**Final Project**

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

Changes Summery

Supervisor: Yoav Ziv

Ori Fogel 315729293

Iris Birman 208111377

26/08/2025

Throughout the development of the project, and particularly towards its conclusion, several practical changes were made compared to what was originally described in the submitted documents. As a result, there are differences in the model’s outcomes and in the evaluation of its performance.

**DR\_NO Attribute**

In the Data Understanding report, the **DR\_NO** attribute was described as:

*"Appears to be a unique identifier for each report and does not offer predictive insights"*

*"Division of Records Number: Official file number made up of a 2-digit year, area ID, and 5 digits"*

However, it was not sufficiently emphasized that this attribute is constructed from other attributes in the dataset (such as **Area ID**). Consequently, **DR\_NO** is not an independent variable but a composite identifier that inherently encodes information already present in other features. This explains its limited predictive value and the risk of overfitting when included in the models. Despite this, **DR\_NO** was not removed from the models described in the research documents. The derived nature of this variable led to very high results, raising concerns of overfitting, which were noted during the process.

Additionally, the variable **date reported** was identified as less relevant given the existence of the **occurrence date** variable. Nevertheless, when running *SelectKBest* with 4, 5, 6, and the full set of features remaining after preprocessing, the highest performance was achieved using all features rather than a limited subset.

**Premis\_cd Variable**

Although initially considered irrelevant, the variable **premis\_cd** was found to provide meaningful predictive information. As part of data preparation, the digit **0** was assigned as a representative code for “unknown” in samples where this field was missing.

**Class Imbalance**

During the analysis, an imbalance between classes of the target variables was identified, particularly in the target variable **category**, where the class *sexual assault* accounted for only 2% of the data. Initial balancing was performed manually using undersampling (random removal of samples from other classes). Later, we explored alternative balancing methods, including **SMOTE**, **RandomOverSampler**, and **RandomUnderSampler**, and additionally applied **class\_weight** to ensure maximal balance. The *sexual assault* class was deliberately preserved due to its business relevance, despite potential reductions in model performance metrics.

**Modeling Process**

Following submission of the initial documents and addressing issues identified during data preparation, we performed **5-fold cross-validation** for each target variable. Six machine learning models were tested: **Logistic Regression, Decision Trees, Random Forest, XGBoost, KNN, and SVM**, using the three balancing methods and the four K values mentioned above.

We also evaluated a 5-fold cross-validation with a **temporal train-test split**, where the training set contained older samples and the test set more recent events. This approach resulted in a decline in performance metrics, and therefore a **random split** was adopted instead.

The best-performing combination for both target variables was **Random Forest** with all relevant features remaining after preprocessing:

* **Category (8 features):** Date Rptd, Date OCC, Time OCC, Area, Vict Age, Vict Sex, Vict Descent, Weapon Used
* **Bureau (9 features):** Date Rptd, Date OCC, Time OCC, Crm Cd, Vict Age, Vict Sex, Vict Descent, Premis Cd, Weapon Used

The **RandomOverSampler** balancing method provided the best results. Detailed results of all runs are presented in the appendices.

**Performance Metrics**

**Bureau:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| 0 | 0.41079 | 0.33495 | 0.36901 | 24574 |
| 1 | 0.40875 | 0.56384 | 0.47393 | 21068 |
| 2 | 0.48453 | 0.44540 | 0.46414 | 29279 |
| 3 | 0.46886 | 0.44885 | 0.45864 | 25599 |
| **accuracy** | 0.44410 | 0.44410 | 0.44410 | 100520 |
| **macro avg** | 0.44323 | 0.44826 | 0.44143 | 100520 |
| **weighted avg** | 0.44663 | 0.44410 | 0.44153 | 100520 |

**Category:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| 0 | 0.71041 | 0.93493 | 0.80735 | 20272 |
| 1 | 0.41164 | 0.43558 | 0.42327 | 12209 |
| 2 | 0.68452 | 0.69878 | 0.69158 | 33162 |
| 3 | 0.07879 | 0.28125 | 0.12309 | 416 |
| 4 | 0.81754 | 0.60695 | 0.69668 | 34461 |
| **accuracy** | 0.68123 | 0.68123 | 0.68123 | 100520 |
| **macro avg** | 0.54058 | 0.59150 | 0.54839 | 100520 |
| **weighted avg** | 0.69969 | 0.68123 | 0.68173 | 100520 |

**Summary and Recommendations**

In summary, the model for predicting **crime type** performs relatively well but shows room for improvement. Adding features or applying alternative data cleaning methods could potentially enhance predictive performance.

In contrast, the model for predicting the **geographic area (Bureau)** exhibited limited predictive performance, suggesting that the data used is insufficient for accurate prediction of this variable. It is possible that more **specialized or localized machine learning models**, tailored to specific regions or patterns, could achieve better performance for predicting Bureau.

Future work could explore **feature engineering** and **specialized modeling approaches** to further improve predictive accuracy, particularly for the Bureau variable.

Appendix A- Modeling Process Evaluation

A.1 [Modeling Process for Bureau](https://github.com/IrisBirman/crime-dashboard/blob/main/BUREAU_cv_results%5B1%5D.csv)

A.2 [Modeling Process for Category](https://github.com/IrisBirman/crime-dashboard/blob/main/Category_cv_results%5B1%5D.csv)