**Final Project**

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

Modelling Report

Supervisor: Yoav Ziv

Ori Fogel 315729293

Iris Birman 208111377

2/5/2025

# Selecting Modelling Techniques

The modelling techniques were selected based on the characteristics of the data and the objectives of the analysis. The selection process considered the type of variables to be predicted, the structure and quality of the dataset, and the requirements of different modelling approaches. The goal was to choose models that would provide accurate, reliable, and interpretable results.

# 1.1 Choosing the Right Modelling Techniques

To meet the project’s predictive goals, modelling techniques were selected based on the structure and quality of the data, the types of target variables, and the need for reliable and generalizable results.

The analysis focused on two targets:

**Category** **of crime type**– originally a high-cardinality categorical variable with approximately 140 distinct crime codes. These were grouped into five broader crime categories using keyword-based mapping to enhance interpretability and reduce complexity.

**Bureau** – created by aggregating the original 21 LAPD reporting areas into four geographic zones, based on the official division of the Los Angeles Police Department, to enable clearer spatial analysis.

Both targets are multi-class categorical variables, which require appropriate classification models. Although the dataset included over 1,000,000 records, the class distribution was highly imbalanced. Therefore, random under-sampling was applied to the training data to equalize the number of samples across classes and reduce bias.

Given the structured nature of the dataset and the modelling goals, Random Forest and XGBoost were chosen as the primary techniques due to their robustness, flexibility, and proven performance on tabular data. These models also support multiclass classification, handle both numerical and categorical inputs, and are relatively resilient to noise and overfitting.

We also experimented with both K-Means clustering and logistic regression models; however, due to their poor performance across key evaluation metrics, we decided to focus on the tree-based models detailed above. K-Means, as an unsupervised algorithm, is not designed for supervised classification tasks and struggles with the labeled and categorical nature of our crime data, especially after one-hot encoding increasing the feature space. Logistic regression, on the other hand, is a linear model that likely failed to capture the complex, non-linear relationships and interactions present in the data—particularly when dealing with multiple classes and high-dimensional input features.

# 1.2 Modelling Assumptions

During the model selection and preparation process, several assumptions were made to align the dataset with the requirements of the chosen models (Random Forest and XGBoost):

It was assumed that both models require complete, structured, and fully numeric input data.

To meet this condition:

* Missing values in the Weapon Used column were treated as absence of a weapon and encoded into a binary feature indicating weapon involvement.
* In the Vict Age feature, anomalous values such as negative numbers were replaced with the mean victim age.
* For categorical features like Vict Sex and Vict Descent, missing values were interpreted as ‘Unknown’ and encoded as 'X' to ensure consistency.

It was assumed that all categorical variables must be encoded numerically to be usable by the models. Therefore:

* Categorical features were transformed using label encoding or one-hot encoding, depending on their nature and cardinality.
* It was assumed that temporal features hold predictive value and should be represented in a way that the model can learn from:

The crime date was converted into the day of the week.

The crime time was converted to the hour of the day.

It was also assumed that class imbalance could lead to biased learning. Hence:

* The training set was balanced using random under-sampling of majority classes to match the size of the minority class (around 4,000 records).

In addition, it was assumed that only meaningful and non-redundant features should be included in the model. Therefore:

* Random Forest feature importance was used to identify the most impactful variables.
* Highly correlated features were removed after correlation analysis to prevent redundancy and reduce noise.

Lastly, it was assumed that simplifying complex categorical features would enhance model interpretability and generalization:

* The original 140 crime categories were grouped into five broader classes based on keyword patterns.
* The 21 LAPD areas were grouped into four larger zones (bureaus) based on LAPD’s official geographical divisions.

These assumptions guided key preprocessing decisions, ensuring the data was well-prepared and suitable for high-performing, robust classification models.

# Test Design

# The project will aim to predict future crime hotspots based on geographical areas and classify crime types. The main focus will be on predicting two specific targets: Category (crime type) and Bureau (geographical division based on consolidated areas).

To test the models, the dataset will be split into training and testing sets using stratified sampling, ensuring that the distribution of the Category variable is maintained in both sets. Only the training set will be balanced using under-sampling to equalize class sizes, while the test set will retain its natural distribution to allow for a realistic evaluation of model generalization and to avoid overfitting.

Model performance for the supervised tasks will be assessed using key classification metrics: Accuracy, Precision, Recall, and F1-Score.  
The evaluation will be based on a classification report that includes these metrics for each target variable (Category and Bureau), as well as for each class within them. Based on the results, model performance will be reviewed and refined iteratively. Multiple training iterations will be conducted, with adjustments made to both the model types (Random Forest and XGBoost) and their hyperparameters. If no significant improvement is observed after tuning, alternative modelling strategies or data preprocessing techniques will be considered. Since the project will focus solely on supervised learning tasks, evaluation methods for unsupervised models will not be required.

# Model Description

This section presents the modelling process, including the experimentation with multiple parameter settings and the evaluation of the resulting models. Both Random Forest and XGBoost were tested using cross-validation to optimize predictive performance on the two targets: Category (crime type) and Bureau (geographic area). The focus was on assessing the impact of various hyperparameter configurations, monitoring model behavior under different conditions, and selecting the best-performing settings for each algorithm.

# 3.1 Parameter Settings

In this project, we focused exclusively on tree-based models—XGBoost and Random Forest—both of which offer a variety of parameters that significantly influence performance. To explore the impact of different settings, we conducted a grid search across several key hyperparameters. Specifically, we tested different training set sizes (train\_size = 0.7, 0.8, 0.9) to observe how the amount of training data affects model accuracy. We varied the number of trees (n\_estimators = 50, 100, 150) and the maximum tree depth (max\_depth = 6, 10, 15 for XGBoost; None, 10, 20 for Random Forest). For XGBoost, we also experimented with different learning rates (learning\_rate = 0.05, 0.1, 0.15), which control how much each tree contributes to the overall prediction.

Parameter tuning was performed using stratified 5-fold cross-validation to ensure consistent and fair evaluation. The training data was balanced concerning both target variables, Category and BUREAU, before model fitting. For each combination of parameters, we calculated the average F1-scores for both targets across folds. This process helped us better understand which configurations were more effective and may serve as a basis for future modelling efforts.

# 3.2 Model Description

We focused on two ensemble models for predicting the target variables Category and BUREAU: Random Forest and XGBoost, each wrapped in a MultiOutputClassifier to support simultaneous multi-label classification. The optimal hyperparameters were chosen based on their performance on the validation sets. For Random Forest, the best results were achieved with n\_estimators=100, max\_depth=20, and train\_size=0.9, while for XGBoost, the best-performing configuration used n\_estimators=150, max\_depth=6, learning\_rate=0.10, and train\_size=0.7.

To make sure the model learns fairly and avoids overfitting, we balanced the classes only after splitting the data into training and test sets. This helped prevent the model from seeing information it shouldn't have during training. Class balancing was done by under-sampling the majority classes in the training set separately for each target variable.

From a practical standpoint, both models ran without technical issues, and training times were reasonable.

The dataset used had already undergone preprocessing, so no major problems related to missing values or data quality were encountered.

# Model Assessment

XGBoost produced particularly impressive results, especially on Category, where it achieved F1-scores nearing 0.999. However, the significant performance gap between Category and BUREAU raised concerns about potential overfitting. The near-perfect accuracy on Category and the relatively lower performance on BUREAU (F1 ~0.93) indicated that the model might be overly focused on patterns specific to Category, potentially sacrificing generalization to other target variables.

On the other hand, Random Forest demonstrated a more balanced approach, with Category achieving an F1-score of approximately 0.96 and BUREAU at ~0.49. While Random Forest exhibited lower performance for BUREAU, it showed better generalization, with less risk of overfitting compared to XGBoost. This suggests that, while there is room for improvement in predicting BUREAU, Random Forest may be more suitable for broader deployment, particularly when generalization across multiple targets is a priority.

To illustrate the comparison between the models, we will present the results in a table:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of model | Classification Metrics | Target variables | |
| Category | Bureau |
| XGBoost | Accuracy | 1.00 | 0.93 |
| Precision | 0.994 | 0.925 |
| Recall | 1.00 | 0.925 |
| F1-Score. | 0.998 | 0.925 |
| Random Forest | Accuracy | 0.96 | 0.49 |
| Precision | 0.96 | 0.4925 |
| Recall | 0.964 | 0.49 |
| F1-Score. | 0.964 | 0.49 |

# Considering the models differing performance across the target variables, we considered separating the models and applying different algorithms to each target: using Random Forest for the Category variable and XGBoost for BUREAU. This decision was driven by concerns about overfitting, as well as the model evaluation metrics discussed above. Below are the results of the separated model approach:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of model | Classification Metrics | Target variables | |
| Category | Bureau |
| XGBoost | Accuracy | - | 0.93 |
| Precision | - | 0.9325 |
| Recall | - | 0.9325 |
| F1-Score. | - | 0.9325 |
| Random Forest | Accuracy | 0.96 | - |
| Precision | 0.958 | - |
| Recall | 0.968 | - |
| F1-Score. | 0.962 | - |