Analyzing the Relationship Between Weather, Unemployment Rate, and Violent Crime in Chicago (2011-2023)

Daisy Harris, Caroline Jarman, Camilla McKinnon

April 2024

Introduction

This project aims to analyze crime reports, specifically focusing on incidents of violent domestic and nondomestic nature in Chicago spanning from 2011 to 2023. Our goal is to identify potential indicators or correlations between weather conditions and unemployment rates on the incidence of violent crimes. Through in-depth examination of official crime reports provided by the Chicago Police Department, in conjunction with weather data and unemployment rates from reputable sources, our objective is to provide insights that can effectively inform crime prevention strategies and interventions. While we acknowledge the influence of various socio-economic and demographic factors on crime rates, our analysis will primarily focus on these specific factors due to time constraints. Success will be measured by the identification of statistically significant correlations and the practical applicability of insights in informing evidence-based crime prevention strategies and policies.

Exploratory Data Analysis

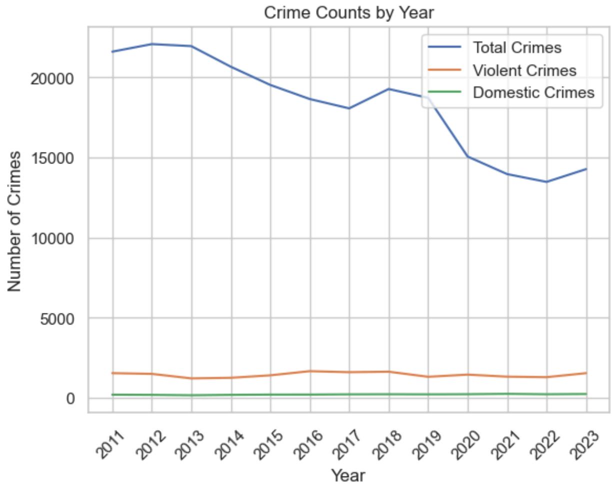
Our analysis delves into a subset of crime reports obtained from the CLEAR system of the Chicago Police Department, which provides insights into a specific area of Chicago. Typically, crime rates are assessed in relation to the population, often expressed as crimes per 1000 individuals. However, due to the absence of precise population data for this particular area and its fluctuations over the years, we recalibrated our approach. Our primary response variable became the daily count of violent crimes, derived from aggregating reports into daily totals, offering a more localized perspective on violent crime trends from 2011 to 2023. We calculated the yearly average for violent crimes per day, as shown in the table below:

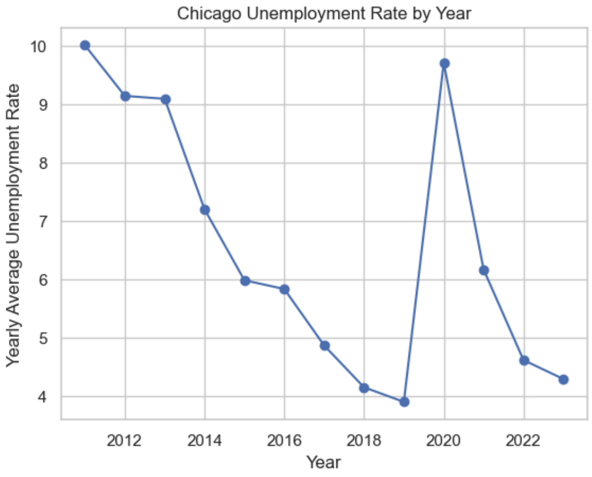
Table 1: Yearly Average for Violent Crimes per Day (2011-2023)

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Avg	4.27	4.25	3.51	3.60	3.95	4.60	4.42	4.56	3.73	4.11	3.78	3.60	4.28

The following visualizations showcase annual trends in total crimes, violent crimes, and domestic crimes during this period, as well as the unemployment rates.

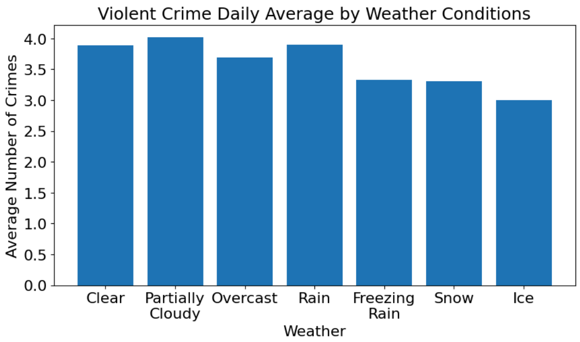
Figure 1: Number of Crimes per Year for Total, Violent, and Domestic Crimes
Figure 2: Average Chicago Unemployment Rate by Year

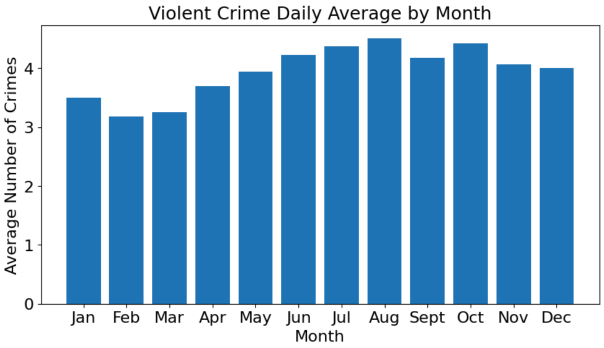




Additionally, our inquiry extends beyond socioeconomic factors to consider the influence of weather conditions. Through an examination of weather data encompassing variables such as precipitation, maximum temperature, and UV index, we aim to uncover potential associations with daily violent crime rates. Below are two plots that show the daily average violent crime by precipitation type and month.

Figure 3: Violent Crime Daily Average by Weather Conditions
Figure 4: Violent Crime Daily Average by Month of the Year





To augment our understanding of crime dynamics, we integrated data on Chicago's unemployment rates throughout the same timeframe, as seen in the graph above. Recognizing the significant correlation between economic conditions and crime prevalence highlighted in existing literature, we added the unemployment rate as an additional indicator. There were lots of indicators to choose from, but we went with unemployment rates because data for that was readily available.

Our Findings

Our exploratory analysis revealed intriguing patterns in the daily crime count. We observed that clear days consistently recorded the highest average daily crime counts, while days with poor weather tended to have lower counts. Moreover, we observed a distinct seasonal trend in daily violent crime, with fewer incidents during the early months of the year followed by a notable increase during the summer months. Interestingly, this pattern was not mirrored in domestic violent crime, which exhibited a more consistent frequency throughout the year. This divergence suggests that factors influencing domestic violence may differ from those driving general violent crime trends. As far as the yearly average for violent crimes per day, there wasn't a clear upward or downward trend over the years, but it stays between 3 and 5 crimes per day.

Methods

Feature Engineering

We applied a variety of different feature engineering methods to our datasets in order to get the most accurate results possible and account for variables that weren't directly available in the datasets. Since we were interested specifically in the number of violent crimes in Chicago we decided to sum up the reported crimes per day in order to make a variable for the count. In addition, since the crimes dataset had the dates and times of each crime reported together and the weather and holiday datasets had only the dates listed, we needed to standardize each date column into the same format. This allowed us to merge them together smoothly.

We created additional variables in our dataset as well. Some of these variables related to the date, including a new variable for month and day of the week so that we could see their effects on the crime rate. We also created binary encoded variables on whether each day was a holiday or had a full moon. Finally, we imputed missing values using the most frequent values in the column.

Exploring Models

We looked at various different models for this project. We spent our time tuning the random forest and XGBoost models more so than the others, since they seemed the most promising. Note that all models used standard scaling, simple imputing, and select percentiles between 40-60%. Table 2 shows an overview of what we tried.

Table 2: Model Description and Results

Model	*Description*	*Hyperparameters*	*Performance*
Linear Regression (Elastic Net)	Linear regression model that combines both the lasso and ridge regression methods. It learns from their shortcomings to improve the regularization of statistical models.	Alpha: 1.0 L1_ratio: 0.5	MSE: 107.503 MAE: 8.1533
AdaBoost	An ensemble method that uses weighted sums to pick the best performers. It combines weak base learners to make a strong final model.	N_estimators: 50 Max depth: 3 Learning rate: 1.0	MSE: 103.59 MAE: 7.956172
Random Forest	Combines multiple decision trees into one final model. It is good at reducing overfitting, and provides flexibility.	N_estimators: 100 Criterion: 'squared_error' Min_samples_split: 2 Max_features: 1 Bootstrap: True	MSE: 4.75 MAE: 1.73
XGBoost	Extreme Gradient Boosting is an ensemble learning method that combines several weak decision trees into a model. It has efficient regularization, predictive accuracy, and flexibility.	Polynomial features: 1 Max depth: 3 N_estimators: 98 Learning rate: 0.05	MSE: 4.37 MAE: 1.66

We also looked at the feature importance ranking from each model. Each model produced a slightly varied list of the most important features, but in general we found that solar radiation, temperature minimum, temperature feels like, unemployment, and similar features were almost always in the top five to ten best features (out of 20 features). We think that the models picked different variables in part due to the different structures of the algorithms, but also due to how little significance the predictors had on crime rate. For example, in the XGBoost model, the feature importance values were all less than 0.2, meaning the predictions would change by 0.2 without that feature in the model. We still used these weather features for our analysis since that is what the case study competition recommended, but for better results we would suggest looking into other types of factors, such as economic indicators, for predicting crime.

Model Selection

Since our data set was so large, overfitting was not a huge problem. We did see that the elastic net linear regression model was performing poorly on the validation data. We concluded a more flexible model was needed, and moved on to testing ensemble methods. The Adaboost model did slightly better, but the tree-based Random Forest and XGBoost did substantially better. The XGBoost ran faster than the random forest model and still had a comparable performance, so we selected the XGBoost model to finalize our analysis with. We used random-search cross validation to find the best parameters, which worked smoothly.

Discussion on Best Model (XGBoost)

From our initial exploration, the XGBoost model was one of the best performers. We decided to select this model for our analysis, and finalize it with more parameter tuning and cross validation.

Tuning Parameters

We explored polynomial features, n_estimators, the max depth, and the learning rate in the XGBoost model. We applied randomized-search cross validation to find the best values for the hyperparameters. Performance was measured by the mean absolute error. Our finalized parameters are as follows:

Table 3: Final XGBoost Model Hyperparameters

Hyperparameters Explored	*Description*	*Value*
Polynomial features	Captures the degree of non-linearity	3
N_estimators	The number of trees in the mode	98
Learning rate	Controls how much each tree contributes	0.05
Max depth	Controls the depth of the trees in the model	3

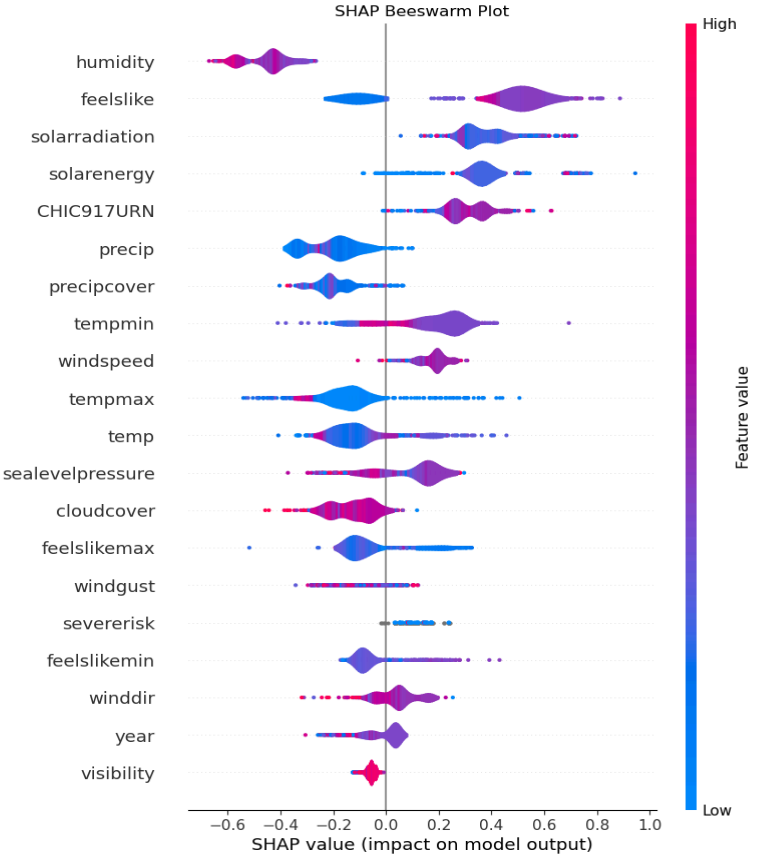
Performance Metrics

By applying the new values for the hyperparameters, the MSE dropped from 5.01 to 4.37. The MAE also dropped from 1.82 to 1.66. This means that by using mostly weather patterns we can predict the number of crimes daily happening in Chicago with an error of less than 2 crimes per day.

Feature Importance

In addition to the importance rankings from the models that we mentioned earlier, we used SHAP to explore which features contributed most to our XGBoost predictions. Figure 5 shows the impact of each variable. We found that humidity, feels like temperature, and solar radiation were among the variables with the biggest impact. Feels like temperature and solar radiation were also among the highest in XGBoost's feature importance method. The slight differences in top features come from how each model measured the importance. We can also see that even though some of the weather variables were better predictors overall, our economic indicator, CHIC917URN, had one of the higher impacts on our model.

Figure 5: SHAP Feature Beeswarm Plot



Conclusion

Our top-performing model emerged as the XGBoost model, utilizing the hyperparameters outlined in Table 3. Demonstrating superior performance in both Mean Absolute Error (MAE) and Mean Squared Error (MSE) testing, this model excelled in forecasting the incidence of violent crimes on specific days. Some key features were identified as Weather Conditions, Min Feels Like, UV index, and Solar Energy. Although the Random Forest model also showed promise, we opted to prioritize the XGBoost model due to its improved performance and faster runtime.

Noteworthy findings include the observation that clear weather days witness a higher occurrence of violent crimes compared to adverse weather conditions. Additionally, the data reveals a seasonal trend, with the lowest average violent crime counts observed in the first few months of the year, peaking during the summer months. On New Year's Day, the average violent crime count spikes to 8.8, significantly surpassing the general average of 3-4 crimes.

Based on these insights, we recommend the Chicago Police Department allocate additional resources for patrolling during New Year’s Day and possibly during the summer months. Moreover, the relatively low crime rates during the early months of the year present an opportune time for implementing and testing new policies or patrolling routes. Furthermore, since domestic crime appears unaffected by weather patterns, we should combat rising trends through raising awareness about this issue.

Our analysis drew insights from weather and unemployment rate data, but we acknowledge the potential of additional indicators such as poverty levels, demographic information, or the proportion of single-parent households in the area that could be more significant in predicting crime rates. Additionally, utilizing a Time Series model could enhance predictive capabilities by capturing year-to-year and seasonal trends in crime data.

References

Crimes - 2022: City of chicago: Data Portal. Chicago. https://data.cityofchicago.org/Public-Safety/Crimes-2022/9hwr-2zxp/data

VC Corporation. *Total weather data -history & forecast data in CSV or JSON*. Historical Weather Data & Weather Forecast Data | Visual Crossing. https://www.visualcrossing.com/weather-data

Unemployment rate in Chicago-naperville-elgin, IL-in-wi (MSA). FRED. (2024, April 3). https://fred.stlouisfed.org/series/CHIC917URN

Mi, Joyce (2019). *Variable Selection Methods with Applications to Crime Predictions.* [Thesis, California State Polytechnic University]. https://scholarworks.calstate.edu/downloads/z029p6976#:~:text=Variables%20which%20were%20associated%20with,poverty%2C%20ethnicity%20and%20family%20structure.