Chicago Crime Risk Prediction Using a Random Forest Model

Jenya Pandu (Section 03), Yanzhen Yun (Section 03), Vivian Zhao (Section 03)
Rutgers University - Introduction to Data Science 439
GitHub link | Dataset link

Abstract — This project aims to help individuals relocating to Chicago make safer housing decisions by predicting the crime risk level of various city areas. Using a dataset of reported crimes, we apply a Random Forest Model (RFM) algorithm—a collection of decision-tree models—to assign area codes a discrete risk score on a 1–10 scale. The model evaluates patterns based on location, time of day, and day of the week—factors known to influence crime likelihood. By combining these features, the system generates a crime risk score for each area, offering an approach well-supported by data for assessing urban safety.

INTRODUCTION

Urban safety remains a critical concern for individuals relocating to unfamiliar cities. In metropolitan areas such as Chicago, where over 300,000 crimes are reported annually, the volume and complexity of crime data make it difficult for potential residents to evaluate the relative safety of different neighborhoods. While public crime statistics are widely available, they are often fragmented and lack contextual interpretation. These raw figures do not provide personalized or actionable insights for individuals who wish to assess the risk of living in a particular area at a specific time.

This project seeks to address that gap by developing a supervised machine learning model that predicts localized crime risk levels in Chicago. The objective is to construct a data-driven tool capable of evaluating how risky a geographic district is—at a given time and day—based on patterns observed in historical crime reports. The model is designed to assist civilians, particularly those relocating to Chicago, in making informed decisions about neighborhood safety.

To achieve this, we aggregate crime incident records by District, Weekday, and Time of Day, and extract features that capture temporal trends (e.g., hour, day, month), categorical variables (e.g., crime type, location description), and aggregated metrics (e.g., violent crime rate, arrest rate). These features are used to train a supervised learning model, specifically a Random Forest Classifier (RFM). The Random Forest leverages multiple decision trees trained on bootstrapped samples to improve accuracy and robustness through ensemble voting.

The output of the model is a Risk Rating on a 1–10 scale, derived by quantile-binning total crime counts into deciles. This discrete score serves as a proxy for how "risky" a district is during a specific temporal window. By operationalizing these predictions into a user-facing format, our system offers practical utility for end users who require digestible and location-specific safety information.

In summary, this project demonstrates how interpretable supervised learning techniques can be applied to urban crime data to generate meaningful safety insights. The resulting tool supports informed decision-making by providing clear, evidence-based evaluations of neighborhood risk, grounded in empirical patterns extracted from historical records.

I. PURPOSE

Safety is one of the most important factors people consider when deciding where to live. Yet for many, especially those relocating to large urban centers like Chicago, evaluating neighborhood safety is difficult. Official crime data is often scattered across reports and dashboards, and interpreting it requires contextual knowledge that the average person may not have. This creates a barrier to informed decision-making—leaving individuals to rely on general perceptions or outdated stereotypes rather than data.

Our project is driven by the goal of making crime data more accessible, interpretable, and useful to the

public. Rather than just reporting past incidents, we aim to forecast risk in a way that is personalized to time, place, and setting. This is especially valuable in a city like Chicago, where crime rates vary widely not only by district, but also by hour of the day, day of the week, and type of location.

We are particularly excited about this project because it applies interpretable machine learning—Decision Trees and Random Forests—to a socially impactful problem. It not only allows us to work with real-world public safety data, but also gives us the opportunity to build a tool that could directly benefit individuals making important life choices about where to live or travel.

This work also contributes to a growing field of research that asks:

- Can machine learning models accurately and fairly predict crime risk?
- How can models be designed to assist civilians rather than only law enforcement?
- What features of time, place, and crime type are most predictive of public safety?

By addressing these questions, our project sits at the intersection of data science, urban planning, and public safety—demonstrating how analytics can help people make more confident and informed decisions about their environment.

II. RELATED WORK

Recent studies have demonstrated how machine learning can be used to uncover and predict crime patterns using historical and socioeconomic data.

One project by Miller [1] used economic indicators—such as unemployment and income inequality—to forecast crime rates across Ohio counties. The results showed these factors had significant predictive power when used in supervised models. Another notable study from the

University of Chicago [2] built a model capable of predicting future crime incidents with up to 90% accuracy. However, it also revealed potential bias in police data, raising important concerns about fairness and data integrity in algorithmic predictions.

In line with standard definitions, we use the FBI's classification of violent crime—which includes murder, rape, robbery, and aggravated assault [3]—to guide our feature engineering and interpretation of results.

Our work builds on these studies by applying an interpretable machine learning model, Random Forests, to assess neighborhood-level risk across Chicago. Unlike prior work that focuses on law enforcement applications, our model is designed as a tool for individuals moving to new areas, helping them assess local safety conditions in an understandable and actionable way.

III. PROBLEM FORMATION

The primary goal of this project is to build a machine learning model that estimates a Risk Rating for a given combination of geographic and temporal factors within the city of Chicago. Rather than predicting individual crimes, the model classifies areas (defined by police District) into one of 10 risk levels based on crime frequency, type, and context. This allows for a scalable, interpretable system that can guide individuals—especially those moving into the city—on neighborhood safety.

A. Input

(1) Historical Crime Dataset ($\mathcal{D}$):

Each observation (row) in $\mathcal{D}$ represents a single crime instance and contains:

- Temporal Features Year, Month, Day, Hour, Weekday, Is_Daytime

- Categorical Features

$$PrimaryType \in \{Homicide, Battery, Theft\}$$
$$LocationDescription \in \{Street, Apartment, Gas\ Station\}$$

- Aggregated Metrics by Group (per District × Weekday × Time of Day)
    - Total crime count: C
    - Violent crime count: V
    - Property crime count: P
    - Arrest rate: A ∈ [0, 1]
    - Domestic crime rate: D ∈ [0, 1]
    - Location District ∈ {1, 2, ..., 25}

B. Output

(1) Risk Rating R ∈ {1, 2, ..., 10}:

A discrete score indicating how risky a district is during a particular time period, derived by binning total crime volumes C into deciles:

$$R = QuantileBin(C, q = 10)$$

C. Approach

Feature Engineering:

Each timestamp was decomposed into interpretable components (hour, day, month, weekday). A binary feature Is_Daytime was added:

$$Is_{!"\#\$\%\&'} = \begin{cases} 1\ if\ 6 \le Hour < 18 \\ 0\ otherwise \end{cases}$$

Categorical features such as District and Weekday were one-hot encoded, creating sparse binary vectors to avoid ordinal assumptions.

Model Training

A Random Forest Classifier was implemented, which leverages the strengths of individual decision trees while mitigating their tendency to overfit. The Random Forest builds an ensemble of *T* trees, which are each trained on a bootstrap sample of the data and a random subset of features at every split. Final predictions are made by majority voting across all trees, yielding greater stability and accuracy than any single tree.

1. Basic Learning with Decision Trees

Each decision tree in the forest learns by splitting the data into smaller parts to reduce error. Recursive splitting continues until it reaches a maximum depth or minimum leaf size, producing a set of simple if-then rules.

$$\text{gini(t)} = 1 - \sum_{i} p(i|t)\text{^}2$$

where $p(i \mid t)$ is the proportion of samples of class i at node t.

2. Random Forest Classifier

Random Forest trains many trees using different random bootstrap samples and random feature choices. Then, it combines their results through majority voting. This helps make predictions more accurate and less likely to overfit.

$$R = 1/T \sum_{t=1}^{T} F\_t(X)$$

where f_t is the prediction from the t-th tree, and T is the total number of trees.

This model uses T=100 trees, each with a maximum depth of 8. While Random Forests can be harder to understand than a single decision tree, the model maintains interpretability by looking at feature importance scores. These scores show how much each factor—like violent crimes, property crimes, arrest rates, and time-related info—affects the predicted risk level.

D. Implementation Structure

The implementation of the risk classification model followed a machine learning pipeline to ensure data integrity, interpretability, and reproducibility. The key stages were as follows:

1. Data Preprocessing:

The raw dataset was cleaned by removing null entries and duplicate rows. Columns unrelated to modeling (e.g., ID, block number, or FBI code) were dropped to reduce noise.

2. Feature Engineering:

Temporal variables, such as date and time, were broken down into parts like hour, day, month, year, weekday, and whether it was daytime (6:00–5:59 PM). These additions allowed the model to find patterns based on time.

3. Aggregation:

Crimes instances were grouped by District × Weekday × Is_Daytime. Then, totals such as number of crimes, types of crimes, arrest rate, and domestic cases were calculated based on this aggregation. This allowed the model to understand patterns in each district.

4. Encoding:

Categorical variables like District and Weekday were one-hot encoded to convert them into a machine-readable numerical format while avoiding any unintended ordinal interpretation.

5. Train-Test Splitting:

The dataset was split into training and testing subsets using an 80/20 split. Stratified sampling was applied to maintain the original distribution of the Risk Rating across both sets.

6. Model Training:

A Random Forest classifier was used for training. The model builds many decision trees using random subsets of the data, then combines their predictions. This approach helps improve accuracy while still providing some interpretability through simple decision rules within each tree.

7. Model Evaluation:

The model was evaluated on the test set using standard error metrics to assess its predictive accuracy and ability to generalize on unseen data.

E. Evaluation Methodology

Model performance was quantitatively assessed using two standard metrics:

Mean Absolute Error (MAE):

This metric measures the average absolute difference between predicted and actual risk scores:

$$\mathrm{MAE} = 1/n \sum |p - r|$$

where p is the predicted rating and r is the actual Risk Rating for sample i.

R-squared Score (R^2):

This metric captures the proportion of variance in the target variable explained by the model:

$$R^{-} = 1 - \frac{\mathbf{d}\ (r-p)\text{^}2}{\mathbf{d}\ (r-r)\text{^}2}$$

where r is the mean of the actual Risk Ratings across the dataset.

These metrics provide insight into both the accuracy and robustness of the model's predictions on unseen data.

F. Generalization and Testing Strategy

To test how well the model works on new data, 20% of the data was set aside and not used during training. The split maintained the same distribution of risk scores for a fair performance check. This method helps ensure that the model stays reliable when used on future crime patterns or in new locations and times that are not present in the training data.

IV. RESULTS

This section presents the performance outcomes of the supervised learning model applied to the historical Chicago crime dataset. The objective was to classify spatial-temporal groupings (District × Weekday × Daytime/Nighttime) into one of ten discrete crime risk levels, based on crime count, type, and contextual indicators.

Model Performance Metrics

The final Random Forest Classifier was trained on 80% of the data and evaluated on the remaining 20%, with stratified sampling to preserve the distribution of the Risk Rating target variable. The following key metrics were used to assess model accuracy:

Mean Absolute Error (MAE): 0.587

This indicates that, on average, the model's predicted risk rating deviates from the true rating by less than 0.6 points on the 1–10 scale. Given the discrete nature of the target, this is a strong result—

especially since a ±1 deviation is generally tolerable in risk classification.

R^2 Score: 0.916

The R^2 score demonstrates that 91.6% of the variance in risk levels is explained by the model. This high value suggests excellent generalization and minimal overfitting, especially considering the categorical and context-rich nature of the input features.

These results confirm that the Random Forest model can reliably capture complex, non-linear interactions between location, time, and crime characteristics to predict area risk.

Feature Importance Analysis

To interpret model behavior, we examined feature importance using the feature_importances_ attribute of the trained Random Forest:

- The top three features—violent_crimes, property_crimes, and arrest_rate—contributed the highest predictive value.
- Domestic crime rate also ranked highly, reflecting the importance of private or interpersonal violence in shaping community-level risk.
- One-hot encoded Districts and Weekdays varied in importance, with some geographic areas (e.g., Districts 5, 10, and 15) and temporal patterns (e.g., weekends) influencing the model more significantly.

This aligns with real-world criminological literature, where violent incidents and law enforcement response rates are consistent predictors of neighborhood safety.

Visualization of Patterns

A. Feature Importance Chart

Figure 1. Random Forest Feature Importances: This bar chart shows the relative importance of each input feature in predicting crime risk levels. Features like violent_crimes, property_crimes, and arrest_rate are the most influential.

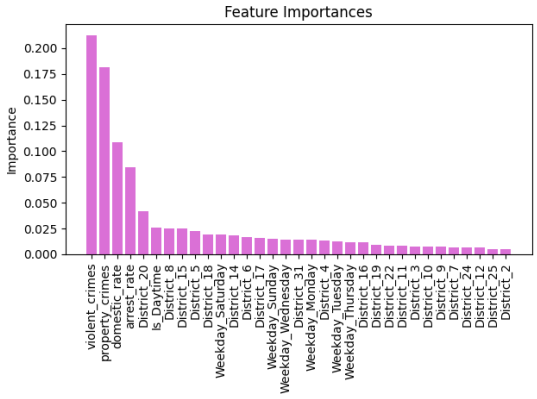

Figure 1: Feature Importance Plot

The feature importance plot visually demonstrates how the model prioritizes different features when making predictions. Notably, numerical indicators (crime counts and rates) dominate categorical indicators (day of week, district ID), supporting the design of our feature space.

B. Temporal Heatmaps

We produced comparative heatmaps showing total crimes per District across each day of the week, separated into daytime (6AM–5PM) and nighttime (6PM–5AM) periods. These heatmaps reveal:

- Temporal Clusters: Certain districts consistently exhibit higher risk during nighttime hours, particularly on Fridays and Saturdays, suggesting correlations with nightlife or social gatherings.
- Spatial Inequality: Some districts (e.g., 7, 10, 15) show elevated crime volumes regardless of weekday, indicating persistent structural vulnerabilities.
- Asymmetric Patterns: Crime risk is not symmetrically distributed between day and night, which highlights the importance of incorporating Is_Daytime as a feature.

Figure 2. District vs. Weekday Heatmap – Daytime Crimes: Aggregated crime counts for each district-day pairing during daytime hours (6AM–5PM).

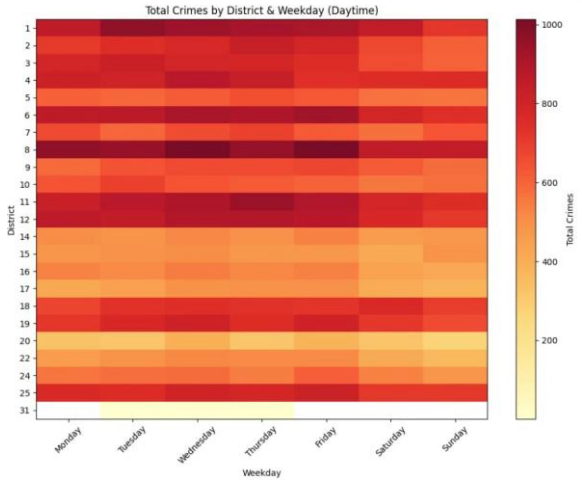

Figure 2: District vs. Weekday Heatmap – Daytime Crimes

Figure 3. District vs. Weekday Heatmap – Nighttime Crimes: Aggregated crime counts for the same pairings during nighttime hours (6PM–5AM), revealing different spatial and temporal patterns.

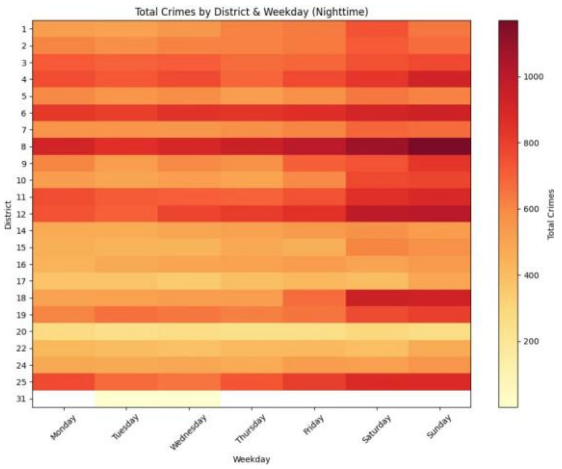

Figure 3: District vs. Weekday Heatmap – NightTime Crimes

Together, these visualizations aid in translating raw model predictions into actionable insights for users or policymakers.

Error Analysis and Generalization

The low MAE and high R^2 scores suggest strong performance, but a closer inspection of prediction errors reveals slight underperformance in rare edge cases—such as unusually low-crime districts during weekends. These cases may be the result of data sparsity, which weakens the signal-to-noise ratio for classification.

To improve generalization, future iterations of this model might incorporate:

- Temporal Validation: Train/test splits by year or season to evaluate robustness across time windows.
- Cross-district Testing: Hold out entire districts during training to assess how well the model extrapolates to unseen locations.

Comparison to Baseline Models

While we did not train a standalone Decision Tree Classifier, the structure of the Random Forest is built on individual decision trees. However, Random Forests outperform single-tree models in practice by reducing overfitting and improving stability through aggregation. Based on our findings:

$MAE \approx 0.587$
$R^2 \approx 0.917$

These results reflect the benefits of this Random Forests ensemble approach, especially when modeling complex, high-variance data like urban crime patterns.

Summary of Findings

- The model can predict risk levels with high accuracy and low deviation, making it useful for civilians, researchers, or policy analysts.
- Violent and property crimes—along with enforcement indicators like arrest rate—are the most impactful features.
- The framework supports interpretability and scalability, making it adaptable to other urban areas or extended datasets.

This model demonstrates that supervised learning can go beyond raw statistics to generate contextualized and interpretable safety assessments that meet the needs of real users.

References

Federal Bureau of Investigation. (2019). Violent crime definition – Crime in the United States 2019. https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/violent-crime

MiddleHigh. (n.d.). Chicago crime data from 2000 [Data set]. Kaggle. https://www.kaggle.com/datasets/middlehigh/los-angeles-crime-data-from-2000

Miller, T. (2022). Predicting crime using economic indicators. Bowling Green State University Honors Projects. https://scholarworks.bgsu.edu/honorsprojects/1011

University of Chicago Biological Sciences Division. (2022). Algorithm predicts crime but reveals police bias. https://biologicalsciences.uchicago.edu/news/algorithm-predicts-crime-police-bias