Time-Series Forecasting of Crime in Los Angeles, California

Aryan Patidar
Data Science Institute, Brown University

GitHub Repository: *github.com/apatidar/data-1030-project*

---

1 Introduction

1.1 Background and Motivation

Los Angeles, the second most-populous city in the country, is known for it's vast size and diversity, which contribute to a broad spectrum of urban crime. These characteristics make it ideal for studies in crime prediction and prevention. This project is motivated by the need to understand historical crime trends and regression techniques to forecast future criminal activity as this issue has not been previously addressed. By predicting incidents of crime, this analysis aims to provide actionable insights to guide law enforcement agencies in efficiently allocating resources and enhancing public safety.

1.2 Description of Dataset

The backbone of this study is a dataset sourced from Kaggle and originally derived from the Open Data repository maintained by the city. It spans incidents from January 2010 to April 2023, encompassing over 2 million data points. This data was compiled from manually-transcribed crime reports, leading to a significant presence of missing values. Although there are a wide array of features, we will focus on those that can serve as predictors in our model. This selection is guided by the data's inherent characteristics and the goal of developing an insightful forecast model for criminal activity in Los Angeles.

2 Exploratory Data Analysis

2.1 Visualization of Target Variable

Our primary goal is to predict future crime rates based on historical data, positioning the total number of crimes as our target variable. We first examine the distribution of crime types in Los Angeles, helping us understand the prevalence of various crimes. Figure 1 highlights the most common crimes; although the dataset contains 144 types, only 15 account for a significant proportion of incidents. We observe a drop in frequency to "Grand Theft from Vehicle," followed by a gradual decline, which suggests a potential focus area for the

predictive model, indicating that prioritizing these common crime types could enhance the effectiveness of law enforcement strategies.

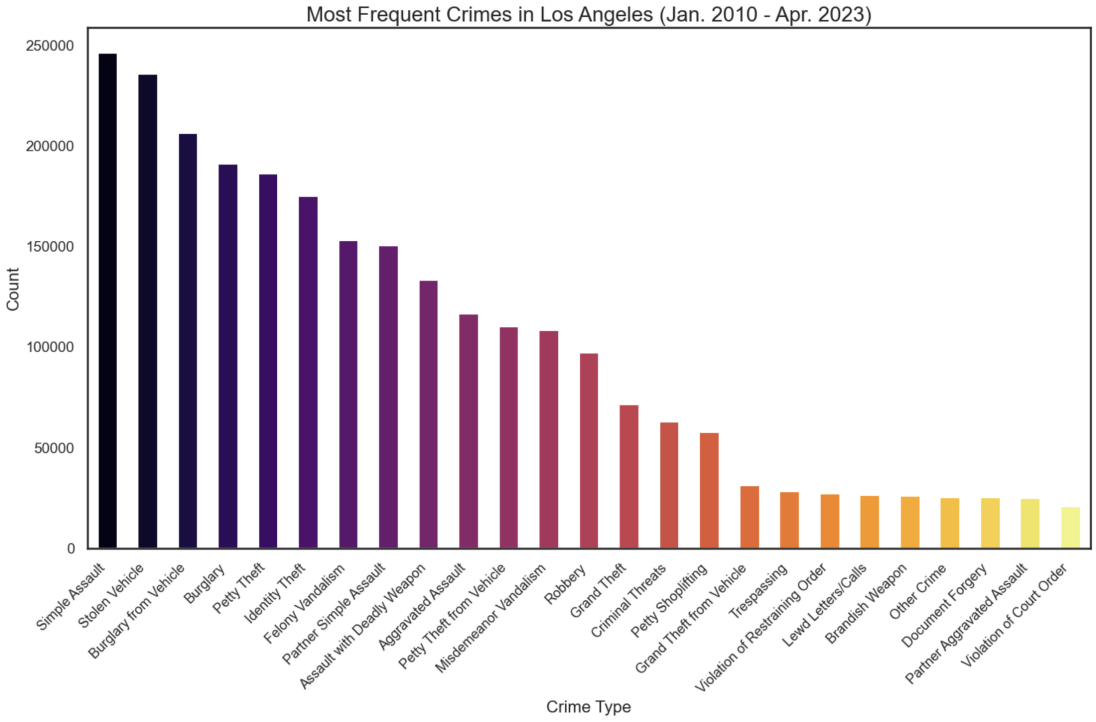

Figure 1: Most common crime types in Los Angeles

Another critical aspect to consider is potential seasonality in total crime numbers over the years, introducing unnecessary noise into our model. Figure 2 illustrates crime occurrences over the specified period, with a steep drop toward the end of the range due to incomplete data collection. Excluding the anomaly observed during the pandemic time, a consistent yearly fluctuation pattern is evident. This indicates the presence of seasonality, which must be addressed when preprocessing to ensure the reliability of the models.

Figure 2: Crime in Los Angeles over time

2.2 Missing Values

The dataset, compiled from manually-transcribed crime reports, naturally contained several missing values. This was partially alleviated by dropping columns that are irrelevant to the analysis. For predictors like "Victim Race" and "Victim Sex" where missing values were minimal, rows with gaps were removed to avoid introducing inaccuracy through imputation. The dataset was also refined to focus on the top-15 most common crimes. Over 1.78 million data points remained after these adjustments, offering a substantial basis for analysis while ensuring data integrity.

3 Methods

3.1 Lag Features

The next phase involves aggregating the data to facilitate the prediction of future crime totals. To achieve this, we must decide the appropriate time interval for aggregation, along with the determination of the optimal number of lag features to be included.

Various time intervals were considered, including daily, weekly, monthly, and quarterly periods. After aggregating at each interval, autocorrelation plots were generated and autoregressive models were fitted. These models assessed how the number of crimes is correlated with itself over different time lags. However, this revealed the challenge of non-stationary data across all periods. The autocorrelation plots also displayed periodic peaks, suggesting the presence of seasonality. To counter these issues, differencing was employed to subtract previous values from current ones, rendering each time-series stationary, as confirmed by an ADF test. Post-differencing, the autocorrelation plots exhibited the anticipated decreasing trend as lags increase, as displayed in Figure 3.

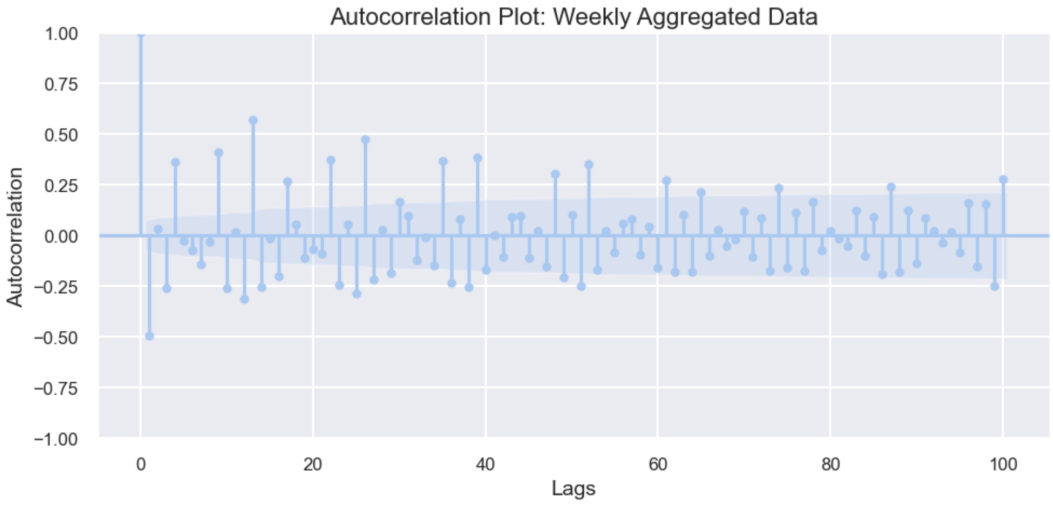

Figure 3: Autocorrelation of data aggregated weekly with 100 lags

After inspecting the autoregressive model outputs, we aggregated at the weekly level to bal-

ance granularity and model stability. The larger intervals aggregated too much data, leading to a scarcity of points for a robust analysis and a lack of significant lag features. Conversely, daily aggregation proved too granular, introducing high volatility. From a practical standpoint, weekly forecasts are more feasible for resource allocation by law enforcement, as daily adjustments are impractical on a large scale.

3.2 Data Splitting and Preprocessing

After adding four statistically significant lag features, the preprocessing phase addressed the predictors: "Area," "Victim Sex," "Victim Race," "Weapon Use," and "Victim Age." The first four were subjected to one-hot encoding as categorical features, while the last was standardized via standard scaling as a numerical feature. The total number of features rose to 46. The dataset was also split in the ML pipeline, employing "TimeSeriesSplit" cross-validation with five splits to ensure robust model training and validation of time-series data.

3.3 ML Algorithms and Evaluation

A simple linear regression model was first fitted to the data, utilizing the preprocessing and splitting pipeline above. We then introduced regularization techniques through Lasso and Ridge regressions. In both models, the alpha hyperparameter controlling the strength of the regularization was tuned between 0.0001 to 100.

The analysis progressed to more sophisticated, nonlinear models. The Random Forest model, known for versatility and robustness, included hyperparameters like 'max depth,' ranging from None for unrestricted tree growth to 100, and 'max features,' varying from None to consider all features to 1.0 for some features. Support Vector Regression varied the 'C' hyperparameter, determining regularization strength from 0.1 to 10, and the 'gamma' hyperparameter affecting the influence of individual training samples from 0.001 to 100,000. Finally, K-Nearest Neighbors was adjusted for the number of neighbors from 1 to 11 and weighted between uniform and distance metrics to evaluate the impact of neighbor proximity on predictions.

Model performance was evaluated using root mean square error, a suitable metric for its sensitivity to the magnitude of errors and penalization of larger errors. This is critical in the context of crime prediction as larger prediction errors can have worse implications than smaller ones. RMSE's units are the same as the predicted variable, simplifying interpretability, and it can be consistently compared across various models. The uncertainty in this metric was measured by changing the random state parameter. These changes did not significantly impact RMSE: stability that can be attributed to the use of "TimeSeriesSplit" for cross-validation. This approach sequentially splits time-ordered data, mitigating the effects of random state variations. Furthermore, GridSearchCV's systematic approach of tuning hyperparameter combinations for optimal performance likely had a more pronounced influence than any random element introduced.

4 Results

4.1 Model Performance

As a benchmark for model performance, a baseline score was determined by calculating the RMSE if the model predicted the average value of the target variable for every instance, without considering any input features. If a model fails to surpass the baseline, it indicates its inability to capture and interpret underlying patterns in the dataset.

	RMSE Test Score		
	Mean	Standard Deviation	Std. Deviations from Baseline
Baseline	236.71	–	–
Linear Regression	223.49	87.18	+0.15
Lasso Regression	220.15	87.39	+0.19
Ridge Regression	221.64	87.65	+0.17
Random Forest Regression	263.58	74.34	−0.36
Support Vector Regression	221.87	86.85	+0.17
K-Nearest Neighbors Regression	241.50	76.58	−0.06

Figure 4: RMSE test scores of different algorithms

As shown in Figure 4, the linear models generally experienced higher performance than the nonlinear models. All linear algorithms outperformed the baseline, with Lasso performing strongest at +0.19 standard deviations. This indicates that the regularization effect of Lasso, which can shrink some coefficients to zero, was effective. On the other hand, Random Forest struggled with a significant deviation below the baseline. This suggests the model may be overfitting to the training data or not generalizing to the unseen data, which is surprising given the model's typical strength in handling complex datasets. Figure 5 visualizes the scores. KNN was slightly below the baseline, indicating the reliance on the proximity of similar cases did not capture the patterns for crime forecasting, likely due to high dimensionality or noise in the data. All the mean scores are logical as they represent a reasonable number of crime counts, with the RMSE metrics in the same unit as the prediction variable.

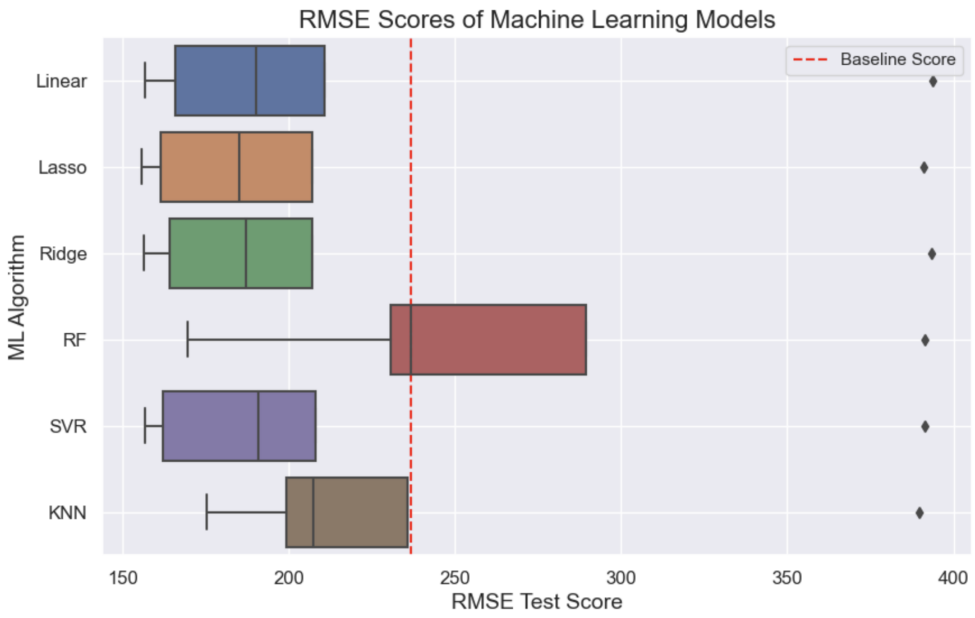

Figure 5: Visualization of model performance

Further inspection was conducted by plotting the residuals as a function of time for each model. Every model experienced significant residual variance near the end, signifying a possible change in the underlying crime patterns or an external influence not accounted for by the models. This could reflect real-world events or shifts in reporting practices that the models fail to capture. Figure 6 illustrates these residual plots, showcasing the roughly similar trend in the predictions made by the best and worst-performing models.

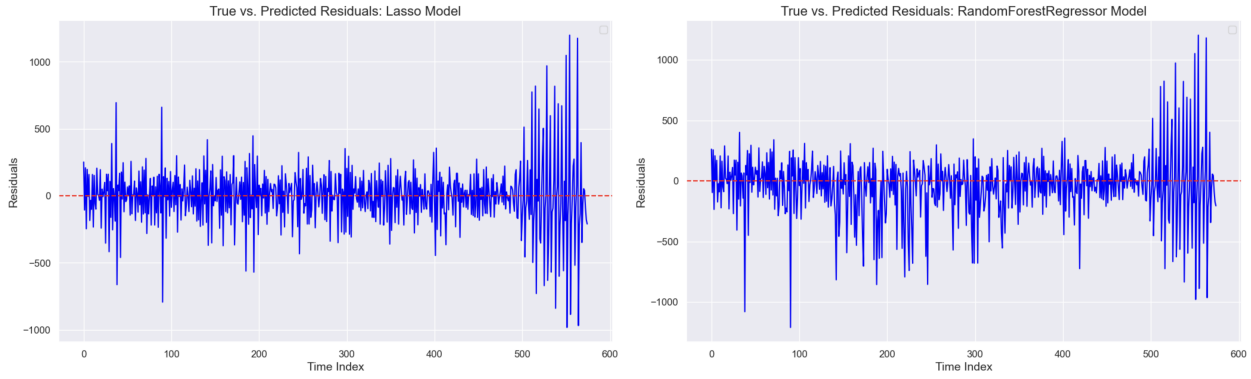

Figure 6: Residual plots for Lasso and RF models

4.2 Global Feature Importance

Global feature importance was assessed using the metrics provided by XGBoost, which is particularly suitable for time-series forecasting. We primarily concentrated on the 'gain' metric, which quantifies the contribution of each feature to the model's predictive power. 'Weight' was also taken into account, reflecting the frequency of a feature's use in making

splits within the model. Features with higher 'weight' values are indicative of a broader influence on the model's predictive outcomes.

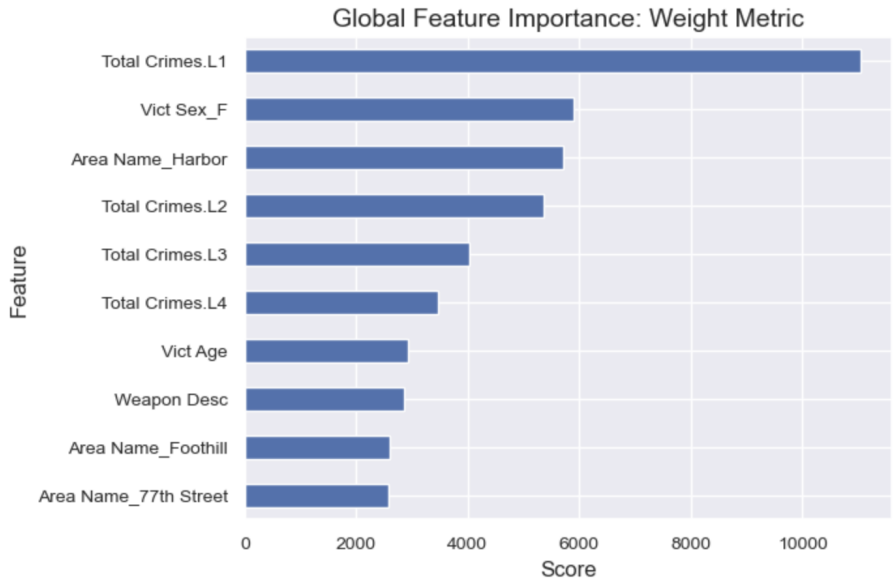

Figure 7: Feature importance by weight

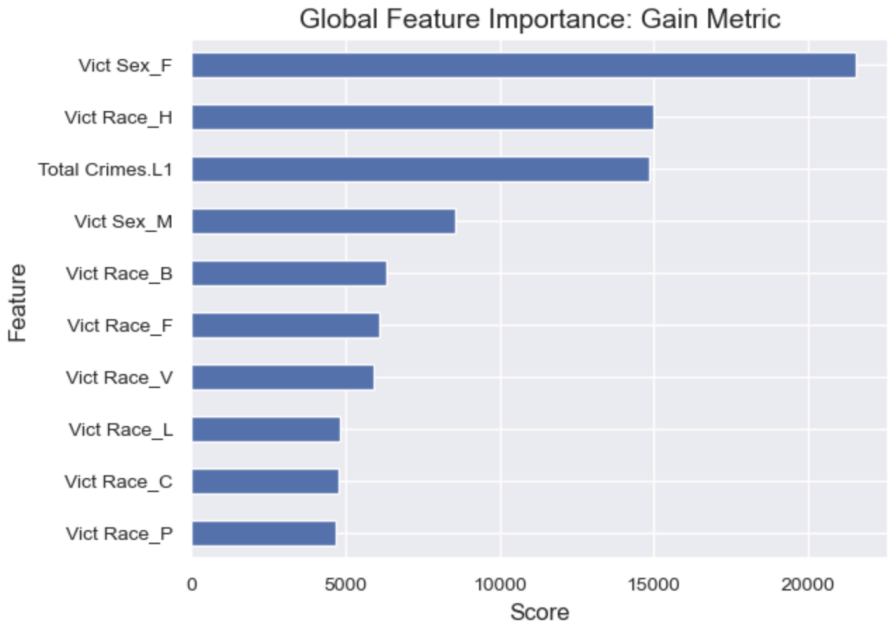

Figure 8: Feature importance by gain

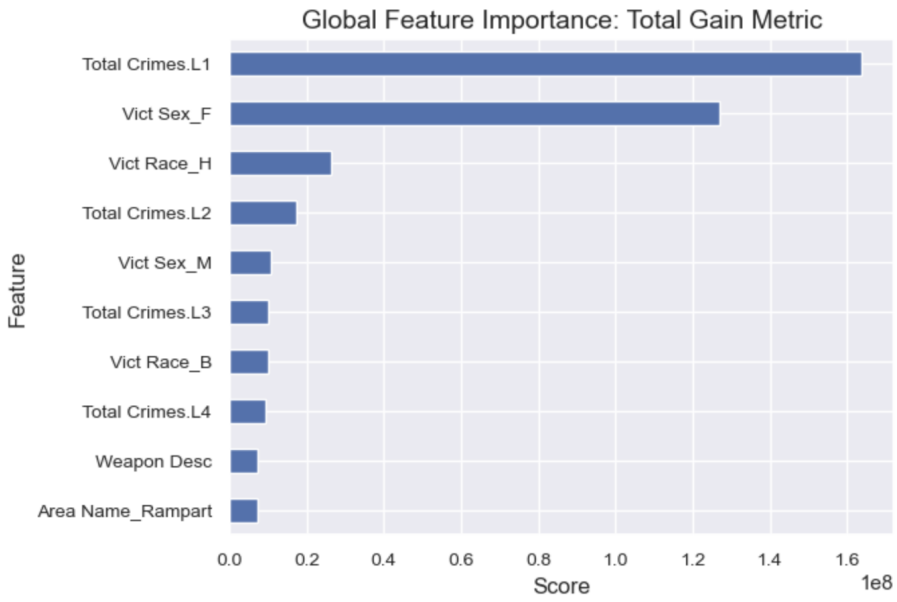

Figure 9: Feature importance by total gain

Figure 7 reveals that the lag features were frequently used for making splits, underscoring their importance to the model's decision-making process. In contrast, Figure 8 indicates that features relating to the sex and race of the victim contribute significantly to increasing the model's predictive accuracy. Overall, Figure 9 consolidates these findings, underscoring the first lag feature as having the highest overall contribution to model performance and suggesting that recent crime history is the strongest predictor of future incidents. Victims who identify as female and Hispanic also appear to be important whereas area names are less impactful according to 'weight' and 'total gain,' indicating that location may not carry as much predictive weight as temporal or demographic factors.

4.3 Local Feature Importance

Figure 10 displays the SHAP values, with a more pronounced dominance of geographic features than the global feature importance plots. One interesting observation is that 77th Street appears to be strongly associated with lower crime predictions; however, in Figure 11, we can see that it experiences the most crime in the entire city. The greater presence of geographic features in the SHAP plot might reflect localized patterns that do not generalize well across the dataset or are overshadowed by stronger, more consistent predictors.

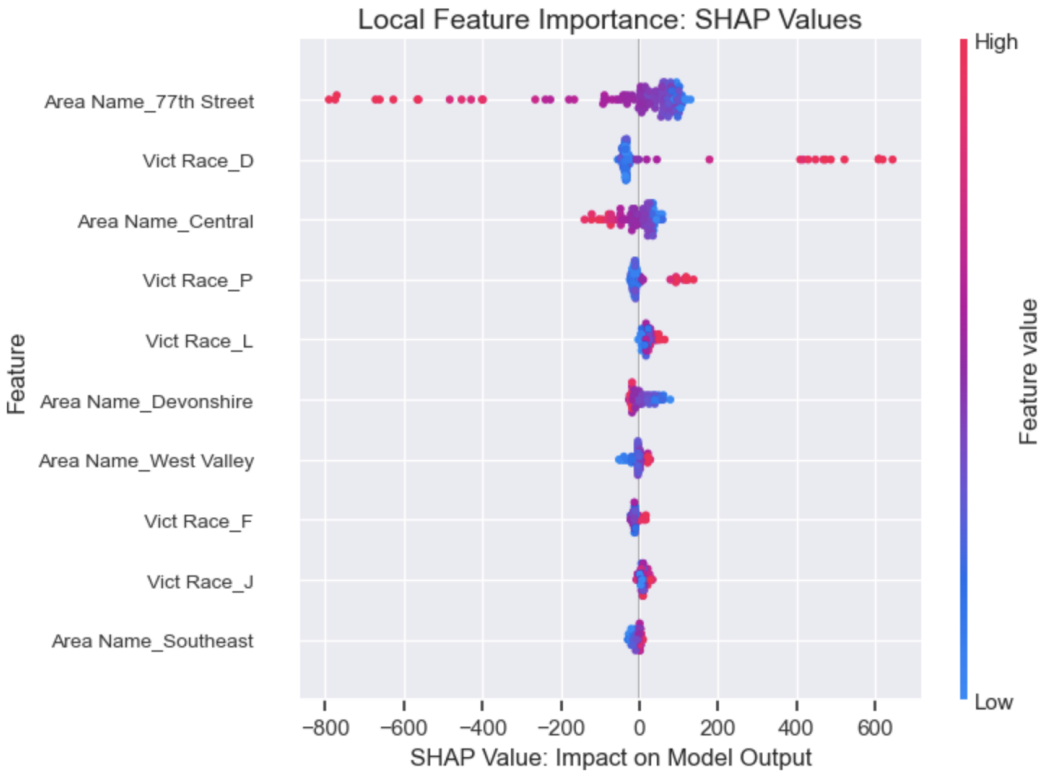

Figure 10: SHAP values for local feature importance

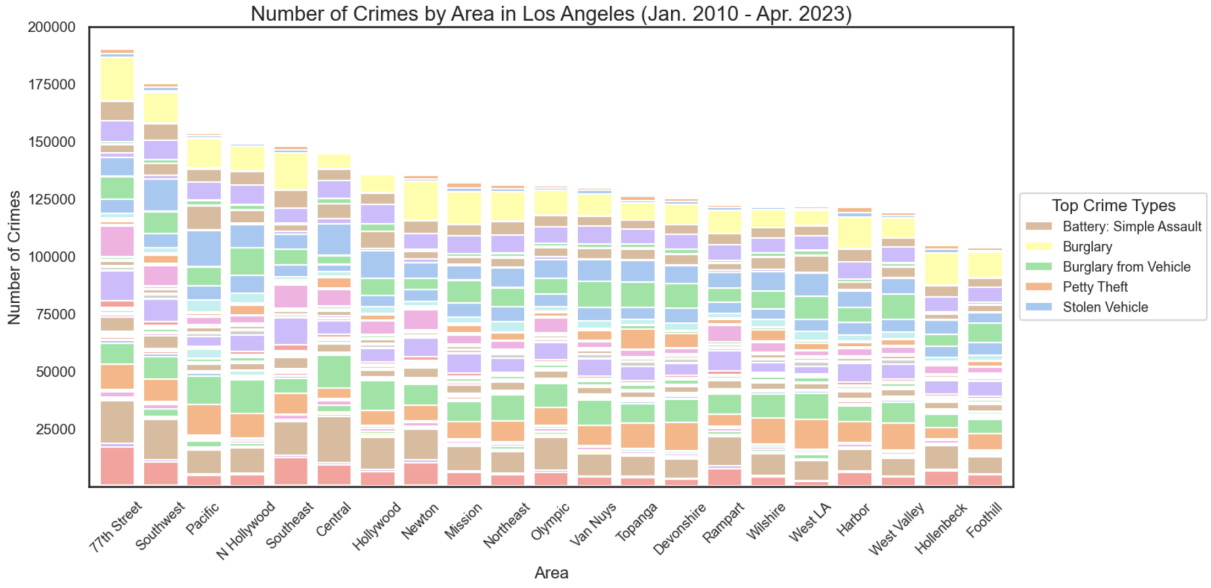

Figure 11: Geographic distribution of crime in Los Angeles

5 Outlook

The Lasso model's robust performance highlights its potential as a valuable tool for law enforcement agencies to forecast crime trends. Its ability to simplify complexity by eliminating non-contributing features through regularization stands out as an effective method in this context.

Looking ahead, several enhancements could bolster the predictive capability of the model. One critical concern to consider is potential data leakage. Current preprocessing involves aggregating variables like victim sex and race, which are not determined until the time of a crime and may inadvertently introduce bias. Identifying and integrating additional predictors could mitigate this issue, and the inclusion of new features could also provide a more nuanced understanding of the underlying trends and seasonality affecting crime rates. Furthermore, the notable peaks in residual plots towards the end of each time-series necessitate further investigation. These deviations could be indicative of changes in crime patterns or external factors not currently captured by the model, such as evolving social conditions or alterations in crime reporting mechanisms. Evaluating outliers that could skew the performance of the model is another essential step. This would aid in refining the predictive accuracy. On the other hand, the elimination of noisy or irrelevant features could streamline the model, focusing on the most informative predictors and enhancing the overall efficiency and interpretability.

Word Count: 1989

6 References

1. Chaitanya, Krishna Kasaraneni. "Crime Data in LA City From Beginning of 2010 to Present." *Kaggle*.

2. "Crime Data from 2010 to 2019." *Los Angeles Open Data*, City of Los Angeles.

3. "Crime Data from 2020 to Present." *Los Angeles Open Data*, City of Los Angeles.

4. "The 300 Largest Cities in the United States by Population 2023." *U.S. Cities*, World Population Review.