"Toronto Crime Data Intelligence"

Final Report

Capstone Project

(BIA-5450)

Submitted to:

Dr. Samer Al-Obaidi

Professor,

Humber College

Submitted by:

Prappan Batra (N01579150)

Dhairya Dangi (N01580705)

Param Panchal (N01579822)

Vrajkumar Patel (N01581006)

Pravina Prajapati (N01579926)

Submitted On:

August 16, 2024

Table of Contents

Executive Summary

The Toronto Major Crime Indicators Analysis Project, spearheaded by the Toronto Police Service, is a data-driven initiative aimed at enhancing public safety through a deep understanding of crime trends across Toronto from 2019 to 2024. This project employs a combination of historical crime data analysis, advanced data visualization techniques, and machine learning models to uncover significant patterns in crime occurrences, with a focus on major crime indicators like assault, robbery, and burglary.

The analysis utilized various machine learning models, including K-Nearest Neighbors (KNN) and Random Forest classifiers, to predict crime trends and identify potential hotspots. The Random Forest model emerged as the most effective, offering superior predictive accuracy, which supports strategic resource allocation and more targeted law enforcement interventions. In addition to predictive modeling, time series forecasting techniques such as Exponential Smoothing and LSTM were applied to project future crime rates, providing valuable insights for proactive policing strategies.

Key findings from this project include the identification of high-risk neighborhoods and temporal patterns in criminal activities, such as spikes during specific times of the year. These insights are crucial for the Toronto Police Service in optimizing patrol schedules, deploying resources efficiently, and engaging with the community to prevent crime.

The project's success is attributed to a comprehensive approach that integrates data science, machine learning, and domain expertise, offering a robust framework for continuous crime monitoring and prevention efforts in Toronto. Through detailed reporting and collaboration with community stakeholders, the project underscores the importance of data-driven decision-making in enhancing urban safety and fostering a safer environment for all residents.

1. Introduction

The Toronto Major Crime Indicators Analysis Project represents a meticulous investigative effort by the Toronto Police Service to comprehensively understand and address crime dynamics within the city. Since its inception in 2014, the project has entailed the systematic collection and categorization of data pertaining to significant criminal incidents, encompassing offenses such as assaults, burglaries, vehicle thefts, robberies, and high-value thefts. The overarching objective is to furnish stakeholders with actionable insights aimed at enhancing public safety and awareness. Central to this endeavor is the preservation of data integrity and the protection of individual privacy, achieved through stringent anonymization protocols. Leveraging advanced analytical techniques and geospatial visualization tools, the project endeavors to discern spatial-temporal patterns of criminal activity, thereby informing targeted interventions and resource allocation strategies. As a testament to its commitment to evidence-based policing and community engagement, the project serves as a cornerstone in the ongoing pursuit of urban security and well-being.

Project Importance

a) Enhance Public Safety by deploying resources effectively.
b) Optimize Resource Allocation for cost savings and operational efficiency.
c) Foster Community Engagement through data-driven crime prevention.
d) Work towards Crime Reduction and improved quality of life in affected areas.

2. Business Problem

Business Problem: Due to a lack of thorough information about crime patterns and trends across various neighborhoods and types of establishments, law enforcement agencies in Toronto are having difficulty in the effective allocation of resources and putting crime prevention initiatives into practice.

Business Requirements:

a) **Data Collection and Integration**:
   - Collect historical crime data from various sources, including offense types, occurrence dates, and geographic locations.
   - Integrate data from different divisions and precincts into a centralized database.

b) **Crime Pattern Analysis**:
   - Identify trends in crime occurrences, focusing on assault incidents.
   - Analyze seasonal variations, day-of-week patterns, and time-of-day trends.

c) **Spatial Investigation**:
   - Explore crime distribution across neighborhoods and premises types.
   - Identify high-crime areas (hotspots) and low-crime areas.

d) **Resource Allocation and Deployment**:
   - Develop predictive models to forecast crime hotspots.
   - Allocate law enforcement resources strategically based on predicted crime patterns.

e) **Reporting and Visualization**:
   - Create dashboards and reports for law enforcement officials.
   - Visualize crime data on maps to aid decision-making.

f) **Community Engagement**:
   - Share crime statistics with community stakeholders.
   - Collaborate with community organizations for crime prevention initiatives.

g) **Compliance and Privacy**:
   - Ensure compliance with data privacy regulations.
   - Protect sensitive information related to victims and witnesses.

3. Analytics Questions

a) What are the key crime trends and patterns in Toronto over the past few years?
Finding general patterns in Toronto's crime rates over the past few years, including rises or falls in particular categories of crimes, is the goal of this inquiry. We can learn more about how crime dynamics have changed over time and pinpoint probable causes by examining these patterns.

b) What are the seasonal variations and day-of-week patterns in crime occurrences?
Understanding the temporal patterns in crime, such as how crime rates vary on particular days of the week or fluctuate with the seasons, is the main goal of this question. This research can assist determine whether specific times are more likely to see criminal activity and can help guide focused preventive actions.

c) Which specific Premise in Toronto experiences the highest incidence of crime?
This question seeks to pinpoint the location or type of premises (e.g., residential areas, commercial properties) in Toronto where crimes are most frequently reported. Identifying these hotspots is crucial for law enforcement to allocate resources more effectively and for public awareness.

d) Which year experiences the lowest number of crimes and highest number of crimes?
This question aims to compare crime data across different years to identify the years with the lowest and highest crime rates. Understanding these fluctuations can help in assessing the effectiveness of crime prevention strategies and in recognizing external factors that may have influenced crime rates.

e) What is the distribution of crimes depending on offense type?
This question analyzes how different types of crimes are distributed across the dataset. By examining the prevalence of various offense types, we can better understand the composition of crime in Toronto, which is essential for developing specialized interventions and policies.

4. Scope Statement

The project scope includes the following key activities:

a) **Data Collection and Integration:** Collecting historical crime data from the public safety data portal of TPS.
b) **Data Analysis:** Analyzing historical crime trends, including seasonal variations and spatial distribution patterns across neighborhoods and Offense types.
c) **Descriptive Analysis:** Understand patterns and classification of crimes.
d) **Reporting and Visualization:** Creating reports, dashboards, and visualizations to communicate insights effectively to stakeholders.
e) **Community Engagement**: Collaborating with community stakeholders and organizations to share crime statistics, insights, and prevention strategies.

5. Data Source / Key Data Entities and Flows

Data Source:

The data used in this project originated from the Toronto Police Service Public Safety Data Portal. The dataset specifically pertains to Major Crime Indicators (MCI) and is openly available to the public. This data was retrieved from the Toronto Police Service's data portal, which can be accessed via (Toronto Police Service).

Data Source Format:

The data is provided in the form of a CSV file and contains detailed information about various crimes reported in Toronto. The CSV file format ensures ease of access and manipulation for analysis purposes.

The primary business data entities relevant to the analysis of Major Crime Indicators are:

a) Offence Information:

- **EVENT_UNIQUE_ID:** A unique identifier for each offence.
- **OFFENCE:** The title of the offence.
- **UCR_CODE and UCR_EXT:** Uniform Crime Reporting codes used to categorize the type of offence.

b) Timestamp Information:

- **REPORT_DATE, OCC_DATE:** Dates when the offence was reported and occurred.
- **REPORT_YEAR, REPORT_MONTH, REPORT_DAY:** The year, month, and day the offence was reported.
- **OCC_YEAR, OCC_MONTH, OCC_DAY:** The year, month, and day the offence occurred.
- **REPORT_DOW, REPORT_HOUR, OCC_DOW, OCC_HOUR:** Day of the week and hour the offence was reported and occurred.

c) Location Information:

- **DIVISION:** The police division where the offence occurred.
- **LOCATION_TYPE:** The type of location where the offence occurred.
- **PREMISES_TYPE:** The type of premises where the offence occurred.
- **HOOD_158, HOOD_140:** Identifiers for neighbourhoods based on two different neighbourhood structures.
- **LONG_WGS84, LAT_WGS84:** Longitude and latitude coordinates of the offence location.

As we have all the required variables in the same table, we have not created any Fact and Dimensional tables. For our analysis purposes, a flat CSV file containing all the required attributes is more reliable than individual dimensional CSV files.

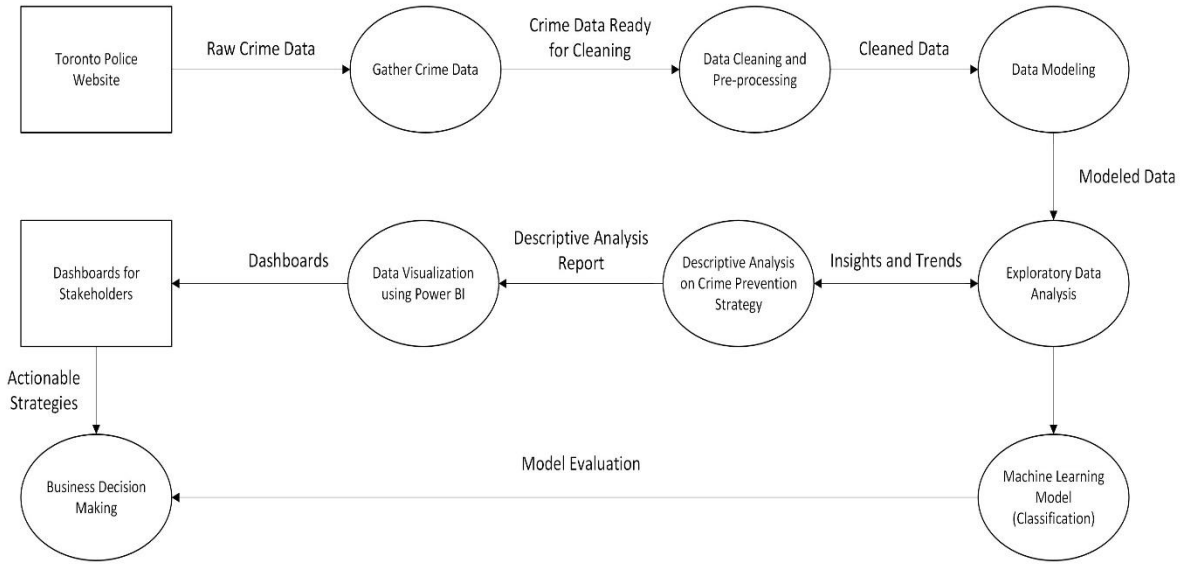
Systems of Record:

The existing systems of record for each data entity are maintained by the Toronto Police Service. These systems include:

- **Crime Reporting Systems:** These systems capture and manage all crime reports, including details about the offence, the date and time of reporting, and the occurrence.
- **Geographical Information Systems (GIS):** These systems manage the spatial data, including longitude and latitude coordinates, and neighbourhood information.
- **Timestamp Databases:** These databases handle temporal data such as dates, days of the week, and hours related to the offences.

Process Data Diagram:

Below detailed graphical representation that integrates the business processes and data flows involved in the analysis of Toronto crime data. This diagram will offer a clear overview of how data moves through the system and how different processes interact with the data.



(Fig 5.1 Process Data Diagram)

6. Data Manipulation Process and Data Output

Filtering Required Data:

The project prioritizes recency; therefore, data from 2019 to 2024 have been selected for analysis. Consequently, data from 2014 to 2018 have been excluded. The next section will outline the cleaning steps for the chosen data.

Dropping Null Values:

To maintain the dataset's integrity and completeness, we identified and removed missing values. Using the isnull() method, we created a Boolean DataFrame to locate missing values, and then used the sum() method to count them.

After determining the extent of missing data, we applied the dropna() method to remove rows with any missing values. This ensured the dataset was complete and ready for accurate analysis, preventing issues like biased results or statistical errors.

Dropping Unnecessary Records:

Neighbourhood and coordinate information (latitude and longitude) will appear to be Not Specified Area (NSA) and (0,0), respectively, if any of the following conditions are met: (1) Division is NSA OR (2) Originating X/Y values are 0 OR (3) Originating X/Y values are outside the City of Toronto.

If an event occurs within 5,000 meters outside the City of Toronto, it is snapped to an intersection and will have coordinates. Neighbourhood values for these events would be NSA.

We have dropped all the records with NSA values for the neighbourhood columns. Also, we have dropped all the records having latitude or longitude values ‘0’. For analysis, these values seemed to deviate the results from the real analysis, and these records constituted approximately 2% of the total data, so we dropped them.

Data Storage:

The data for the Toronto Major Crime Indicators Analysis Project originates from the Toronto Police Service website and is initially stored locally on a PC for analysis. Using Jupyter Notebook, data cleaning and preprocessing are performed to ensure data integrity and consistency. The cleaned dataset is then exported as a CSV file, which remains stored locally on the PC. Subsequently, analytical insights, including dashboards and reports, will be generated using Power BI, serving as the primary platform for visualizing and presenting findings. These local storage locations facilitate efficient data manipulation and analysis while ensuring compliance with data privacy and security protocols.

Data Output:

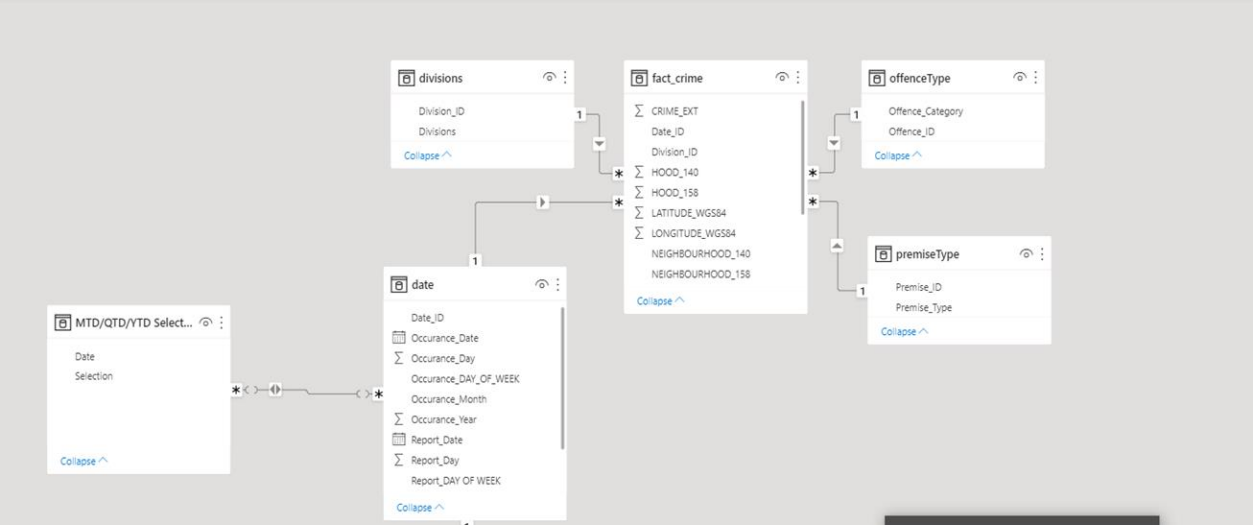
In order to prepare the data for additional analysis, extensive data modification procedures are used throughout the Toronto Major Crime Indicators Analysis Project's data output phase. The dataset will be refined and organised using a variety of algorithms and programming techniques after it has been cleaned and pre-processed in Jupyter Notebook. To obtain valuable insights, this may involve statistical analysis, feature engineering, and the use of machine learning techniques like classification. After the data has been analysed, Power BI will be used to produce comprehensive reports, interactive dashboards, and visualisations. Stakeholders will find these outputs crucial as they offer practical insights into Toronto's crime trends, patterns, and spatiotemporal dynamics.

7. New Solution Design and it's Fit into the Existing IT Architecture

Two design solutions were implemented: **(1)** Data Visualization and **(2)** Machine Learning Models using KNN and Random Forest classifiers.

Using Visualization Approach

Data Model: Creating a snowflake schema for crime data involves centralizing quantitative measures in a fact table, surrounded by dimension tables detailing offense types, divisions, premises types, and dates. Below is the snowflake schema for the crime dashboard:



(Fig 7.1: Data Model – Snowflake Schema)

The Power BI data model for analyzing crime data consists of several interconnected tables. The central fact table, **fact_crime**, holds core crime data, including fields such as NEIGHBOURHOOD_140, NEIGHBOURHOOD_158, OBJECTID, Offence_ID, Premise_ID, UCR_CODE, and coordinates (X, Y). This table is linked to several dimension tables to enrich the data analysis. The **date** table provides date-related information for filtering, with fields like Date_ID, Occurrence_Date, Occurrence_Month, Occurrence_Year, and Report_Date. The **divisions** table contains division-related information with fields such as Division_ID and Divisions. The **offenceType** table classifies types of offences, and the **premiseType** table categorizes different premises where crimes occurred. An additional table, **MTD/QTD/YTD Selection**, seems to be used for selecting different time frames for analysis. These relationships facilitate comprehensive crime data analysis and visualization in Power BI.

Dashboard:

(Fig 7.2: Dashboard)

(Fig 7.3: Time Dashboard)

Below is a detailed description of the Dashboard:

a) **Crime Distribution by Neighborhood**

Visual: Matrix

Purpose: A matrix was used to analyze crime distribution by year, month, and type, with neighborhoods on one axis and time periods on the other. Each cell displays crime frequencies, helping to identify high-crime areas and trends for better resource allocation and crime prevention.

b) **Crime Distribution by Division**

Visual: Matrix

Purpose: It involves analyzing crime data by segments within a broader area to understand how crime rates vary across these divisions. This method reveals patterns and trends, aiding strategic decisions on patrolling.

c) **Crime by year and month**

Visual: Clustered column chart

Purpose: Tracking crime data by specific years and months reveals trends, patterns, and fluctuations, offering insights into seasonal variations and long-term changes in crime rates.

d) **Crime by various Premises**

Visual: Bar chart

Purpose: It involves examining crime data based on different types of locations where crimes occur. This includes outside areas such as streets and parks, apartments in multi-unit residential buildings, commercial properties like shops and offices, etc.

e) **Interactive Statistics Cards**

Purpose: To enhance crime data analysis, developed interactive statistics cards that summarize key crime metrics, including the total number of crimes and a detailed breakdown by crime types, such as robbery, assault, and theft.

f) **Dropdown Filters**

Purpose: To enable detailed and customizable crime data analysis, and integrated dropdown filters into the system. These filters allow users to refine data by time periods, premises types, offense types, and geographical locations, facilitating focused analysis and identification of trends and patterns.

g) Time dashboard for Crime Analysis

At the top, there are buttons labeled "MTD" (Month-to-Date), "QTD" (Quarter-to-Date), and "YTD" (Year-to-Date) for filtering the data by different time periods. Additionally, dropdown menus allow users to filter the data by Year, Month, Day, and Day of the Week.

The line chart in the center shows the "Number of Crimes by Month," depicting the monthly trend of crimes. The chart indicates a peak in January and then a decline in the following months, with minor fluctuations.

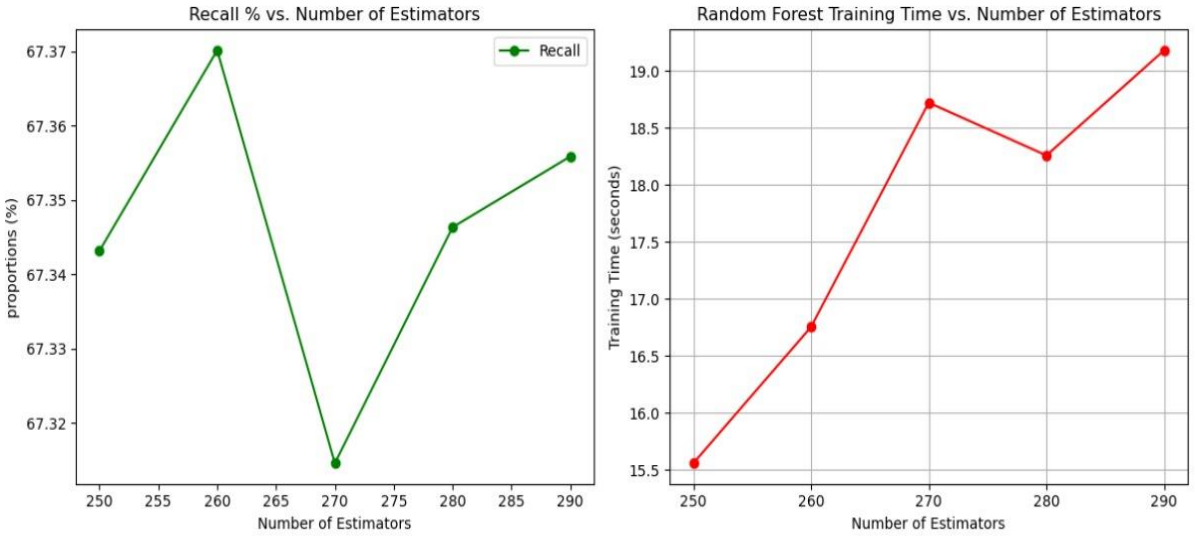
The bar chart on the right displays the "Number of Crimes by Day of Week." It shows that the highest number of crimes occurred on Fridays (31.4K), followed closely by Saturdays (30.3K) and Thursdays (30.1K). The numbers are relatively consistent throughout the week, with a slight drop on Monday and Sunday.

This dashboard provides an overview of crime trends over time and by specific categories, helping users to identify patterns and make informed decisions.

- **Using Machine Learning Approach**

  - **Random Forest Classifier**: The Random Forest Classifier is an ensemble learning method that combines multiple decision trees to improve prediction accuracy. It aggregates the results of several decision trees to make a final prediction, which helps in reducing overfitting and enhancing model performance.

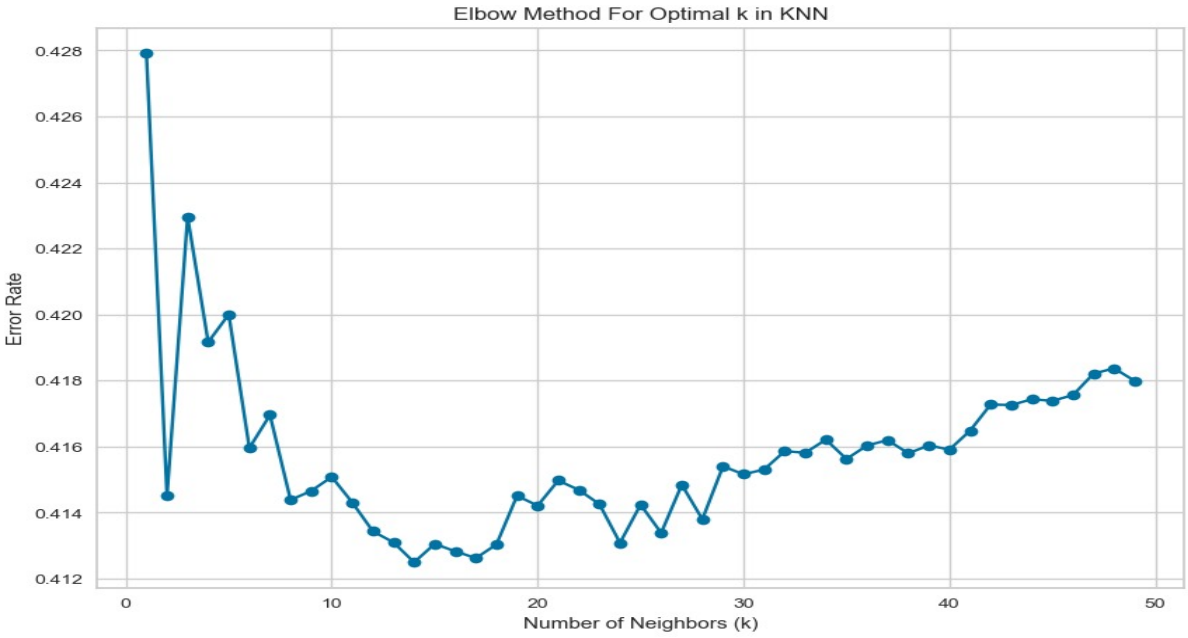
    - **Methodology**: We tested the Random Forest model with varying numbers of estimators, ranging from 250 to 290, with a step of 10, to find the optimal configuration.
    - **Metrics**: The classifier's performance was evaluated using a confusion matrix and classification report, providing insights into precision, recall, and F1 scores for different crime types. For example, it showed higher precision for 'Assault' but lower performance for 'Theft Over'.



(Fig 7.4: Random Forest: Accuracy & Time vs. Estimators)

The left graph indicates that the recall percentage for the Random Forest model fluctuates with different numbers of estimators, peaking at around 260 estimators and dropping around 270 estimators. The right graph demonstrates that the training time consistently increases as the number of estimators rises. These graphs highlight the trade-off between achieving higher recall and managing training time, suggesting that while more estimators might improve recall up to a point, it also results in longer training times.

- **KNN:** An instance-based learning approach called K-Nearest Neighbors (KNN) is straightforward and uses the majority class among its closest neighbors to classify data points. The class most prevalent among the closest k neighbors is assigned after calculating the distance between data points.



(Fig 7.5: Elbow Method)

This graph demonstrates the Elbow Method to find the optimal number of neighbors (k) for the K-Nearest Neighbors (KNN) algorithm. Initially, the error rate is high with low values of k, indicating poor performance. As k increases, the error rate decreases significantly, reaching its minimum around k = 14. This point, known as the "elbow," suggests the optimal k value.

Best Number of Neighbors: 14; This is the optimal number of neighbors for the KNN classifier, giving the lowest error among the tested range.

- **The Fit of the New Solution into the Existing IT Architecture**
  - **Integration into the existing system:**

    Visualization: Integrating Power BI's interactive visualizations into the TPS IT architecture enhances crime analysis by utilizing existing data from Records Management Services. It provides detailed matrices, charts, and interactive cards, supporting strategic decision-making and public transparency. The IT Services infrastructure ensures reliable performance, making the visualizations accessible to all stakeholders.

    Machine Learning Models: Integrating KNN and Random Forest classifiers into TPS's IT architecture enhances predictive analytics. These models use existing infrastructure for efficient operations and support evidence-based decision-making. Integration with Business and Policing Applications ensures real-time data flow, improving report accuracy and utility. This approach aligns with IT Risk Management practices, ensuring robust performance and system integrity while enhancing crime data analysis and resource allocation.

  - **Benefits of the New Solution:**

    Visualization: The integration of Power BI's interactive visualizations into the existing IT architecture offers numerous benefits. It provides a user-friendly platform for analyzing and interpreting crime data, making it easier for stakeholders to identify trends and patterns. This enhanced analytical capability supports more informed decision-making and strategic planning.

    Machine Learning model: Provides significant benefits in crime analysis by enhancing the accuracy of crime-type predictions. They enable efficient resource deployment and proactive crime prevention, allowing law enforcement to identify patterns and trends more effectively.

- **<u>Steps for Client Company:</u>**
  - **Data Import and Preprocessing**:
    - Import crime data into the system.

- Perform data cleaning to ensure the dataset is accurate and complete.
- Preprocess data by handling missing values, removing duplicates, and normalizing formats.

- **Model Training and Evaluation**:
  - Train the Random Forest and KNN models using the prepared data.
  - Evaluate the models using test data, recording accuracy, precision, recall, and F1 scores.
  - Perform cross-validation and hyperparameter tuning to ensure model accuracy and reliability.

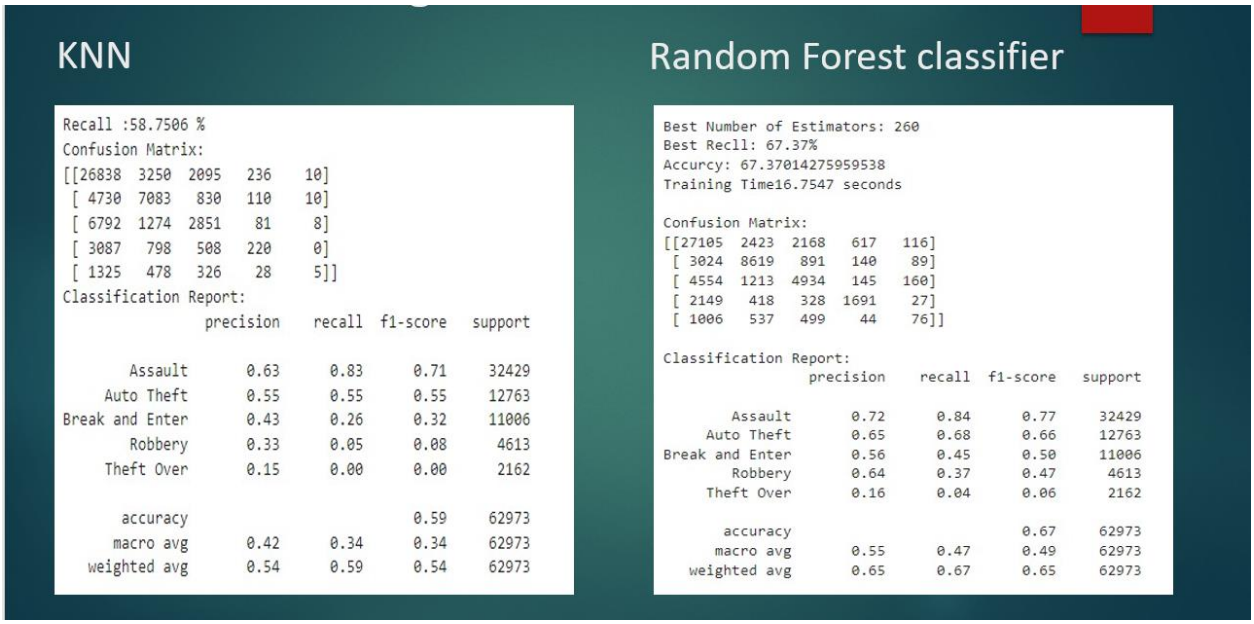
- **Utilizing Power BI for Visualization**:
  - Import cleaned and processed data into Power BI.
  - Create and customize visualizations such as matrices, clustered column charts, bar charts, and interactive statistics cards.
  - Use dropdown filters to refine data analysis based on specific criteria.

- **Continuous Improvement**:
  - Implement a feedback mechanism to capture user input and suggestions.
  - Regularly update and retrain models with new data.

Automate data preprocessing and analysis workflows to enhance efficiency and accuracy.

8. New Solution Implementation and Outcome Testing

- **Comparison between KNN and Random Forest Classification:**



(Fig: 8.1: Comparison between KNN and Random Forest Classifier)

The comparison between the K-Nearest Neighbors (KNN) and Random Forest classifier models reveals significant differences in their performance based on the metrics presented in their classification reports and confusion matrices.

K-Nearest Neighbors (KNN)

The KNN model achieved a recall of 58.75%, indicating that it correctly identified 58.75% of the positive instances across all classes. The confusion matrix shows that KNN struggled particularly with lower recall rates for the "Break and Enter," "Robbery," and "Theft Over" categories, with recalls of 26%, 5%, and 0%, respectively. The overall accuracy of the KNN model was 59%, with a weighted average precision of 0.54, recall of 0.59, and F1-score of 0.54. These results suggest that while KNN performs reasonably well in detecting "Assault" and "Auto Theft," it falls short in accurately predicting less frequent crime types.

Random Forest Classifier

In contrast, the Random Forest classifier demonstrated superior performance with a recall of 67.37%, substantially higher than KNN. The confusion matrix for the Random Forest model indicates better classification across all crime types, with notably higher recall rates for "Assault" (84%), "Auto Theft" (68%), and "Break and Enter" (45%). The overall accuracy of the Random Forest model was 67.37%, with a weighted average precision of 0.65, recall of 0.67, and F1-score of 0.65. This indicates a more balanced and reliable performance across different crime categories.

In summary, the Random Forest classifier outperforms the KNN model in terms of overall accuracy, recall, and precision. It handles the classification of various crime types more effectively, particularly in categories where KNN struggled. However, the significant gains in predictive accuracy and reliability make the Random Forest classifier a better choice for this application.

- **Scenario testing:**

- **Scenario 1:** Ensure that the Power BI visualizations accurately reflect crime data from the existing data files.

  Steps:

  1. Imported data file into Power BI.
  2. Verified crime distribution metrics (e.g., by neighborhood and division) match the source data.
  3. Checked the visualizations (matrix, clustered column chart, bar chart) correctly updated based on different time periods and filters.

- **Scenario 2:** Ensure that Power BI dashboards and reports fit into the current workflows and user processes.

  Steps:

  1. Integrate Power BI dashboards with existing workflows, such as crime analysis and resource allocation.
  2. Test interactive features like dropdown filters and interactive statistics cards within the TPS's operational environment.

3. Confirm that Power BI tools support current reporting requirements and enhance decision-making processes.

- **Scenario 3:** Validated the precision and performance of the KNN and Random Forest models on test data.

  Steps:

  1. Train both KNN and Random Forest models with training data.
  2. Test the models with unseen data and record accuracy, precision, recall, and F1 scores.
  3. Compare the performance metrics with pre-defined benchmarks.

- **Scenario 4:** Performance Testing

  Steps: Conduct load testing to ensure the system can handle the computational demands of the new models. Evaluate the models' performance under various conditions to ensure they meet the required speed and accuracy benchmarks.

- **Optimization:**
  - **For Visualization:**

    Optimized Approach: Using Talend Open Studio for ETL Processes

    To improve the overall result and performance of the dashboard, Talend Open Studio, an ETL (Extract, Transform, Load) tool, was introduced in the data preparation process.

    This approach involved the following steps:

    a) Data Extraction and Transformation:
    - Talend Open Studio was used to extract data from CSV files.
    - Performed necessary transformations on extracted data, such as cleaning, normalization, and aggregation, to ensure consistency and accuracy.

    b) Creation of Dimension and Fact Tables:
    - After transformation, created dimension tables files (e.g., divisions, offenceType, premiseType, date) and fact tables (e.g., fact_crime).
    - These files were stored in optimized formats, ready for import into Power BI.

    c) Importing Data into Power BI:
    - The pre-processed dimension and fact files were imported into Power BI.

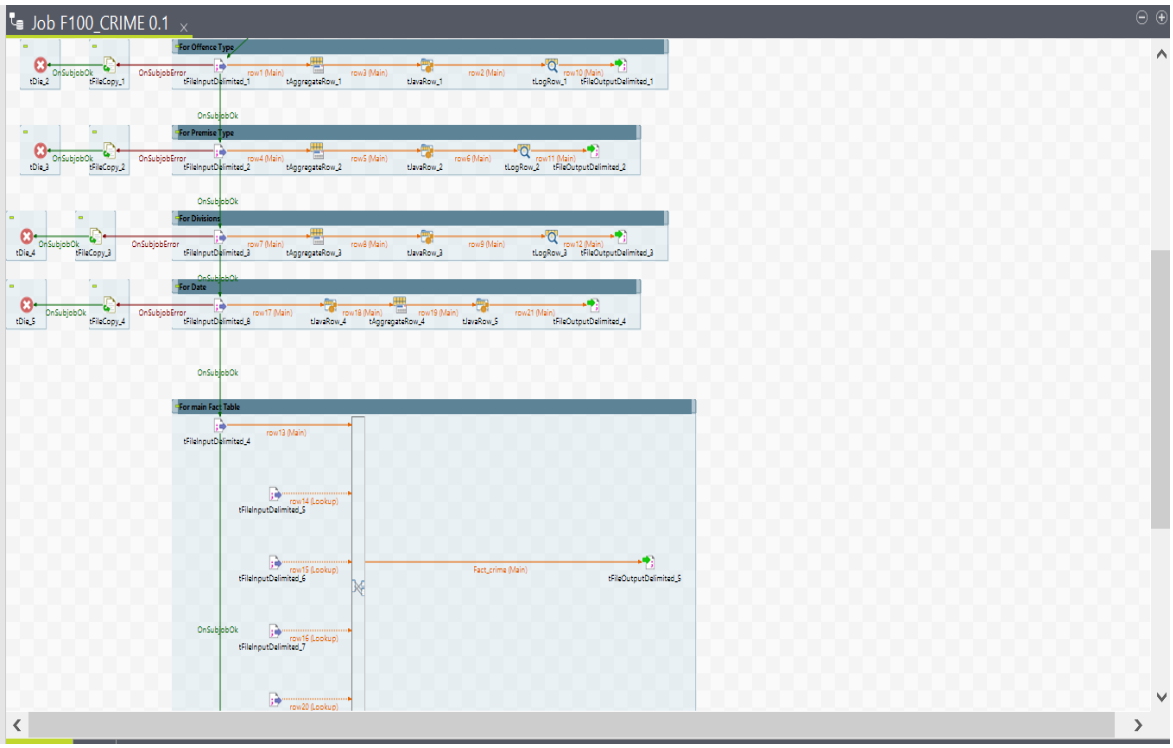
- This separation of concerns allowed Power BI to focus on data visualization and analytics, leveraging the optimized data structures created by Talend Open Studio.

d) Ongoing Monitoring and Data Updates:

- Regularly monitored data sources and ETL processes to ensure data integrity and accuracy.

- Regularly monitored data sources and ETL processes to ensure data integrity and accuracy.

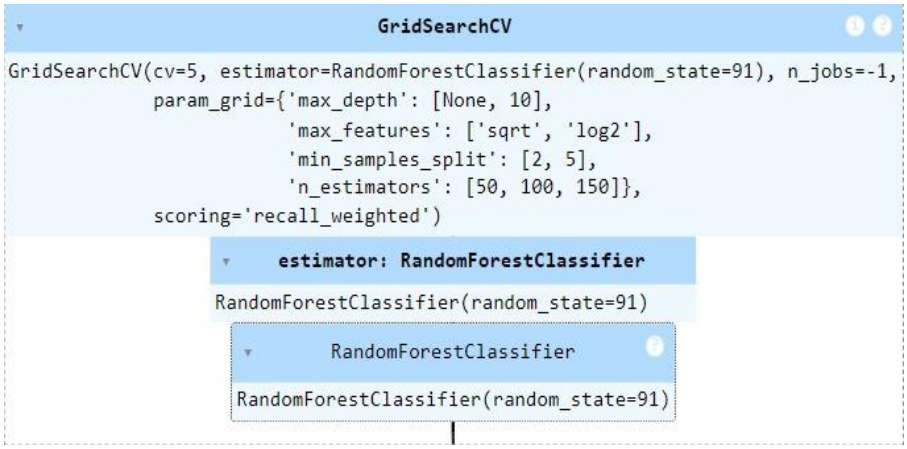
- Set up automated data updates to keep the data current and reflective of the latest information.



(Fig: 8.2: Optimization using ETL Job)

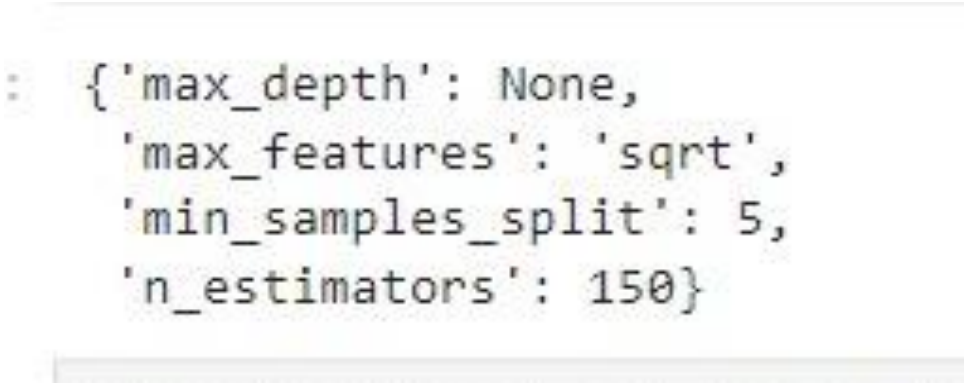
➢ **For Machine Learning**

Our Random Forest model performed better when we used GridSearchCV, a potent machine learning hyperparameter tuning tool. GridSearchCV finds the set of hyperparameters that gives our model the highest performance by conducting an exhaustive search over a given parameter grid.



(Fig: 8.3: GridSearchCV Method)

Hence, the set of best parameters after running the GridSearchCV method.



(Fig: 8.4: Best Parameters)

This shows the hyperparameters used for configuring a Random Forest classifier. Each hyperparameter plays a specific role in determining the behavior and performance of the model:

1. **max_depth: None**
   - This means that there is no maximum depth set for the individual trees in the forest. Trees will grow until all leaves are pure or until they contain less than min_samples_split samples.
2. **max_features: 'sqrt'**
   - This specifies the number of features to consider when looking for the best split. 'sqrt' means the square root of the total number of features will be considered. This is a common choice for classification tasks, as it helps reduce overfitting.

3. **min_samples_split: 5**
    - This indicates the minimum number of samples required to split an internal node. Setting it to 5 means that a node must have at least 5 samples to be considered for splitting. This can help prevent overfitting by ensuring that splits occur only when there is a sufficient number of samples.
4. **n_estimators: 150**
    - This defines the number of trees in the forest. Setting it to 150 means that the Random Forest model will be composed of 150 individual decision trees. A higher number of estimators generally improves performance but also increases computational cost and training time.

Below is the output of the Random Forest Classifier with the best set of parameters. The recall is 67.56%; which is more optimized.

```
Recall :67.5591 %
Confusion Matrix:
[[27572  2389  1990   429    49]
 [ 3101  8613   882   121    46]
 [ 4772  1187  4851   111    85]
 [ 2321   471   360  1448    13]
 [ 1032   539   499    32    60]]
Classification Report:
                precision    recall  f1-score   support

       Assault       0.71      0.85      0.77     32429
    Auto Theft       0.65      0.67      0.66     12763
Break and Enter       0.57      0.44      0.50     11006
       Robbery       0.68      0.31      0.43      4613
    Theft Over       0.24      0.03      0.05      2162

      accuracy                           0.68     62973
     macro avg       0.57      0.46      0.48     62973
  weighted avg       0.65      0.68      0.65     62973
```

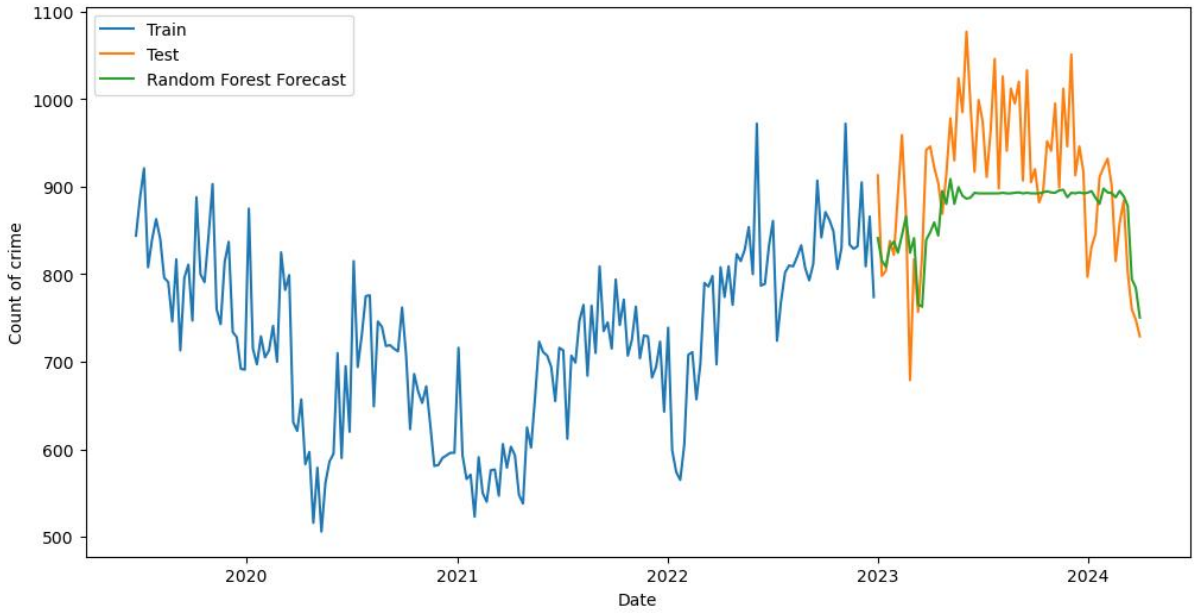
(Fig: 8.5: Optimized Result of Random Forest Classification)

9. Potential Future Solutions

As we continue to enhance our crime data analysis framework, applying advanced forecasting models presents a significant opportunity for optimizing the accuracy and efficiency of our predictions. By integrating models such as Random Forest Forecast, Long Short-Term Memory (LSTM) networks, and Exponential Smoothing Forecast, we can improve our ability to predict future crime trends, enabling proactive measures and more effective resource allocation. Below are potential optimizations for future development:

Random Forest Forecast:

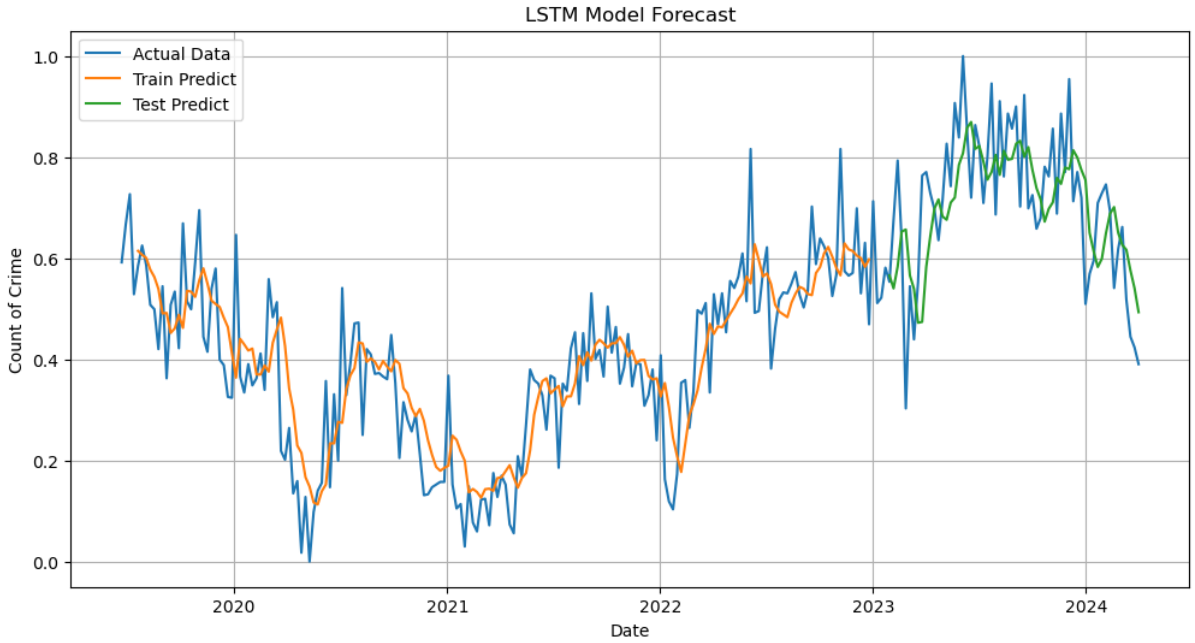
Random Forest Forecasting adapts the classic Random Forest algorithm for time series prediction by building multiple decision trees to capture complex patterns in historical data. This method effectively handles large datasets with features like time-lagged variables and seasonal indicators, making it ideal for forecasting crime frequency across different locations and periods.



(Fig: 9.1 – Random Forest Forecast)

LSTM Forecast:

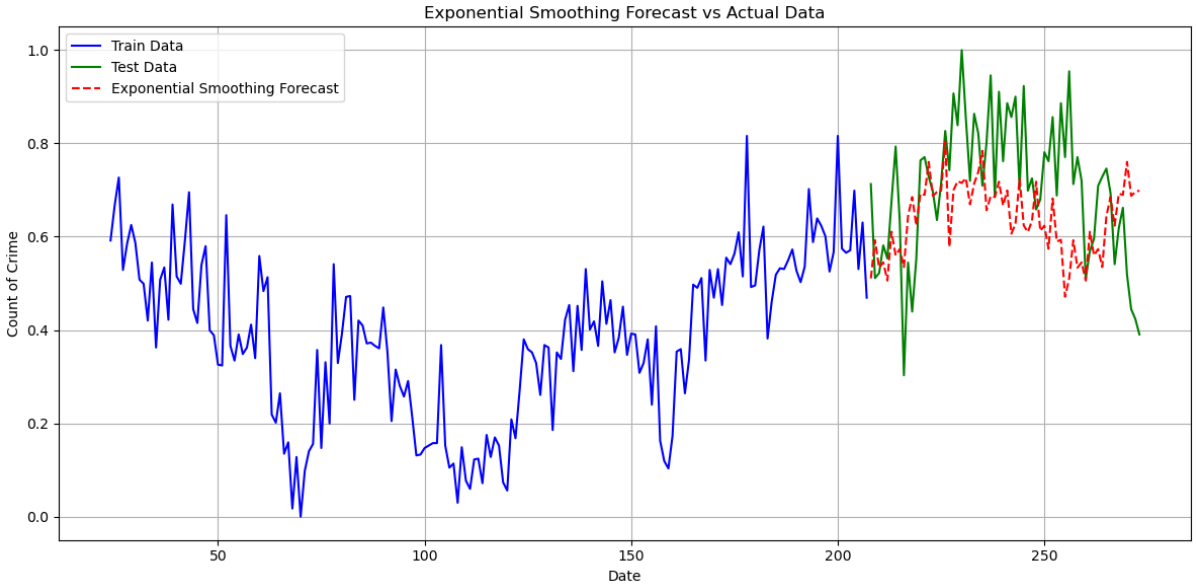
Time series forecasting benefits greatly from the use of LSTM networks, a type of recurrent neural network (RNN) that is specifically engineered to capture long-term dependencies in sequential data. When data shows intricate, long-term patterns, LSTMs—as opposed to conventional RNNs—use memory cells to store information over time, allowing for more precise forecasts of crime trends.



(Fig: 9.2 – LSTM Forecast)

Exponential Smoothing:

By highlighting current observations, exponential smoothing is a time series forecasting technique that removes noise from past data. It modifies predictions to reflect current trends by giving more weight to recent facts. This method produces more accurate projections by taking into consideration both short-term changes and long-term trends, which makes it especially useful for datasets with strong seasonal patterns, such as varying crime rates.



(Fig: 9.3 – Exponential Smoothing Forecast)

10. References

Toronto Police Service. (n.d.). Retrieved from https://data.torontopolice.on.ca/datasets/TorontoPS::major-crime-indicators-open-data/about

Appendix

Individual Reflections:

Prappan Batra

I was responsible for the critical task of data manipulation and scenario testing. My work involved cleaning and preprocessing the dataset to ensure its accuracy and completeness for further analysis. This included identifying and handling missing values, as well as dropping unnecessary records that could skew the results. Additionally, I tested the Power BI visualizations to ensure they accurately represented the crime data and integrated seamlessly with existing workflows. This included verifying that the visualizations, such as matrices and charts, correctly updated based on various filters and time periods. My contributions were essential in laying a solid foundation for the data analysis and ensuring that the final outputs were reliable and actionable.

Dhairya Dangi

I played a pivotal role in implementing new solutions and optimizing the existing processes. My responsibilities included designing and executing the data visualization strategies in Power BI and evaluating the performance of machine learning models. I led the efforts in comparing KNN and Random Forest classifiers, optimizing their performance using tools like GridSearchCV, and refining our approach with Talend Open Studio for ETL processes. These efforts were aimed at enhancing the accuracy and efficiency of our predictive models and visualizations. By optimizing the Random Forest model and streamlining data processing, I ensured that our project delivered high-quality insights and actionable recommendations.

Param Panchal

My focus was on ensuring the data's integrity and exploring advanced forecasting techniques. I handled data cleaning tasks to prepare the dataset for in-depth analysis, addressing any inconsistencies and ensuring that the data was accurate. I also investigated and applied forecasting models, including Random Forest Forecast and LSTM networks, to improve our ability to predict future crime trends. By examining these advanced techniques, I aimed to enhance the precision of our predictions and support more effective resource allocation. My contributions helped in advancing our understanding of crime trends and improving the predictive capabilities of our models.

Vrajkumar Patel

I was responsible for analyzing the results from our visualizations and machine learning models to provide actionable insights. My work involved interpreting data from the Power BI dashboards and extracting meaningful patterns and trends. I also played a key role in recommending improvements based on the analysis of model performance and visualization effectiveness. This included evaluating how well the dashboards and reports met the project's objectives and suggesting enhancements where needed. My efforts ensured that our findings

were clear, actionable, and aligned with the project's goals, ultimately contributing to informed decision-making.

Pravina Prajapati

I focused on the development and documentation of Power BI visualizations. This involved creating and customizing dashboards to effectively represent the crime data and meet the project's requirements. I also ensured that interactive features, such as dropdown filters and statistics cards, were fully functional and added value to the analysis. Additionally, I compiled comprehensive documentation for the project, detailing the processes and results. My work was crucial in presenting the project’s outcomes clearly and ensuring that all deliverables were well-documented and accessible to stakeholders.