**Final Project**

**Development a Predictive Model for Identifying Future Crime Hotspots Using Historical and Geospatial Crime Data**

Evaluation Report

Supervisor: Yoav Ziv

Ori Fogel 315729293

Iris Birman 208111377

30/5/2025

# Evaluating the Results

At the final stage of the project, the developed models successfully received crime-related input data - such as victim age, gender, ethnicity, and weapon usage - and produced predictions for both the type of crime (Category) and its likely location (BUREAU). This capability directly supports the project’s core business objectives of improving crime prevention, enhancing operational efficiency, and enabling informed decision-making through data-driven insights, even without direct involvement from law enforcement bodies.

By providing accurate and actionable predictions, the results empower stakeholders to allocate resources more effectively and take proactive measures, thereby strengthening preventive efforts and operational decision-making.

To ensure the results are presented clearly and in an accessible way, we provide structured tables and summary reports that present key model metrics and findings in a straightforward format for both technical and business audiences. Additionally, visualizations such as graphs and charts are used to highlight important trends and patterns in the data, making the insights easier to understand and communicate.

Looking ahead, we plan to develop an interactive dashboard that will allow users to input new crime data and receive real-time predictions for crime type and location. This tool will enable stakeholders to quickly analyze incidents and make informed decisions, supporting faster and more effective crime prevention and operational management. By combining clear reporting with practical, user-friendly tools, we ensure that the project results directly contribute to achieving the business goals.

During the analysis, we found some important and unexpected results. For example, the number of reported sexual assault cases in the data was noticeably lower compared to other crime categories. This could suggest possible underreporting or limitations in how the data was collected, highlighting a need for further investigation and improvement in data quality.  
Another interesting result was that property crimes tended to happen most often around noon. This pattern raises questions about what social or environmental factors might be influencing the timing of these crimes. Understanding these factors could help improve crime prevention efforts.

Regarding model performance, the XGBoost model demonstrated near-perfect accuracy for the Category target (F1-score ≈ 0.998), which, while impressive, raised concerns of potential overfitting - especially given its lower performance on the BUREAU target (F1-score ≈ 0.93). Conversely, the Random Forest model exhibited more balanced generalization, scoring 0.96 on Category and 0.49 on BUREAU. Although its BUREAU performance was lower, it suggested a reduced risk of overfitting, supporting its suitability for broader deployment.

Based on these results, we ranked and selected models according to their alignment with business goals: Random Forest was chosen for Category prediction due to its robustness, while XGBoost was selected for BUREAU owing to its superior accuracy despite associated risks. This dual-model strategy enables reliable forecasting of both crime type and location, directly addressing the business objectives.

The results also brought up some important questions for future work. One key question is why it was harder to accurately predict the location of the crime (BUREAU) compared to predicting the type of crime (Category). This might mean that we are missing important factors that influence location, or that the current features in the dataset don’t fully capture the geographic aspects of the incidents. Another question is how we can improve our feature engineering process - for example, by adding environmental or social data that might help the model better understand where crimes are likely to occur. These questions will help guide future improvements in both data collection and model development.

In summary, the approved models for deployment are Random Forest for Category and XGBoost for BUREAU, chosen based on their combined data science validity and business relevance.

# Review Process

As part of the CRISP-DM methodology, we took time at the end of the project to reflect on what went well and what could be improved. One of the main strengths of the project was the careful selection of models that fit our goals and data. We also performed thorough statistical analysis, which helped us better understand the data and evaluate model performance. In addition, the strong collaboration between team members made it easier to handle challenges and stay organized throughout the process.  
Another important advantage was the clear structure of the CRISP-DM model itself. We had step-by-step guidance throughout the project, so it was always clear what was expected at each stage and what the next steps should be. This helped us stay focused and work efficiently from beginning to end.

However, the project also faced several challenges. The large number of crime categories and geographic regions initially caused low model accuracy and confusion. Feature selection proved to be complex and time-consuming, requiring careful consideration to balance model performance and how easily the results can be understood and used. In addition, we initially tried to predict the exact location of crimes using geographic coordinates (longitude and latitude) as target variables. This approach led to poor model performance due to the high number of unique coordinates and the lack of repeated location values, which limited the model’s ability to learn patterns effectively. Moreover, predicting both coordinates along with the crime category created a three-target model, adding further complexity. To address these issues, we switched to using predefined areas (Area), and later refined the model further by grouping locations into broader regions (BUREAU), which helped simplify the classification task and improve results.

Looking ahead, we identify opportunities for improvement, including exploring a wider array of modeling techniques and investing more time in the Exploratory Data Analysis (EDA) phase. A deeper initial understanding of the data could have enabled more strategic decisions earlier in the process, potentially improving model outcomes. These lessons will inform future data science projects, particularly those involving multi-target predictions and geospatial data, helping to enhance effectiveness and reliability.