Crime Prediction Using Public Data

Dhruv Hiteshkumar Arora

Devarsh Prashant Kale

Harsh Hiteshkumar Patel

CSP 571 – Data Preparation & Analysis, Spring 2025
Illinois Institute of Technology

1. Abstract

This study explores how supervised learning techniques can be applied to public crime data for the purpose of predicting arrest outcomes. Crime prediction plays a vital role in helping law enforcement improve decision-making and allocate resources more effectively. With crime data sourced from the City of Chicago, we investigated patterns, performed feature engineering, and built classification models. The main objective of the project was to assess whether features such as crime type, location, and time could predict whether an arrest was made.

We used over 750,000 records from 2022 to 2024 and performed extensive preprocessing, exploratory analysis, and model evaluation. Three classification algorithms were trained and compared—Logistic Regression, Random Forest, and XGBoost. Among them, XGBoost showed the highest predictive accuracy of 91.55% on unseen test data. This report details our entire data preparation and analysis pipeline and concludes with recommendations for improving predictive performance and supporting policy interventions.

2. Overview

The increase in availability of open government data has enabled researchers and analysts to derive insights that can assist public services. One such application lies in criminal justice, where data science techniques can enhance understanding of crime trends and aid in preventive strategies. In this project, we examined how arrest prediction can be approached using classification models based on labeled data.

The problem is modeled as a binary classification task, where the target variable is whether an arrest occurred. We hypothesized that temporal features (e.g., hour, month), along with categorical features like crime type and location description, can act as reliable predictors. The methodology involved data ingestion, transformation, exploratory analysis, model training and testing, as well as performance evaluation.

Prior research such as James et al. (2021) and Breiman (2001) has validated the effectiveness of statistical learning in binary classification problems. Inspired by these foundations, we adopted a mix of linear and ensemble learning algorithms to assess which model best captures patterns in arrest outcomes.

3. Data Processing

The raw dataset obtained from the City of Chicago's data portal includes crimes reported from 2001 onwards. For relevance, we filtered the data from 2022 to 2024, resulting in over 750,000 records. The dataset contains detailed information including date, time, location, crime type, whether it was domestic, whether an arrest occurred, and geospatial coordinates.

Initial steps included parsing the datetime column to extract Year, Month, and Hour. Fields with sparse or irrelevant data—such as geographic coordinates—were dropped. To reduce dimensionality and sparsity, we filtered only the top 10 categories in 'Primary Type' and 'Location Description'. This ensured that the resulting feature space was manageable and that we retained categories with sufficient representation.

Categorical variables were encoded using one-hot encoding to make them compatible with machine learning algorithms. Missing data was minimal in the filtered dataset, and no imputation was necessary. A binary flag for 'Domestic' was also retained. After processing, the dataset was split into a training set (80%) and a test set (20%).

4. Data Analysis

The exploratory data analysis revealed that crime rates exhibit noticeable variation by time and type. Crimes are more frequent at night and during weekends. July had the highest volume of crime, suggesting seasonal effects. Theft, Battery, and Criminal Damage emerged as dominant crime types. Street-level crimes were more frequent than those in commercial or public institutions. A correlation heatmap revealed weak correlations among numeric variables, with some moderate linkages between 'Domestic' and 'Arrest'.

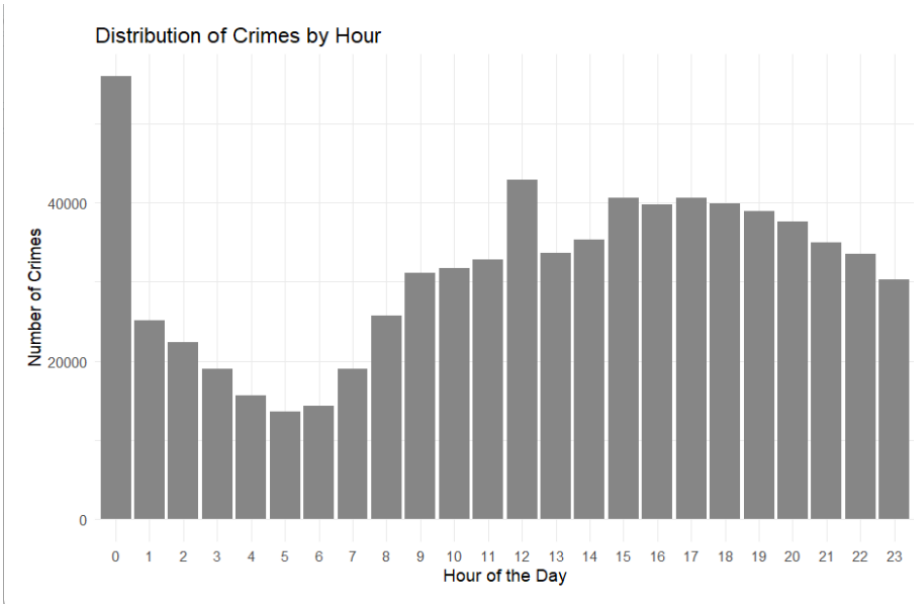


Figure 1: Distribution of crimes by hour of the day.

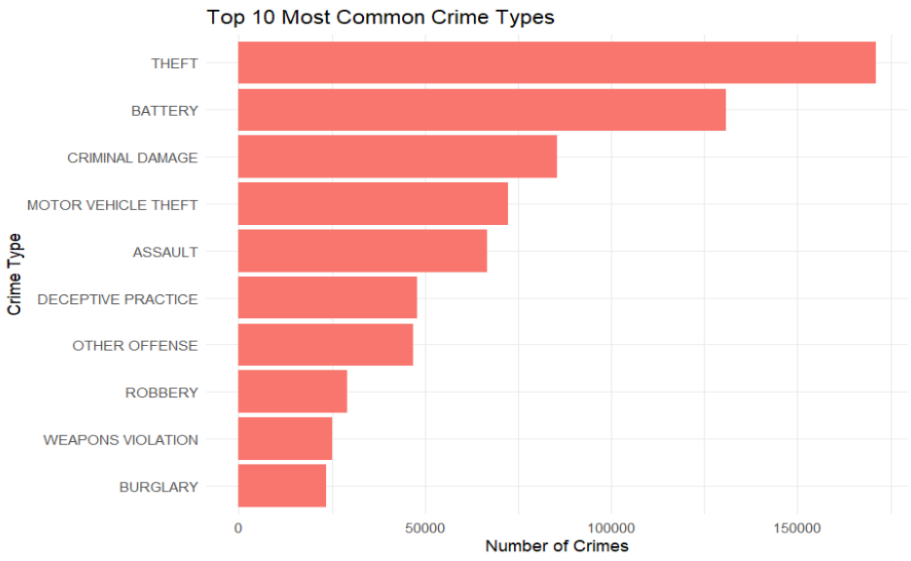


Figure 2: Top 10 most common crime types reported.

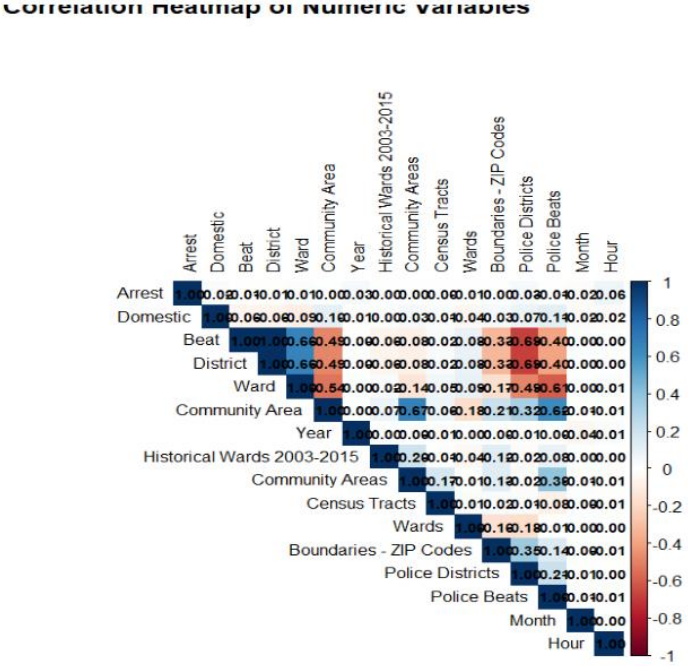
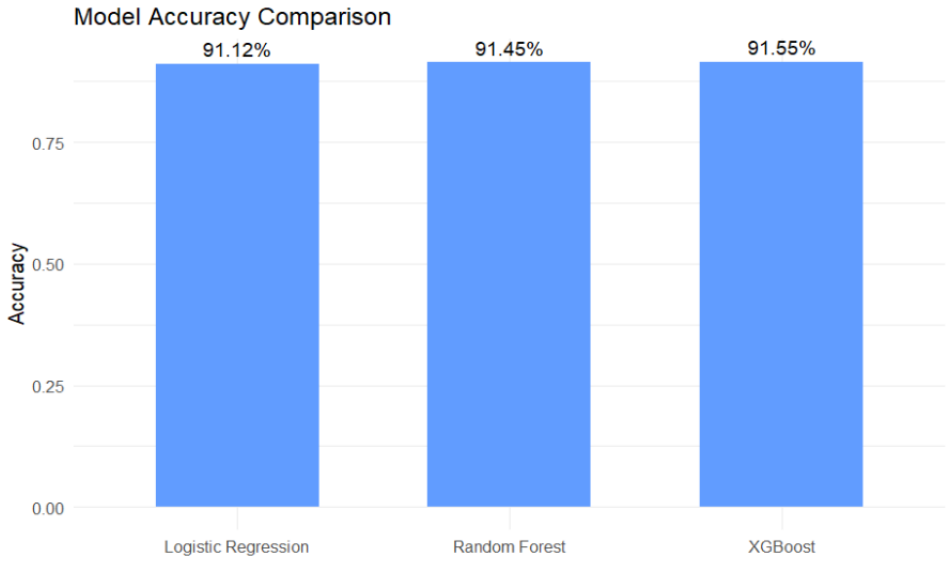


Figure 3: Correlation matrix of selected numeric features.

5. Model Training

To train predictive models, we used supervised learning techniques. Logistic Regression was used as a baseline due to its interpretability. Random Forest and XGBoost were chosen for their superior performance in classification tasks and ability to handle non-linearity. The training and testing data were balanced in terms of proportions. Categorical variables were encoded using one-hot encoding. Hyperparameters such as ntree (for Random Forest) and nrounds (for XGBoost) were selected based on initial cross-validation performance.



e.

Figure 4: Accuracy comparison between Logistic Regression, Random Forest, and XGBoost.

6. Model Validation

The test dataset was used to evaluate the models' ability to generalize. XGBoost showed the highest accuracy, followed closely by Random Forest and Logistic Regression. All models had high precision but relatively low recall due to class imbalance, meaning the models often predicted 'No Arrest' accurately but struggled to detect true 'Arrest' cases. Confusion matrices were plotted to visually interpret the classification errors made by each model.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 99459  8201
         1  1227  2680

               Accuracy : 0.9155
                 95% CI : (0.9138, 0.9171)
    No Information Rate : 0.9025
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3278

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.24630
            Specificity : 0.98781
         Pos Pred Value : 0.68595
         Neg Pred Value : 0.92383
             Prevalence : 0.09753
         Detection Rate : 0.02402
   Detection Prevalence : 0.03502
      Balanced Accuracy : 0.61706

       'Positive' Class : 1
```

Figure 5: Confusion matrix showing true vs predicted labels for the XGBoost model.

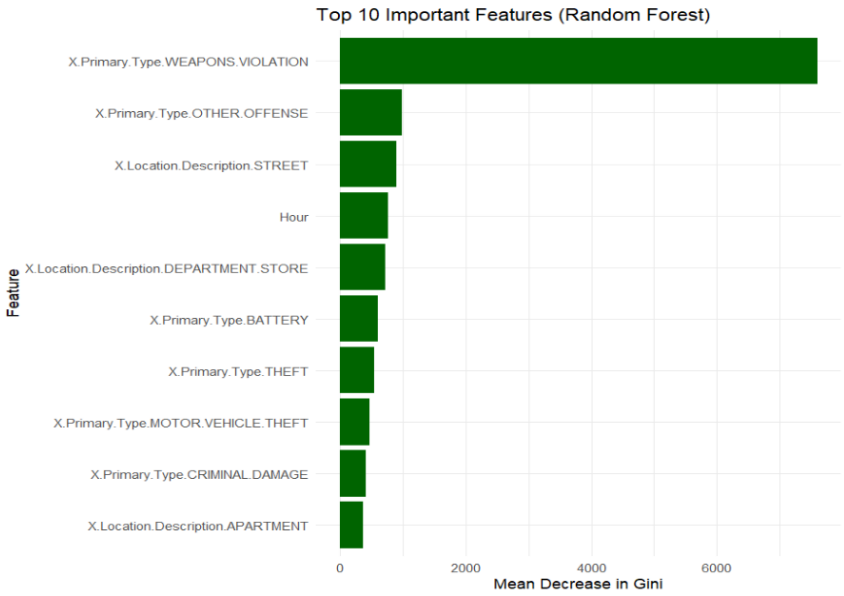


Figure 6: Top 10 most important features identified by Random Forest.

7. Conclusion

In conclusion, machine learning techniques can be effectively applied to predict arrests using time, location, and crime type data. XGBoost was the best-performing model, offering a good balance between accuracy and interpretability. However, limitations such as data imbalance and lack of external features suggest that future iterations should explore synthetic oversampling (e.g., SMOTE), geo-clustering, and integration of socio-economic data to enhance model robustness.

8. Data Sources

City of Chicago. (2024). Crimes - 2001 to Present [Dataset]. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

9. Source Code

The source code and scripts are organized into the following structure:
- data/raw/: Contains the original unprocessed CSV files.
- data/processed/: Holds the cleaned and encoded datasets used in modeling.
- scripts/: Includes R scripts for EDA, feature engineering, and model training.
- visuals/: Contains plots and charts used in reporting.
- report/: Includes RMarkdown, LaTeX, and DOCX documents for final submission.

Dependencies include R (v4.3+), caret, xgboost, randomForest, dplyr, ggplot2, and corrplot.

10. Bibliography

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning. Springer.
2. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. ACM SIGKDD.
3. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
4. Chawla, N. V., et al. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of AI Research.
5. Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.
6. Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning. Journal of Computational & Graphical Stats.
7. R Core Team. (2024). R: A Language and Environment for Statistical Computing. https://www.r-project.org/
8. Wickham, H., & Grolemund, G. (2017). R for Data Science. O'Reilly Media.

Following is the Project.rmd attached: