Group: Insight Ink Crew

DATA CAPSTONE PROJECT

Final Part: Model Development and Deployment

**Submitted by:**

Baskaran, Asmita

Coimbatore Jayaraj, Pavithran

Kaur, Harmanjot

Keshav, Pooja

Nalchigar, Niloofar

Pham, Thi

# Baskaran, Asmita Coimbatore Jayaraj, Pavithran

Kaur, Harmanjot Keshav, Pooja

# Nalchigar, Niloofar

Pham, Thi

**Professor: Zeyad Azem**

Date: 16, December 2023

**1/ Solution Development** **I believe I would be a valuable asset to the JAM team due to my strategic marketing expertise, proficiency in data analysis tools, and experience managing social media channels. My project coordination and management skills, coupled with an entrepreneurial mindset, align well with JAM's dynamic environment. I am confident that my adaptability, excellent communication skills, and passion for effective marketing make me a strong fit for contributing to JAM's mission of getting 1 million people playing**

**Outline Modeling Strategy**

**I am drawn to JAM for several compelling reasons. Firstly, the company's mission to facilitate play and recreation aligns with my belief in the importance of fostering a positive and active community. The emphasis on inclusivity, as reflected in the statement 'JAM Is For Everyone,' resonates with my values of diversity and equal opportunity.**

**Additionally, the dynamic and entrepreneurial environment described in the job description excites me. The opportunity to contribute to a team that values creative thinking, initiative, and a proactive approach is precisely the kind of setting where I thrive.**

**Furthermore, the unique combination of sports leagues and virtual events offered by JAM presents an innovative and comprehensive approach to recreational activities. The company's commitment to creating a diverse and inclusive work environment, as mentioned in the job posting, aligns with my personal commitment to contributing to inclusive workplaces.**

**Overall, JAM's mission, values, and the dynamic work environment it offers make it an appealing and exciting opportunity for me to contribute my skills and passion.**

**In my previous role as a PMO Lead at Matiran, I served as a coordinator for the 'National Electronic ID for Iranian Citizens Program,' a large-scale initiative aimed at modernizing the National ID's IT infrastructure. This experience has uniquely positioned me for success in the Marketing Coordinator role at JAM.**

**As a PMO Lead, I led a team through the strategic planning, implementation, and progress monitoring of this critical program. I developed and maintained comprehensive project plans following PMBOK guidelines, ensuring effective change management and positively impacting the client's objectives.**

**My role involved not only strategic planning but also real-time progress tracking through the development of procedures and project analysis using tools like Power BI and Excel. These skills are directly transferable to the role at JAM, where effective marketing tactics and continuous improvement are essential.**

**Moreover, my experience in data analysis and reporting, both for the National Electronic ID program and a recent Business Research project at Irving Consumer Product, equipped me with the ability to create insightful and impactful project analysis reports and presentations. This skill is directly applicable to JAM's need for clear and accurate communications and marketing content.**

**Overall, my prior experience in coordinating large-scale programs has honed my organizational, project management, and data analysis skills, making me well-prepared to contribute effectively to JAM's marketing objectives. I am confident that my experience will set me up for success in the dynamic and fast-paced environment at JAM.**

**My two young daughters are my little comedians at home, and they've unintentionally become my humor mentors. One day, they decided to give our family cat a 'makeover' with stickers and glitter. The cat, less than thrilled with its newfound sparkle, put on a hilarious, glittery protest. The image of a disgruntled cat covered in stickers still makes us burst into laughter.**

**This parenting adventure has taught me that sometimes the best comedy arises from unexpected situations. While I may not have successfully transformed our cat into a sparkling superstar, the experience brought our family together in fits of laughter.**

**Feature Selection and Engineering:**

* **NEIGHBOURHOOD\_158:** This categorical variable was crucial for spatial analysis. By converting it into dummy variables or applying spatial encoding techniques, we could capture the inherent patterns of homicide occurrences across different neighborhoods. This allowed us to monitor the impact of neighborhood-specific interventions and policy changes.
* **OCC\_DATE:** The datetime feature was decomposed into year, month, day, and weekday components, providing a multi-faceted temporal view of the data. Seasonal decompositions were performed to identify and adjust for patterns related to weather, holidays, and other cyclical events.

**Model Selection:**

For our predictive analysis, the ARIMA model was specifically chosen for its strengths in handling time series data with underlying trends, seasonal patterns, and potential non-stationarity—which are all characteristics derived from the "OCC\_DATE" feature.

**Model Development**

**a. Descriptive Models:**

Created visualizations representing historical trends using the features 'OCC\_YEAR,' 'OCC\_MONTH,' 'OCC\_DAY' and ‘NEIGHBOURHOOD\_158’.

A graph with blue bars

Description automatically generated with medium confidence

A graph of a bar chart

Description automatically generated with medium confidence

A graph of different colored rectangular shapes

Description automatically generated with medium confidence

**A graph with different colored lines

Description automatically generated**

A graph of a bar graph

Description automatically generated with medium confidence

This Histogram reflects the frequency of the incidents for days of the year. As its heading through the end of the year, the frequency of Homicides happening increased.

A graph of a number of crime type

Description automatically generated

This is the trend of each Homicide type for days of the week. As we get through to the end of the week, we find increasing number of Homicides of Shooting.

A graph of a number of people

Description automatically generated with medium confidence

D31 is showing the highest risk in terms of frequency of Homicides. This is frequency of each category.

A graph with blue and white text

Description automatically generated

This pivot chart is reflecting neighborhoods proportion of shooting incidents in D31. By identifying 3 top neighborhoods in D31, resource allocation could be managed in more effective way specifically on weekends.

**b. Diagnostic Models:**

A map of a city

Description automatically generatedA diagram of a crime

Description automatically generated

In our exploratory data analysis, we focused on visualizing the distribution of homicides in Toronto to understand both spatial and temporal trends. This was achieved by leveraging two key features from our dataset: "NEIGHBOURHOOD\_158" and “OCC\_DATE”

1. **Spatial Analysis Using Heatmaps**:
   * **Purpose**: To identify high-risk areas within the city.
   * **Method**: We utilized a heatmap to represent the frequency of homicides across different neighborhoods. In this visualization, each point corresponds to a homicide incident, plotted based on its geographical location using latitude and longitude coordinates.
   * **Findings**: The heatmap revealed clusters of incidents in specific areas, highlighting neighborhoods with higher homicide rates. This spatial concentration of homicides is crucial for identifying zones requiring more focused law enforcement and community intervention strategies.
2. **Temporal Trend Analysis**:
   * **Purpose**: To understand how homicide patterns have evolved over time.
   * **Method**: We analyzed the "OCC\_DATE" feature, which provided the dates of the homicide incidents. This enabled us to track changes and trends in homicide rates over the years.
   * **Findings**: By examining the temporal distribution, we could observe any seasonal patterns, peaks, or declines in homicide occurrences. This temporal insight is vital for predicting future trends and preparing for periods with historically higher rates.

**Combining Spatial and Temporal Data**:

* Our approach of integrating both spatial (heatmap) and temporal (time series) analyses offers a comprehensive view of homicide patterns in Toronto. It allows us to pinpoint not only where the hotspots are but also when they are most likely to occur.
* This dual-perspective analysis is instrumental in devising more effective, data-driven strategies for crime prevention and community safety enhancements.

**c. Predictive Models:**

**A black text on a white background

Description automatically generatedA graph of a graph showing a number of people

Description automatically generated with medium confidence**

The ARIMA (Autoregressive Integrated Moving Average) model is a popular statistical approach for time series forecasting that can account for various patterns observed in historical data. Here's a detailed explanation of how it was employed in the context of forecasting future homicide counts:

1. **Utilization of "OCC\_DATE":** This model takes advantage of the temporal data from the "OCC\_DATE" feature, which includes the dates when homicides occurred. From this information, the model recognizes patterns over time, such as increases or decreases in homicide rates, which often align with seasonal trends (e.g., weather changes, holidays, and social events).
   * **Seasonality:** The model can capture recurring fluctuations in the data tied to specific times of the year. For example, if homicides increase during summer months, this seasonal effect is accounted for in the model's predictions.
   * **Historical Trends:** The model also identifies and integrates long-term trends in the data, such as an overall increase or decrease in homicide rates over the years.
2. **Neighborhood-Specific Models with "NEIGHBOURHOOD\_158":** Rather than creating a single model for the entire city, separate ARIMA models were developed for each neighborhood, as identified by the "NEIGHBOURHOOD\_158" variable. This approach allows for a more granular analysis by understanding the unique patterns in each area.
   * **Localized Predictions:** By stratifying the data by neighborhood, we could make precise forecasts for each area. This helps in understanding the nuances of each neighborhood, as factors influencing homicide rates can vary significantly from one area to another.
   * **Aggregation for City-Wide Insights:** Although each neighborhood model operates independently, their outputs can be combined to form an overall picture for the city. This aggregation is crucial for city-level planning and resource allocation, ensuring that the overall trends are not overshadowed by localized data.

**d. Prescriptive Models:**

The design of a prescriptive model would be to suggest strategies for law enforcement agencies based on historical data, temporal patterns, and types of homicides, the aim is to create a tool that assists law enforcement in making informed decisions and allocating resources effectively.

1. **Historical Data:**
   * The model incorporates historical data on homicide incidents, including details such as date, time, location, and characteristics of each case.
   * It analyzes patterns and trends in historical data to identify areas or time periods with higher incidence rates.
2. **Temporal Patterns:**
   * Temporal patterns refer to trends and variations over time. The model considers temporal factors such as the time of day, month, or year when homicides are more likely to occur.
   * For example, the model may identify if there are seasonal variations or specific days of the week when certain types of homicides are more prevalent.
3. **Types of Homicides:**
   * The model categorizes homicides into different types based on various characteristics such as the method used (shooting, stabbing, etc.) or the motive.
   * Understanding the types of homicides allows for targeted strategies. For instance, if a certain type of homicide is associated with specific demographics or locations, law enforcement can tailor interventions accordingly.
4. **Strategy Recommendations:**
   * Based on the analysis of historical data, temporal patterns, and types of homicides, the model generates recommendations for law enforcement strategies.
   * Strategies could include:
     + Increasing patrolling or presence in specific neighborhoods during identified high-risk periods.
     + Implementing community outreach programs or targeted awareness campaigns to address specific issues contributing to homicides.
     + Allocating resources for specialized units or task forces to address particular types of homicides.
     + Adjusting resource distribution based on identified patterns to enhance proactive policing.
5. **Dynamic Adaptation:**
   * The model should be designed to adapt dynamically to changing circumstances and evolving patterns. Regular updates and retraining can ensure that the model remains effective in providing relevant and timely recommendations.
6. **Feedback Loop:**
   * Incorporating a feedback loop is crucial. Law enforcement agencies provide feedback on the effectiveness of implemented strategies, and this information is used to refine and improve the model continuously.

In essence, the model would act as a decision support system, leveraging data-driven insights to guide law enforcement agencies in deploying resources strategically and proactively to prevent and address homicides effectively.

**e. Models Validation**

**1. Model Evaluation**:

* Rigorous back-testing on historical data was conducted by dividing the data based on "OCC\_DATE," ensuring the model was evaluated on temporal validity. We used a rolling forecast approach to simulate real-world performance where each forecast step was evaluated against actual data.
* For spatial validation, we compared the model's predictions against actual occurrences in each neighborhood denoted by "NEIGHBOURHOOD\_158," measuring the spatial accuracy of our predictions.

**2. Feedback Incorporation:**

Incorporating feedback from stakeholders is a crucial step in refining and improving the predictive model. Stakeholders, who may include law enforcement agencies, policymakers, or subject matter experts, can provide valuable insights and domain knowledge that may not be captured solely by the data. Their feedback can help identify potential biases, enhance feature selection, and improve the model's interpretability. Additionally, stakeholders can offer context-specific information that contributes to a more accurate representation of real-world scenarios. This iterative process of feedback and refinement ensures that the model aligns closely with the practical needs and challenges faced by law enforcement in addressing and preventing homicides in Toronto.

**f. Solution Selection**

After a meticulous process of model selection and validation, we have chosen the Autoregressive Integrated Moving Average (ARIMA) model for its robustness in handling time series data and its capacity to incorporate past events to predict future occurrences.

***2/ Solution Deployment:***

**Deployment Strategy for ARIMA Model to Predict Homicide Counts and Top 5 Neighborhoods in Toronto**

We have undertaken a comprehensive analysis of the Toronto Police dataset on homicides. Our objective was to construct a predictive model that could reliably forecast homicide counts and identify neighborhoods at a higher risk.

The ARIMA model, with its fine-tuned parameters, has demonstrated promising results in not only forecasting overall homicide counts but also pinpointing the top five neighborhoods with the highest predicted homicide rates: Moss Park, Mount Olive-Silverstone-Jamestown, South Riverdale, Glenfield-Jane Heights, and Black Creek. These forecasts empower law enforcement agencies to allocate resources efficiently and engage in proactive community safety initiatives.

Our deployment strategy for the ARIMA model is as follows:

1. **Model Finalization:**
   * We will train the ARIMA model on the entire dataset to leverage the full scope of historical data, ensuring the model has the most comprehensive learning possible.
   * The final model will then be serialized using Python's **‘pickle’** module, allowing us to store the model's state for future predictions.
2. **Infrastructure and API Development:**
   * We will deploy the model on a secure cloud server with scalable capabilities to handle varying loads of prediction requests.
   * An API will be created using Flask, which will serve as an interface for querying the model and retrieving forecasts.
3. **Integration and Testing:**
   * The API will be rigorously tested to validate its performance, ensuring accurate and timely responses to prediction requests.
   * We will integrate the API with an existing system used by the Toronto Police, allowing for seamless access to the predictive insights generated by our model.
4. **Monitoring and Maintenance:**
   * Continuous monitoring will be implemented to track the model's predictive performance and system health.
   * Regular updates and retraining sessions will be scheduled to maintain the accuracy of predictions as new data becomes available.
5. **User Interface and Accessibility:**
   * A simple, user-friendly dashboard will be developed, enabling stakeholders to visualize and interpret the model's predictions easily.
   * Comprehensive documentation will be provided, detailing the usage of the model and API.
6. **Security and Compliance:**
   * Adherence to data protection regulations and implementation of cybersecurity measures will be a priority to safeguard sensitive information.
   * Regular security audits will be conducted to ensure the integrity and confidentiality of the predictive system.

In summary, our deployment plan is designed to be robust, user-centric, and adaptable to evolving data trends. By implementing this ARIMA model, the Toronto Police can not only predict homicide counts with a high degree of accuracy but also deploy preemptive measures in high-risk neighborhoods, ultimately fostering a safer community.

***3/ Performance Management:***

* **Define performance metrics to monitor:**

In defining performance metrics for monitoring and analysing homicide incidents in Toronto, it's essential to consider several key factors that reflect the effectiveness, accuracy, and impact of your analytical models or solutions.

* Accuracy: The proportion of correct predictions over the total number of predictions made. It provides an overall measure of how often the model correctly predicts the type of homicide incidents.
* Precision: The ratio of true positive predictions to the total number of positive predictions made by the model. It measures the accuracy of positive predictions, specifically focusing on the model's ability to correctly identify the actual positive cases.
* Recall (Sensitivity): The ratio of true positive predictions to the total number of actual positive cases. It measures the model's ability to capture all actual positive cases without missing them.
* F1 Score: The harmonic means of precision and recall. It balances precision and recall, providing a single metric that considers both false positives and false negatives. It's useful when the class distribution is imbalanced.
* Confusion Matrix: A table representing the performance of a classification model. It shows the number of true positives, true negatives, false positives, and false negatives, providing a more detailed understanding of the model's performance for each class.
* Area Under the ROC Curve (AUC-ROC): A metric used for binary classification that measures the model's ability to distinguish between classes. It represents the area under the Receiver Operating Characteristic (ROC) curve, with higher values indicating better discrimination between classes.
* Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) (for regression models): These metrics measure the average magnitude of errors between predicted and actual values in regression models. Lower values indicate better model performance.

These metrics will serve as benchmarks to monitor the performance of the developed predictive model. Regular monitoring and evaluation using these metrics will help assess the model's accuracy, reliability, and suitability for predicting homicide incidents in Toronto, allowing for necessary adjustments and improvements if needed.

**Communication plans**

1. Stakeholder Engagement:

* Toronto Police Department: Regular meetings to present model developments, gather feedback, and discuss implementation strategies.
* Local Government Authorities: Quarterly briefings on model progress, highlighting actionable insights for policy-making and resource allocation.
* Community Organizations: Bi-annual workshops showcasing derived insights for educational programs and community initiatives.

2. Reporting Structure:

Monthly concise reports on model performance metrics (accuracy, precision, recall, etc.) and any updates.

Bi-annual detailed presentations illustrating visualizations, diagnostic findings, and predictive trends for strategic planning.

3. Medium of Communication:

Utilize email updates for brief summaries, attaching detailed reports and presentations as necessary.

Conduct in-person or virtual meetings for stakeholder consultations and workshops.

Develop a dedicated online platform for stakeholders to access model updates, reports, and insights.

4. Feedback Mechanism:

Establish feedback loops after each major presentation or report release to gather stakeholder input.

Create surveys or feedback forms to capture specific suggestions and concerns for model improvement.

5. Transparency and Collaboration:

Encourage collaboration among stakeholders through shared access to information and collaborative decision-making.

Maintain transparency in model development, sharing methodologies, data sources, and validation approaches.

**Decision Matrix Definition**

A decision matrix was developed to help law enforcement authorities prioritize measures and responses in response to homicide cases in Toronto.

**Criteria and weight:**

Effectiveness, 4

Possibility of Cost, 3

Feasibility, 2

Community Needs Alignment, 5

**Options:**

* Increase the number of police officers patrolling high-risk areas.
* Use Predictive Policing Models
* Programs for Community Outreach and Education
* Improved Investigation and Forensic Methods

**Assess each choice against each criterion on a scale of 1 to 5, with 5 being the best.**

*Increase the number of police officers patrolling high-risk areas.*

* Effectiveness: 4
* Possibility of Cost: 2
* Feasibility: 3
* Community Needs Alignment: 3

*Use Predictive Policing Models*

* Effectiveness: 5
* Possibility of Cost: 3
* Feasibility: 4
* Community Needs Alignment: 4

*Programs for Community Outreach and Education*

* Effectiveness: 3
* Possibility of Cost: 4
* Feasibility: 5
* Community Needs Alignment: 5

*Improved Investigation and Forensic Methods*

* *Effectiveness: 4*
* Possibility of Cost*: 3*
* *Feasibility: 3*
* Community Needs Alignment*: 4*

**Total Scores:**

Calculate the overall score for each choice by multiplying each criterion's score by its weight and adding the results.

* *Increase Police Patrols in High-Risk Areas*
  + Total Score: (4*4) + (2*3) + (3*2) + (3*5) = 16 + 6 + 6 + 15 = 43
* *Implement Predictive Policing Models*
  + Total Score: (5*4) + (3*3) + (4*2) + (4*5) = 20 + 9 + 8 + 20 = 57
* *Community Outreach and Education Programs*
  + Total Score: (3*4) + (4*3) + (5*2) + (5*5) = 12 + 12 + 10 + 25 = 59
* *Enhanced Investigation and Forensic Techniques*
  + Total Score: (4*4) + (3*3) + (3*2) + (4*5) = 16 + 9 + 6 + 20 = 51

The option with the greatest overall score is deemed the most preferable. In this scenario, the "Community Outreach and Education Programs" option gets the greatest overall score of 59, making it the most favorable strategy or intervention based on the provided criteria and priorities.

As a result, according to the decision matrix, law enforcement agencies should consider prioritizing or executing this specific approach.