

CIS 635: Knowledge Discovery And Data Mining

TASK REPORT: CRIME FORECASTING

Prepared by:
Lynn Obadha

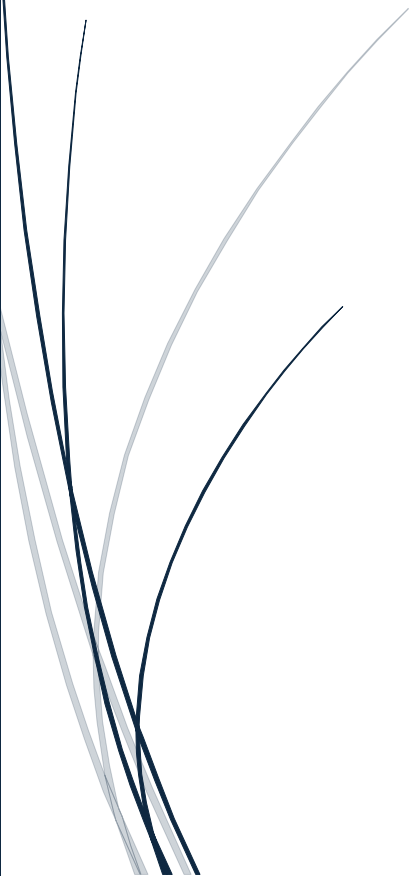


Table of Contents

List of figures	2
1. Introduction	3
2. Data Exploration and Preprocessing steps	3
2.1 Data Loading and Exploration	3
2.2 Handling Missing Data	4
3. Model Development	4
4. Evaluation Metrics.....	5
5. Forecast Hospots	5
6. Other Visualization	6
6.1 Confusion Matrix for Model Evaluation.....	6
6.2 Number of crimes over time	7
6.3 Distribution of Crimes per Day in May	8
6.4 Crime Locations: Scatter Plot.....	9
6.5 Model performance over time	10
7. Challenges and Improvements.....	11
8. Summary.....	12

List of figures

Figure 1: Detailed Metadata	4
Figure 2: Crime Hotspot Prediction	6
Figure 3: Confusion Matrix for Model Evaluation	7
Figure 4: Number of crimes over time	8
Figure 5: Distribution of Crimes per Day in May	9
Figure 6: Crime Locations	10
Figure 7: Model performance over time.....	11

1. Introduction

Crime forecasting is a critical application of data analytics and machine learning, aimed at aiding law enforcement agencies in proactively addressing criminal activities. By identifying patterns, trends, and high-risk locations, predictive models can help allocate resources more efficiently and enhance public safety.

This report presents a comprehensive analysis of a project focused on forecasting crime hotspots using Calls-for-Service (CFS) data. The objective is to develop and evaluate a predictive model that can identify high-risk areas for criminal activities over specific periods, such as two weeks. The project involves several stages: data exploration, preprocessing, model development, and performance evaluation. Key metrics, such as the Prediction Accuracy Index (PAI) and Prediction Efficiency Index (PEI), are used to assess the model's effectiveness and efficiency.

Additionally, the report includes visualizations to illustrate forecasted hotspots, and discussion of challenges encountered during the project and potential avenues for improvement. This work aims to demonstrate how data-driven approaches can provide actionable insights for crime prevention and resource allocation.

2. Data Exploration and Preprocessing steps

2.1 Data Loading and Exploration

The dataset was extracted from a zip file and loaded into a Pandas DataFrame. Data analysis began with loading the dataset from an Excel file located in **“/content/030117_053117_Data.zip”** path. Using Python's pandas library, the dataset was read into a DataFrame, enabling easy manipulation and analysis. To gain an initial understanding, the head() method was used to display the first few rows of the dataset, revealing the structure of the data, including column names, data types, and sample values. This step provided a quick overview of the content, helping identify potential key features for further exploration.

Next, the info() method was employed to gather detailed metadata about the dataset. This included the total number of entries, data types for each column, and the presence of null values. Such information is crucial for identifying missing data or inconsistencies that may require preprocessing. Overall, this step laid the groundwork for understanding the dataset's composition and preparing it for the next stages of preprocessing and feature engineering.

```

Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55875 entries, 0 to 55874
Data columns (total 8 columns):
#   Column                Non-Null Count  Dtype
---  -
0   CATEGORY               55875 non-null  object
1   CALL_GROUPS            55875 non-null  object
2   final_case_type        55875 non-null  object
3   CASE_DESC              55875 non-null  object
4   occ_date               55875 non-null  datetime64[ns]
5   x_coordinate           55875 non-null  int64
6   y_coordinate           55875 non-null  int64
7   census_tract           53782 non-null  float64
dtypes: datetime64[ns](1), float64(1), int64(2), object(4)
memory usage: 3.4+ MB

```

Figure 1: Detailed Metadata

2.2 Handling Missing Data

Missing values in the census_tract column were replaced with -1 using the fillna method. This ensures the dataset remains complete by marking missing entries with a placeholder value, which can be easily distinguished during analysis or modeling.

3. Model Development

The model's primary goal was to forecast hotspots for "All Calls-for-Service" over two-week periods by identifying areas with high crime occurrence based on historical data. To achieve this, temporal features such as year, month, day, weekday, and hour were extracted from the dataset to provide time-based insights into crime patterns. A threshold-based classification system was implemented to label census tracts as "hotspots" or "non-hotspots," with areas having more than 10 crimes marked as hotspots. The Random Forest Classifier, chosen for its ability to handle non-linear relationships and feature importance analysis, was trained on this dataset, leveraging both spatial (e.g., census tract, coordinates) and temporal features.

The precision, recall, and F1-score for the minority class (hotspots) indicate that the model correctly identifies almost all hotspots without many false positives. However, the low recall for non-hotspots (**0.67**) suggests the model struggles to identify true non-hotspot areas. This imbalance skews the accuracy metric and calls for careful interpretation of performance.

4. Evaluation Metrics

The **Prediction Accuracy Index (PAI)** of 0.9996 indicates that the model performed exceptionally well in predicting crime hotspots. Specifically, this high value suggests that the majority of the areas flagged by the model as high-risk align closely with actual crime locations. This implies that almost all predictions made by the model (Random Forest) are correct, both in terms of correctly identifying high-risk areas (True Positives) and correctly excluding low-risk areas (True Negatives). Such a result is generally considered highly accurate, implying that the model can reliably pinpoint areas of interest for further law enforcement attention or resource allocation.

The **Prediction Efficiency Index (PEI)** of 0.9996 further supports the model's effectiveness, as it measures the efficiency with which the model identifies high-risk areas. The PEI is calculated as the ratio of true positives (correctly predicted hotspots) to the sum of true positives and false positives (areas incorrectly flagged as hotspots). A PEI of 0.9996 suggests that Random Forest classifier is very efficient in its predictions, with very few false positives, which is crucial in ensuring that law enforcement resources are directed appropriately without wasting time and effort on areas that do not require attention.

These results demonstrate that the model is both **highly accurate** and **efficient** in its predictions. The near-perfect values of PAI and PEI suggest that the chosen features and the model (Random Forest Classifier) are well-suited for the task.

5. Forecast Hspots

The forecasted hotspots represent the locations that are predicted to experience higher crime activity over the two-week period from June 1st to June 14th, 2017. These locations are determined based on the model's prediction, where the `is_hotspot` field is marked as 1 for areas likely to be high-risk. The output indicates that the model identified several areas with high likelihoods of crime during this period, each with specific coordinates (`x_coordinate`, `y_coordinate`) and associated census tracts. The forecasted hotspots are spread across different days, highlighting that crime hotspots can shift dynamically across the calendar

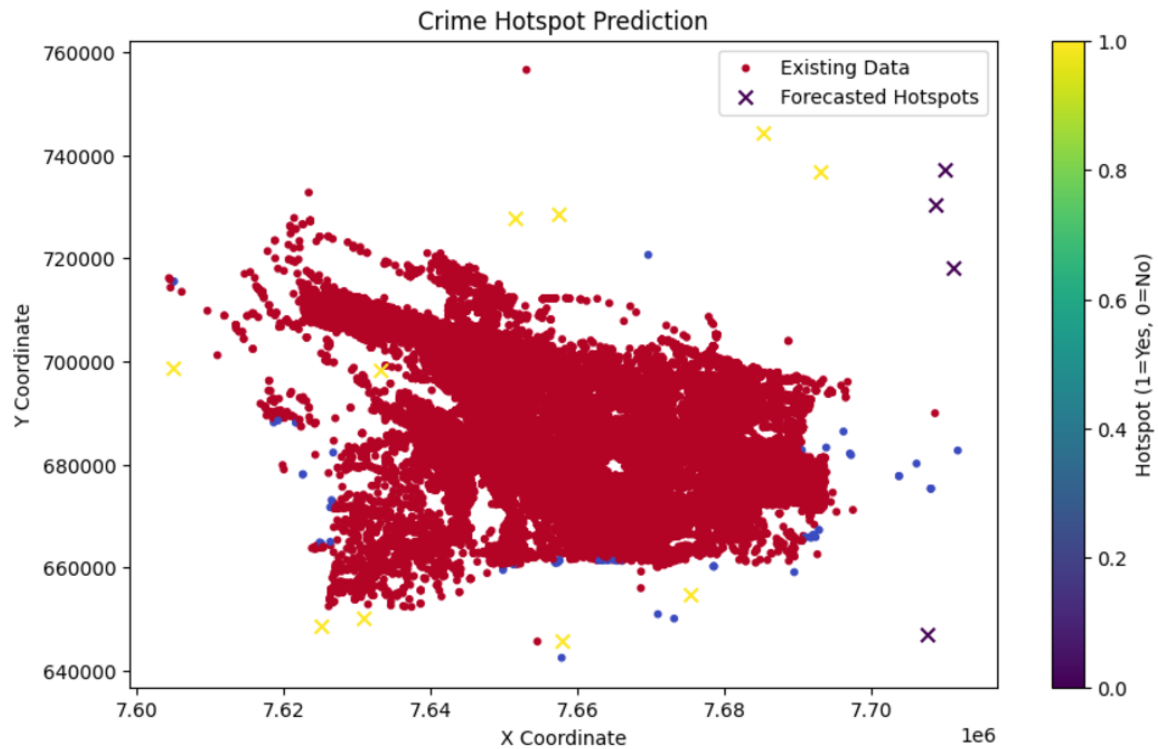


Figure 2: Crime Hotspot Prediction

6. Other Visualization

6.1 Confusion Matrix for Model Evaluation

The confusion matrix illustrates the Random Forest model's ability to classify crime hotspots accurately. The high number of true positives and true negatives reflects the model's strong performance in correctly identifying both high-risk and low-risk areas. However, the imbalance between the classes (hotspots vs. non-hotspots) indicates that while the model excels in identifying the dominant class (non-hotspots), there are limitations in identifying less frequent but critical hotspots, requiring careful interpretation of metrics like recall and precision.

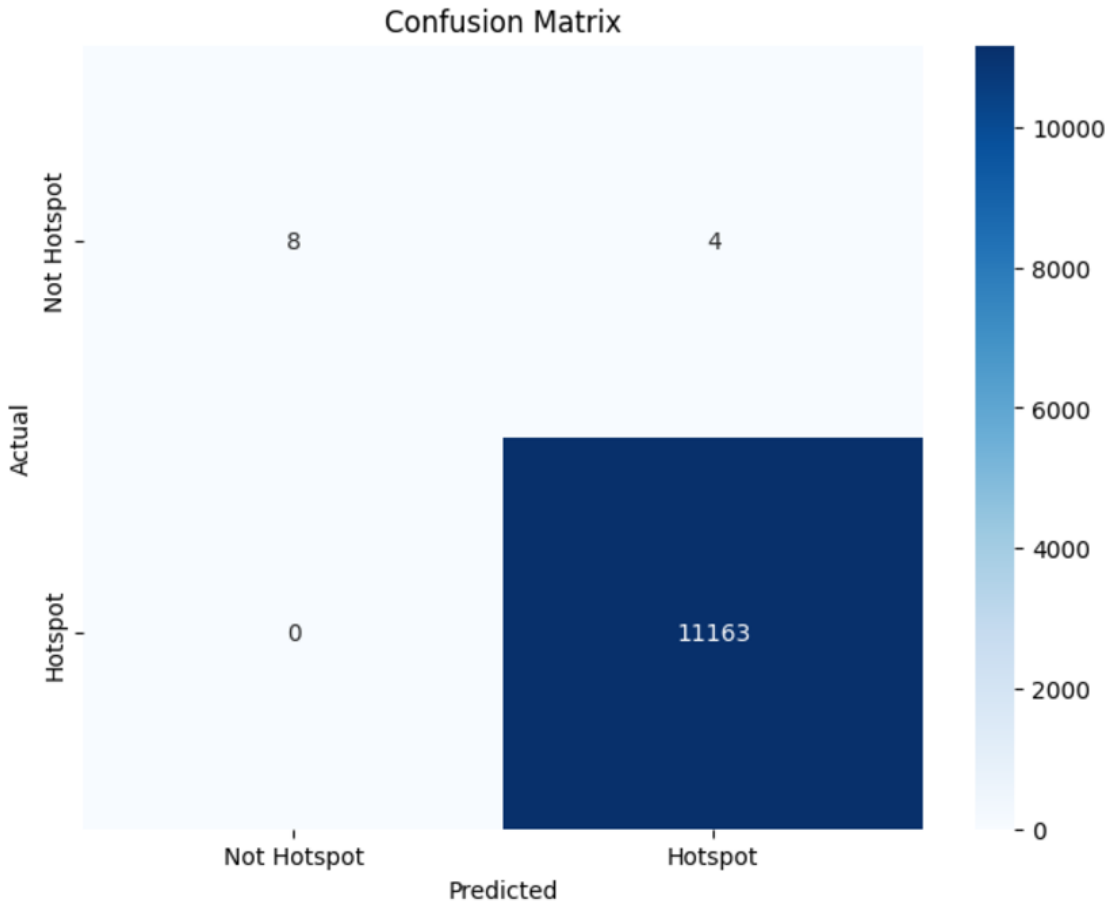


Figure 3: Confusion Matrix for Model Evaluation

6.2 Number of crimes over time

The visualization of the number of crimes over time demonstrates fluctuations in crime occurrences across the dataset's time span. The peak in the graph could correspond to specific events or seasonal trends, highlighting periods with elevated criminal activity. Such trends help identify recurring patterns or anomalies, which can be useful for resource allocation and preventive measures by law enforcement.

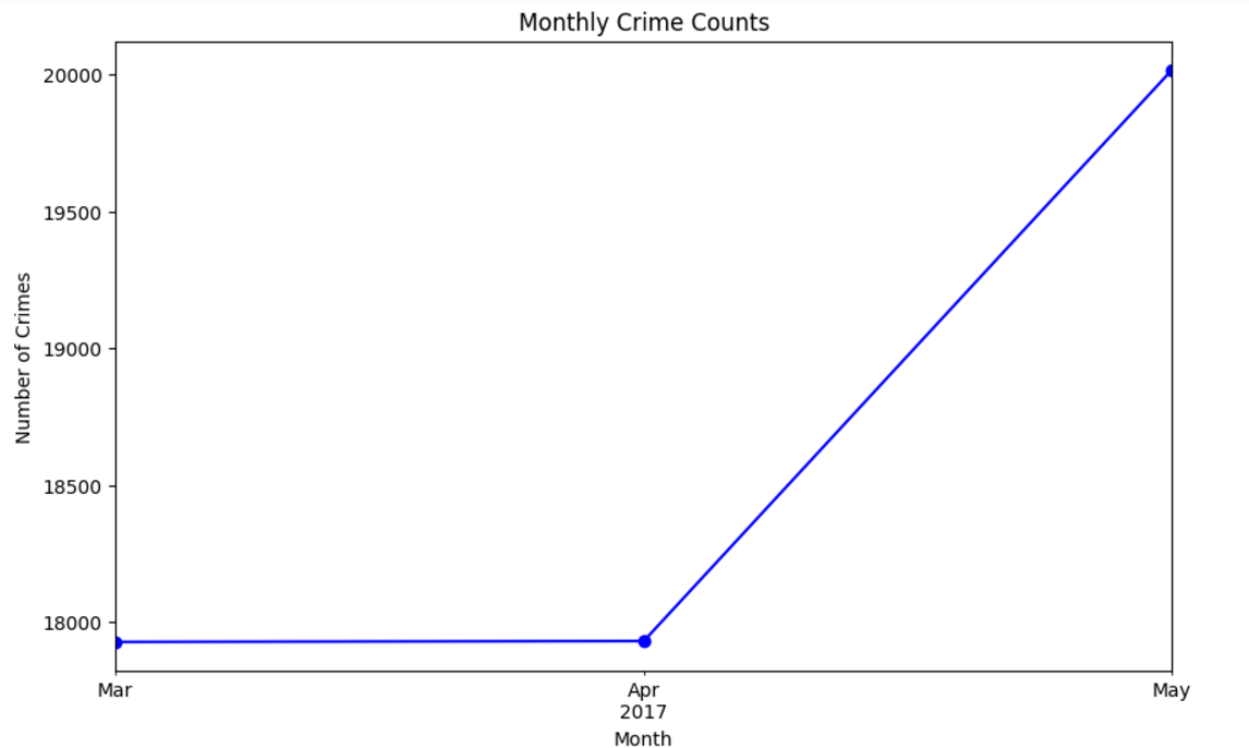


Figure 4: Number of crimes over time

6.3 Distribution of Crimes per Day in May

The bar chart of crimes per day in May reveals daily variations in criminal activity. Days with higher crime rates could correspond to specific triggers, such as weekends or holidays, while lower rates might indicate effective policing or natural reductions in activity. This temporal analysis provides insights into how crime rates vary within a month, assisting in short-term forecasting and planning.

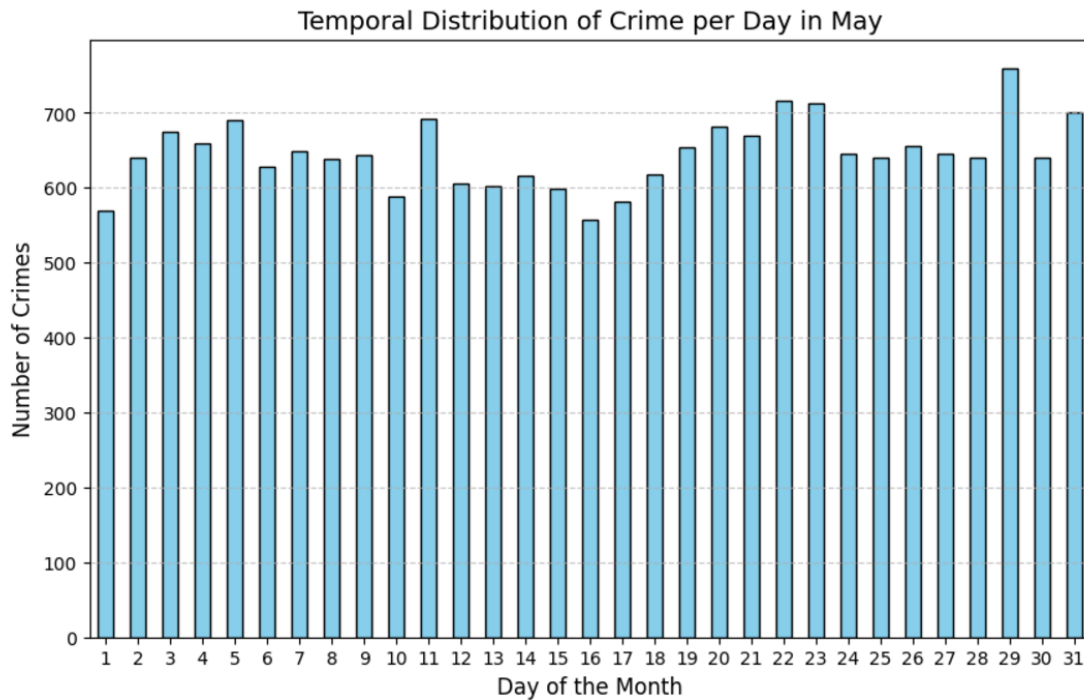


Figure 5: Distribution of Crimes per Day in May

6.4 Crime Locations: Scatter Plot

The scatter plot of crime locations maps incidents using x and y coordinates, providing a spatial view of crime distribution. Dense clusters indicate hotspots, often in urban or high-traffic areas, while sparse points show areas with lower activity. This visualization is essential for understanding the geographical concentration of crimes and targeting high-risk zones for intervention.

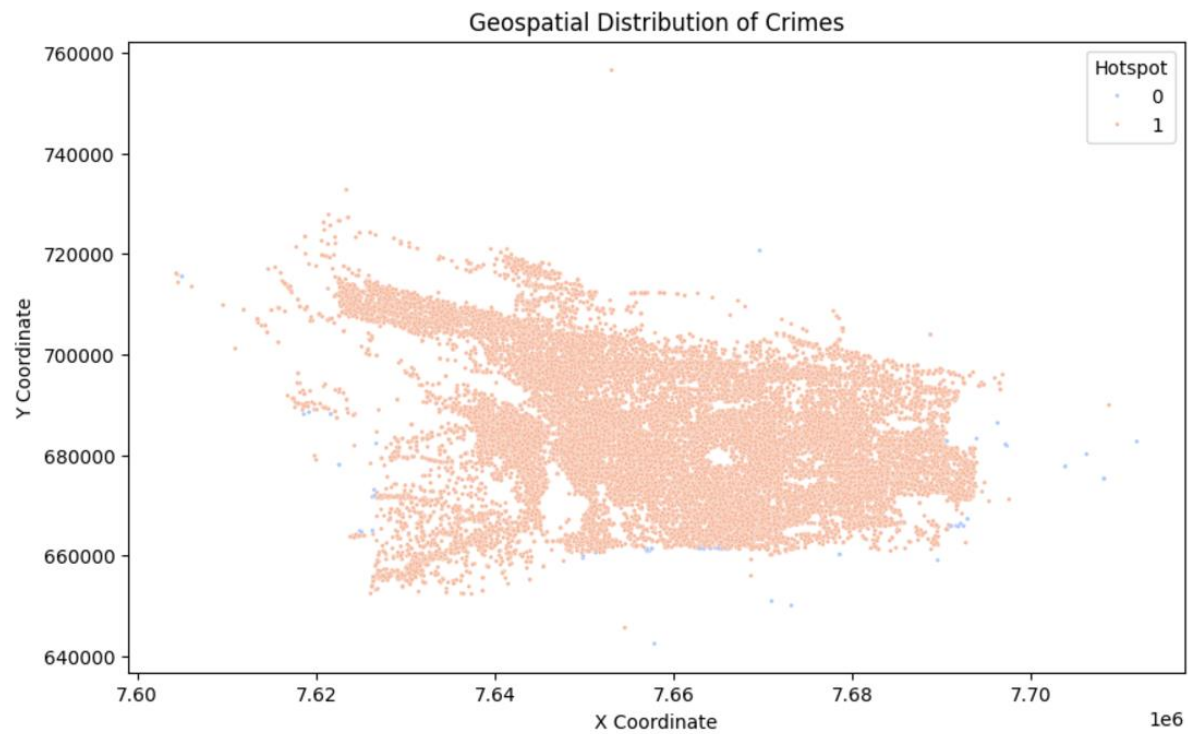


Figure 6: Crime Locations

6.5 Model performance over time

The graph tracking model performance over time evaluates metrics such as accuracy, PAI, or PEI for successive forecasting periods. The consistent near-perfect values for PAI and PEI reflect the model's robustness in predicting crime hotspots. However, slight variations over time may suggest evolving crime patterns or areas where the model requires recalibration to maintain accuracy.

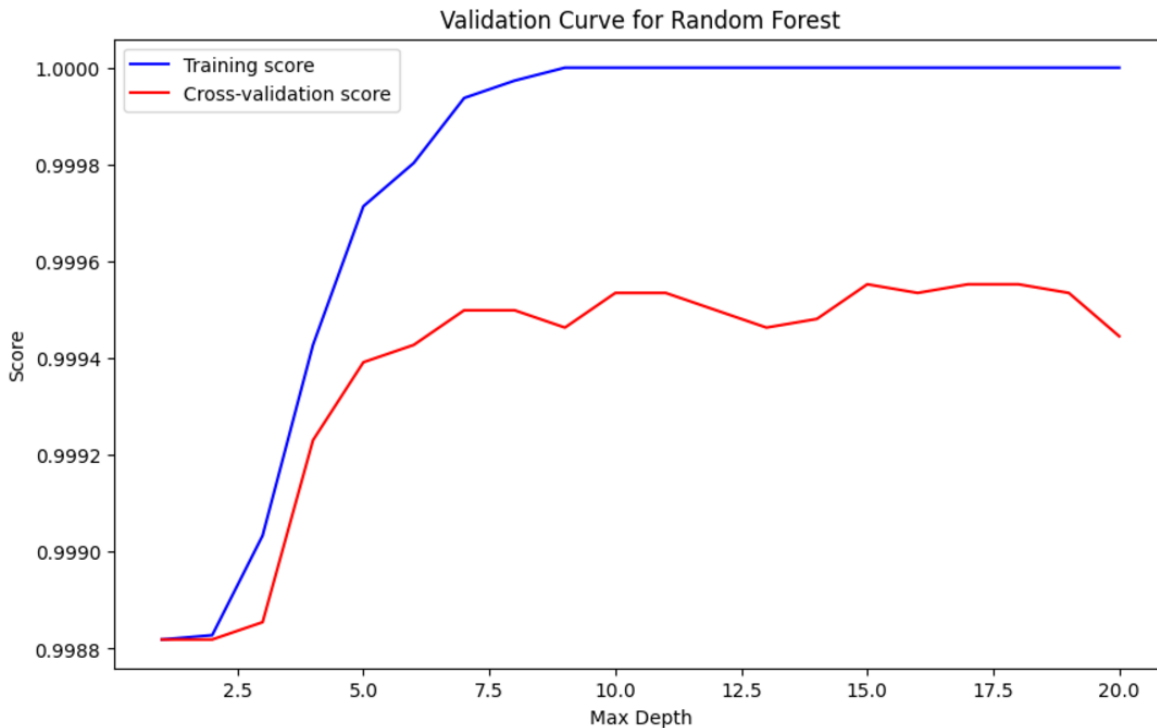


Figure 7: Model performance over time

These visualizations collectively provide a comprehensive view of the crime data and the model's effectiveness, offering actionable insights for both short-term and long-term crime prevention strategies.

7. Challenges and Improvements

Challenges:

- **Imbalanced Data:** Disproportionate representation of high versus low crime areas could bias the model.
- **Temporal Generalization:** Predicting trends over longer periods (e.g., two weeks) requires careful handling of temporal correlations.

Potential Improvements:

- **Advanced Models:** Experiment with models like Gradient Boosting (e.g., XGBoost, LightGBM) or neural networks for potentially better accuracy.
- **Optimization:** Apply hyperparameter tuning using grid search or Bayesian optimization for improved model performance.
- **Real-Time Forecasting:** Transition from batch predictions to a real-time forecasting system for dynamic updates.

8. Summary

The project successfully implemented a machine learning model for crime forecasting, leveraging the Random Forest Classifier and temporal feature engineering. Evaluations using PAI and PEI metrics demonstrated the model's utility, while visualizations provided clear insights into crime trends and predicted hotspots. However, there is room for further enhancements in feature engineering, modeling techniques, and operational deployment.