.63 crimes) is slightly higher but by a small margin (1.29 crimes) over the period of 5 years.

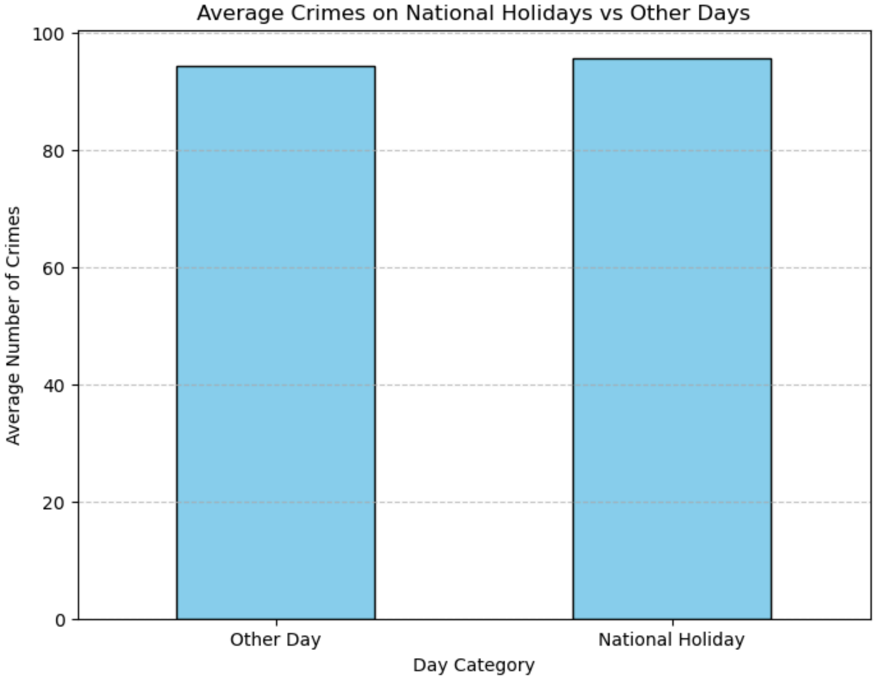


Figure (3)

- From the plots of the location of CFS over the map of Portland (Figure (4), it can be seen that there are some concentrated regions of CFS and also there are regions with no CSF. This spatial feature can be exploited to build a model that consider some regions as hotspot based on the no. of CFS on the region.

- The district with clearly the highest crime count is "district 822"

- CSF data were aggregated over 2-week periods starting from 1st March 2012 to 28th February 2017 for training purposes to predict the CSF for March-May 2017.

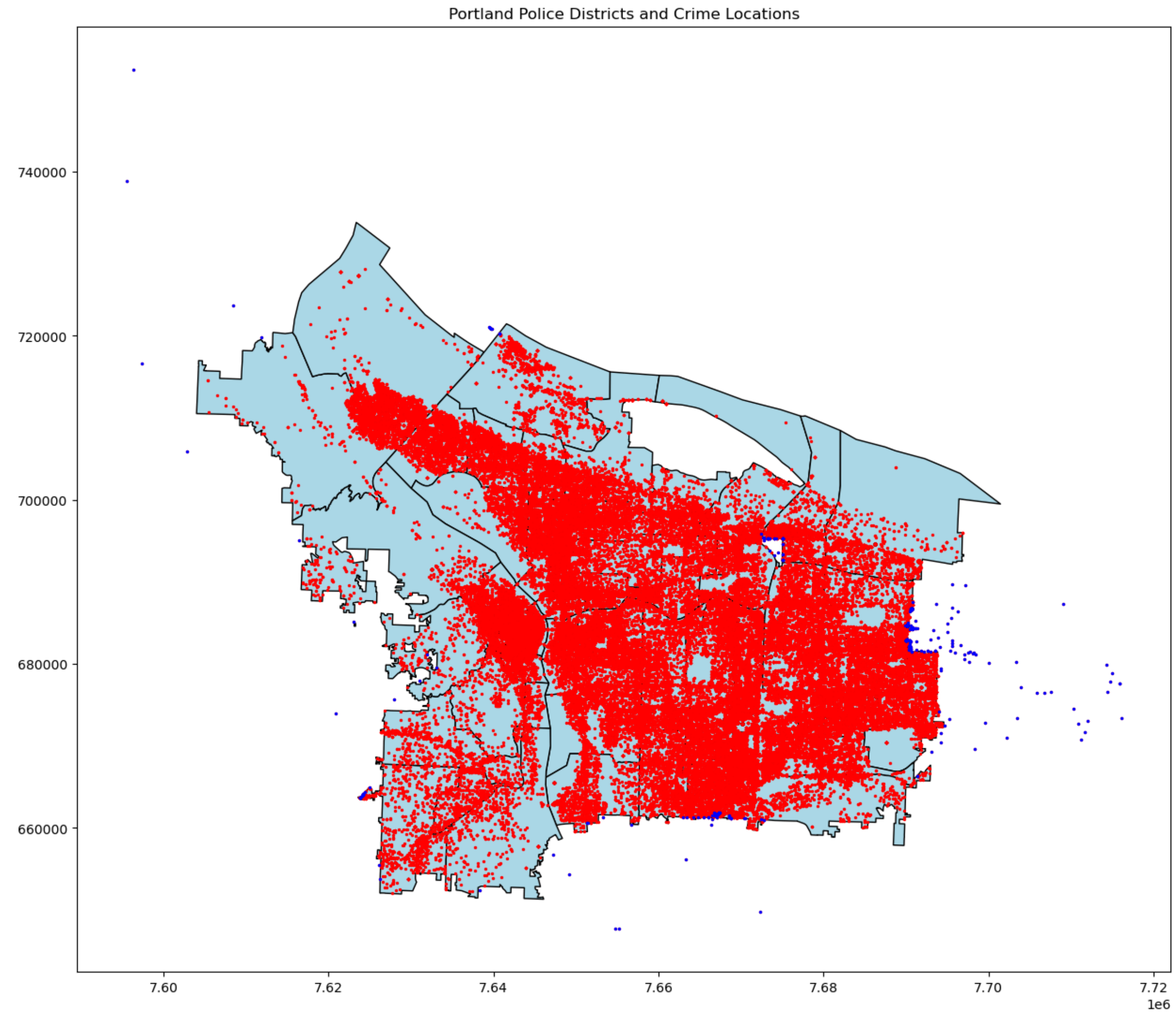


Figure (4): CSF locations on top of Portland city

Spatial Grid Generation

To standardize spatial analysis, I created a uniform grid overlay of 600x600 and 250x250 ft cells.

Model Development

Model Selection

It is well known that RNN models are great for time-series forecasting[4], which is the problem at hand. I used the LSTM-RNN model for predicting the CSF data.

Approach I: Used the total crimes on cells as different features. i.e. If cell_id [5,7,2,8, 10] were identified as potential hotspots, then we have 4 features and the label is also equivalent to the feature-length predicting the no. of crimes on each of those cells at once. It is framed as a regression problem. This required a large no. of units in the final LSTM layer as 351 cells were identified as potential hotspot cells. Also, the performance was not good based on the PAI and PEI. Also, balancing the dataset was difficult in the approach. So this approach was not used for prediction.

Approach II:
Used crime_count in cells for different time steps as a single feature as described in [1]. The problem is framed as a classification task of predicting whether a cell is a hotspot for a particular time stamp. The following description is for this approach.

Data Modeling

The CSF data from February 1st, 2012 to February 28th, 2017 were aggregated in 2-week intervals for each cell in the grid. A sequence length of 26 (365/14) is defined to feed the model with the data for the whole previous year for the next time-step to be predicted.

Features

The models used the following features:

- Historical crime counts in grid cells.
- Spatial features: Historical crime counts of surrounding cells (work in progress)

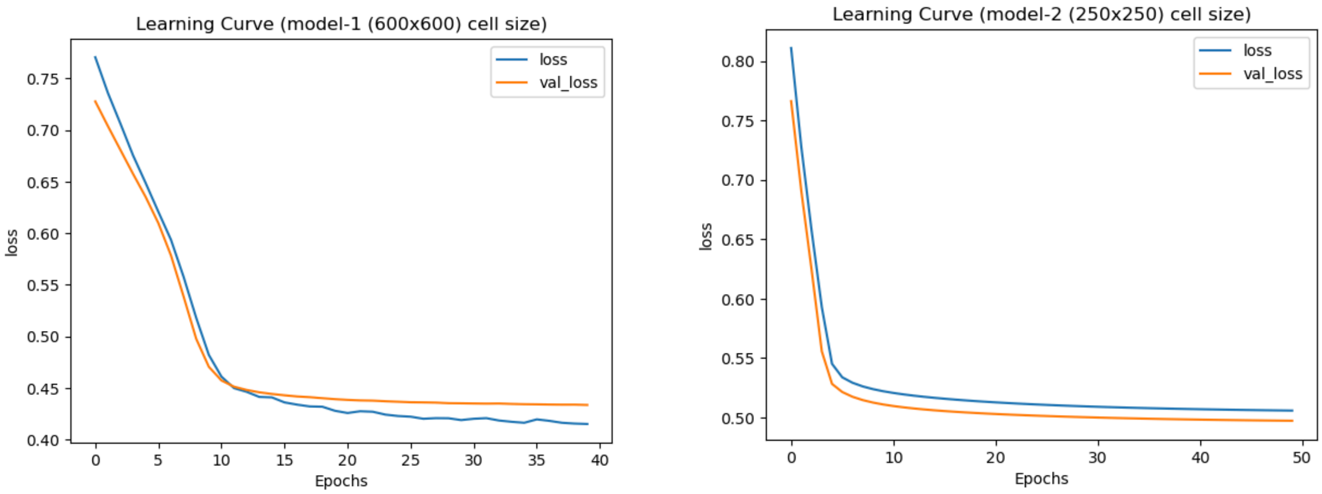
Prediction Process

1. **Hotspot definition**: Defined hotspots as the top 2% of grid cells with the highest crime counts. (threshold = 5 for a cell of 600x600ft, and threshold=3 for cell of 250x250ft)
2. **Balanced Dataset:** There were a large number of negative samples as compared to positive samples (2248). So, to make the training dataset balanced, 2248 random samples were drawn from the large negative samples without replacement.
3. **Validation**: 10% of the data from 2012 to February 2017 was used for validating the model.
4. **Model Testing:** The labels from March 2017 to May 2017 were selected for testing the model which includes the last sequence from the training data as predictors.

Evaluation of Model Performance

As I increased the probability-threshold value for both classification models, I expected the PAI to increase and PEI to go down as raising the threshold decreases Recall. And, as Recall decreases, fewer actual crimes are captured within the predicted hotspots. As expected, the PAI increased, however, in both the models, fortunately, PEI didn’t go down. This may be because the crime data is imbalanced (ie. few locations account for most crimes). As a result, the smaller predicted areas capture a large proportion of actual crimes, maintaining or even increasing PEI, despite lower Recall.

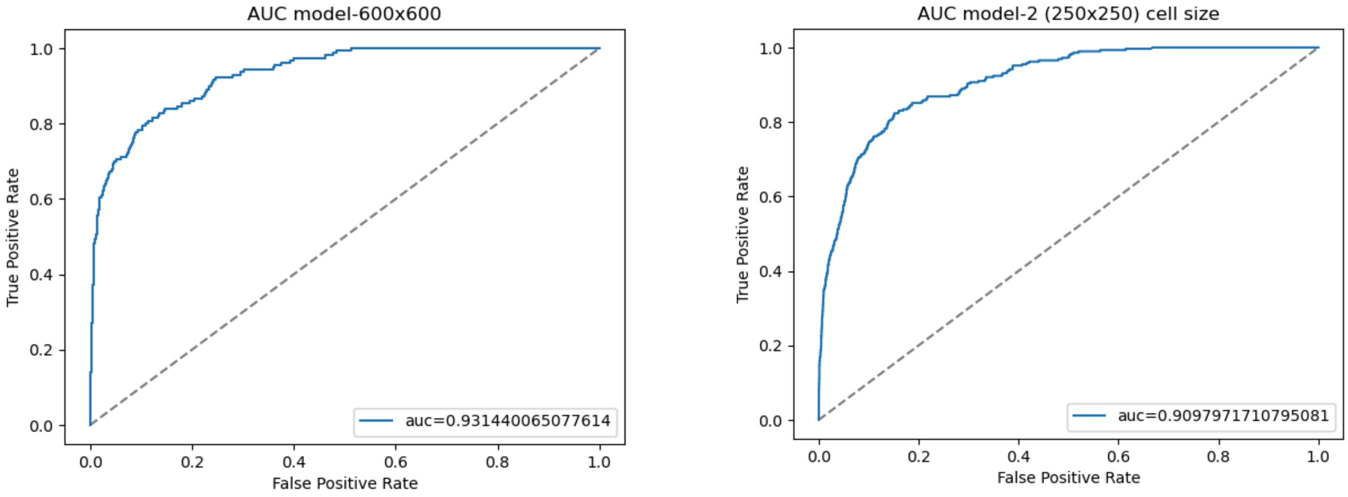
Metrics ↓		**Model 1 (600x600 cell size)**		**Model 2 (250x250 cell size)**	
		Probability Threshold = 0.5	**Probability Threshold = 0.9**	**Probability Threshold = 0.5**	**Probability Threshold = 0.9**
Accuracy		*87.06*	*87.06*	*85.05*	*85.05*
Precision		*30.53*	*73.15*	*13.73*	*45.81*
Recall		*81.41*	*50.64*	*80.97*	*35.99*
f1-score		*44.41*	*59.85*	*23.48*	*40.31*
PAI	**Mar 1 - Mar 14**	*35.96*	*69.02*	*57.17*	*155.05*
	Mar 1- May 31	*23.72*	*43.97*	*36.23*	*88.89*
PEI	**Mar 1 - Mar 14**	*0.78*	*0.82*	*0.54*	*0.75*
	Mar 1- May 31	*0.79*	*0.79*	*0.68*	*0.73*



Learning Curve (model-1 (600x600) cell size)

Learning Curve (model-2 (250x250) cell size)

Figure (5) Learning Curves for Model-1 and Model-2



Figure(6) AUC for Model-1 and Model 2

Challenges and Improvements

Challenges

- **Creating grids over the city boundary.** Small lines exist between the districts even after creating a single polygon from the multi-polygon file provided by the Portland Police Department, which led to the cells being divided even after falling totally inside the city boundary.
- **(Non-excusable):** Lack of time and resources for testing and evaluating different models. (Got the task on Nov 25)

Potential Improvements

- Incorporate demographics features
- Including the crime-count of neighbouring cells (work in progress)
- External spatial and temporal features could be used (like the nearest distance to a police precinct, and sunlight hours in a day) to enrich predictions.

Visualizations

Included plots:

1. Heatmaps of actual vs. predicted hotspots for
   - 600x600 grids in ./Data/visualizations/600/
     i. crime-forecast-grid-mar-01-mar-14-threshold-0.5.png
     ii. crime-forecast-grid-mar-01-mar-14-threshold-0.9.png
     iii. crime-forecast-grid-mar-01-may-31-threshold-0.5.png
     iv. crime-forecast-grid-mar-01-may-31-threshold-0.9.png
   - 250x250 grids in ./Data/visualizations/250/
     i. crime-forecast-grid-mar-01-mar-14-threshold-0.5.png
     ii. crime-forecast-grid-mar-01-mar-14-threshold-0.9.png
     iii. crime-forecast-grid-mar-01-may-31-threshold-0.5.png
     iv. crime-forecast-grid-mar-01-may-31-threshold-0.9.png

   This visualization is made through the **QGis** Software using the shape file filled with information about predicted hotspots.

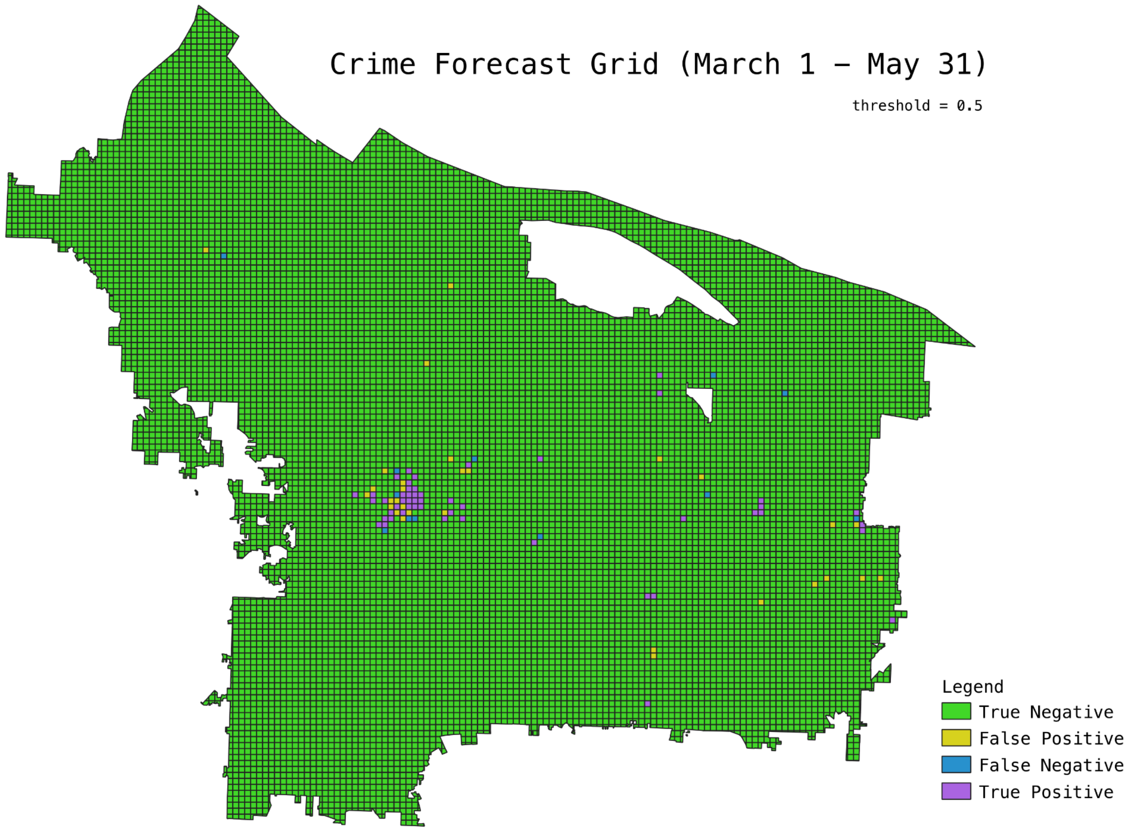


Figure (7) Crime Forecast visualization for 600 x 600 ft cells

2. Visualizations for EDA on *Exploratory Data Analysis.ipynb* notebook.

References

[1], Yong, et al. "Crime hot spot forecasting: A recurrent model with spatial and temporal information." *2017 IEEE International Conference on Big Knowledge (ICBK)*. IEEE, 2017.

[2] “Holidays.” *U.S. Department of Commerce*, 4 Sept. 2024, www.commerce.gov/hr/employees/leave/holidays.

[3] Y. Hua, Z. Zhao, R. Li, X. Chen, Z. Liu, and H. Zhang, "Deep Learning with Long Short-Term Memory for Time Series Prediction," in IEEE Communications Magazine, vol. 57, no. 6, pp. 114-119, June 2019, doi: 10.1109/MCOM.2019.1800155.

[4]“Real-Time Crime Forecasting Challenge Posting.” *National Institute of Justice*, nij.ojp.gov/funding/real-time-crime-forecasting-challenge-posting.