Predicting Future Drug Trafficking Hotspots in Colombia

Natalia Espinosa Dice
Advisor: Xiaoyan Li
January 9th, 2025

Abstract

Colombia plays a central role in the global cocaine trade, making the identification of drug trafficking hotspots vital for disrupting supply chains and shaping effective policy. Despite its real-world significance, machine learning approaches have not been widely applied to this problem. This paper addresses that gap by leveraging regression models to predict trafficking hotspots across Colombia's 32 departments and capital district, using time-series data from the United Nations Individual Drug Seizures Database. The best-performing model, XGBoost, achieves a mean squared error of 3.288 on the log scale. Feature importance analysis highlights key factors such as crime rates, government anti-narcotics operations and urban-rural population distributions. Engineered time-dependent features - such as exponential moving averages and rolling statistics - are particularly important for capturing trends, though sudden shifts in activity remain challenging to predict. Ultimately, this paper underscores the value of combining rich socioeconomic data with advanced feature engineering to model drug trafficking, offering key insights into its dynamics and laying a strong foundation for future research.

1. Introduction

1.1. Motivation

Drug trafficking is a pervasive global issue, posing significant threats to public health, security and governance. In 2022, the global supply of cocaine reached an all-time high of over 2,700 tons, reflecting a 20% increase from the previous year.[20] Colombia plays a central role in this crisis, accounting for 65% of global coca bush cultivation in 2022, covering an estimated

230,000 hectares.[19] Despite intensified coca eradication efforts, cocaine seizures continue to rise, underscoring the resilience of trafficking organizations.[19] Addressing trafficking in Colombia is therefore critical not only for national security but also for global efforts to curtail the cocaine trade.

Beyond identifying trafficking hotspots, this study also explores the socioeconomic conditions associated with drug trafficking, offering insights into the structural vulnerabilities that trafficking organizations exploit. High levels of poverty and unemployment, for instance, are believed to facilitate the recruitment of ordinary citizens into criminal activities.[8] Trafficking has also been linked to high levels of violence, providing "elements that facilitate the lethality of violence" including access to weapons, training of personnel to use lethal violence and reliance on intimidation tactics to control populations.[8] Finally, with access to vast financial resources, trafficking organizations can facilitate the corruption of governments, undermining institutional efficiency and justice systems.[15] Examining these socioeconomic dynamics is a critical question within political science and provides an important framework for developing predictive models.

1.2. Goal

This paper employs regression models to address two primary goals: (1) to identify trafficking hotspots and (2) to explore the socioeconomic conditions most correlated with drug trafficking. Existing research in related domains serves as a foundation for this study. Cipriano et al. examined the determinants of illegal coca production in Peru, focusing primarily on the influence of government policies, such as eradication efforts.[2] Zuckerman Daly analyzed the conditions driving organized violence in Colombia, adopting a subnational approach but focusing on predicting violence rather than drug trafficking.[24] Bazzi et. al used machine learning to predict violence in Colombia and Indonesia, highlighting the distinct challenges of modeling highly fluctuating time-series data.[1]

Building upon these studies, this paper is novel in four distinct ways. First, it shifts focus to predicting drug trafficking levels by employing regression models and leveraging the United Nation's Individual Drug Seizures Database for the years 2012-2022. Secondly, following Zuckerman Daly's work, it adopts a subnational lens, analyzing Colombia's 33 administrative departments (including

its district capital) rather than viewing the country as a single unit. Third, it dramatically expands the scope of socioeconomic features considered by aggregating data from five distinct databases, thus developing a novel and comprehensive dataset. Finally, to address the complex time-series prediction task at hand, it leverages robust feature engineering to account for temporal dependencies and improve the performance of Linear Regression, Random Forest and XGBoost.

This paper evaluates model performance using Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-squared (R^2). The baseline linear regression model established an MSE of 6.01 on the log scale and an R^2 of 0.42, while the best-performing developed model, XGBoost, achieved an MSE of 3.288 on the log scale and an R^2 of 0.68. Important socioeconomic features include homicide rates, historical government enforcement activity and urban and rural populations. Key engineered features include exponential moving averages (EMAs) alongside rolling minimum, maximum and average values. These findings underscore the potential of combining rigorous feature engineering with comprehensive socioeconomic data to capture complex patterns in drug trafficking patterns. The remainder of this paper is structured as follows: Section 2 reviews related work and identifies gaps in existing methodologies; Section 3 outlines the study's rationale and novel methods; Section 4 details the implementation process; Section 5 evaluates model performance; and Section 6 discusses conclusions, limitations and pathways for future research.

2. Background and Related Work

While predicting future drug trafficking levels based on the UN Individual Drug Seizures Database is a novel task, this paper draws on an abundance of previous work to inform its approach. Existing literature contributes in two key areas: (1) identifying factors associated with trafficking and (2) outlining methodologies for predictive modeling.

2.1. Identifying Factors Associated with Drug Trafficking

Existing research offers valuable insights into the socioeconomic, geographic and governance-related conditions that shape trafficking patterns. For instance, Jiménez-García et al. explore the relationship between drug trafficking, violence and socioeconomic vulnerabilities in Pereira,

Colombia.[8] The study aggregates government statistics from 2010 to 2019, drawing on data from the Colombian National Police, the National Planning Department and the Mayor's Office of Pereira. Using regression-based supervised learning models, it identifies strong correlations between violence, poverty and trafficking.[8] However, their analysis is limited to a single city, restricting its generalizability to broader regional or national contexts. Furthermore, they focus primarily on correlation rather than prediction. I address these limitations by adopting a predictive modeling framework that generalizes across all administrative departments and incorporates temporal features to capture dynamic changes in trafficking patterns. Despite these limitations, Jiménez-García et. al's findings underscore the relevance of crime, poverty and governance indicators, which I therefore include as features in this paper.

Several other studies also provide valuable insights into the factors associated with drug trafficking, though they do not directly address the task of prediction, and therefore inform this paper's data collection. For example, Singer explores the societal impacts of drug trafficking, including its role in fostering violence, corruption and declining trust in government institutions.[15] While Singer does not propose methodologies for prediction, the emphasis on corruption as a key enabler of trafficking highlights the importance of incorporating corruption and governance indicators as features in this paper's predictive framework. Similarly, Thoumi examines the structural conditions that facilitate drug economies in Colombia, emphasizing the role of weak governance, economic crises and unemployment as enabling conditions for illicit drug industries.[17] Thoumi also highlights the role of trafficking routes and geography, through which coastal and border regions emerge as high-risk areas.[17] Indeed, the United Nations 2024 World Drug Report also notes a shift toward maritime trafficking in recent years, with more than 80% of cocaine shipments directed toward the coast.[20] On land, it illustrates that trafficking routes span northward towards Venezuela and the Caribbean, eastward toward Brazil and southwest toward Ecuador.[20] These trends confirm that departments along Colombia's coast or borders experience heightened trafficking activity, thus justifying the inclusion of both a border and coastline department feature in this paper's framework. While the UN report provides a global overview of trafficking trends, it lacks specific regional

analyses within Colombia. I address this gap by instead focusing on department-level patterns to provide a more nuanced understanding of trafficking dynamics.

Finally, Saab and Taylor document the extensive historical involvement of both the Revolutionary Armed Forces of Colombia (FARC) and paramilitary groups in drug trafficking.[14] Although the 2016 peace accords have since reduced their influence, the historical presence of these groups may have established social infrastructures that continue to shape trafficking patterns today. I thus incorporate features that account for historical FARC and paramilitary activity to ensure these enduring influences are captured in the predictive model.

Together, these studies provide critical insights into the factors driving drug trafficking, informing this paper's data collection and scope. By integrating socioeconomic, geographic and governance-related variables, I thus shift focus to building a predictive model to empirically test these relationships.

2.2. Outlining Methodologies for Predictive Modeling

Given the lack of existing literature on modeling drug trafficking specifically, I draw heavily on studies that model related political phenomena, such as violence and illegal crop cultivation.

Cipriano et. al offers perhaps the closest parallel by analyzing the determinants of illegal coca production in Peru.[2] This study uses the Peruvian government's National Commission for Development and Life without Drugs (DEVIDA) Database, which includes extensive data on government eradication efforts and coca cultivation levels. It applies Lasso regression, Ordinary Least Squares and Vector Autoregression to assess the impact of government enforcement policies. Its findings reveal a weak negative relationship between eradication efforts and coca cultivation and a positive correlation between coca base paste confiscations and coca cultivation, suggesting that enforcement policies alone may be insufficient to deter illicit activity.[2] While Cipriano et. al's work focuses narrowly on government enforcement policies, I expand the scope of structural factors examined by also incorporating economic, governance and crime-related indicators into its predictive framework. Furthermore, whereas Cipriano et. al models coca *cultivation*, a more stationary

phenomenon, I predict *trafficking* volumes, which are inherently more dynamic and susceptible to fluctuations over time and space. Addressing these complexities thus requires advanced feature engineering to incorporate temporal dependencies alongside socioeconomic trends.

Zuckerman Daly investigates the conditions favoring organized violence at the subnational level on Colombia.[24] Using a dataset of 274,428 municipality-month observations, she applies regression techniques and incorporates spatial lag features to capture how prior violence influences future outbreaks. The study finds that areas with strong pre-existing organizational structures and histories of past mobilization are more prone to persistent violence.[24] A key takeaway of Zuckerman Daly's work is its subnational lens, which emphasizes the importance of analyzing variations at the municipality level rather than treating Colombia as a homogeneous entity. This approach acknowledges that political, economic and social factors vary substantially across regions, and that patterns of violence often emerge in pockets rather than uniformly across the state.[24] Since drug trafficking operates under similarly localized conditions, I adopt Zuckerman Daly's subnational approach. However, while Zuckerman Daly employs binary classification for violence, I predict trafficking levels on a continuous scale, necessitating richer feature engineering of time-dependent variables and an expanded scope of socioeconomic features.

Finally, Bazzi et al. examine the use of machine learning to predict violence in Colombia and Indonesia.[1] For Colombia, the study draws on violence data from 1988 to 2005 provided by the Conflict Analysis Resource Center and combines it with socioeconomic data, including population density, government revenues, military presence and geographic features. Testing a range of machine learning algorithms - including Lasso regression, random forests and neural networks - the authors predict violence hotspots one year ahead based on historical patterns and socioeconomic variables. The study finds that machine learning models are effective at identifying persistent hotspots of violence but face challenges in forecasting sudden outbreaks or escalations.[1] The inclusion of lagged dependent variables, socioeconomic features and geographic characteristics, particularly terrain ruggedness, improves predictive accuracy, highlighting the importance of accounting for both historical patterns and regional characteristics. However, the study also underscores the difficulty of

predicting abrupt deviations, reflecting the limitations of current modeling techniques.[1] I adapt Bazzy et al.'s methodology, shifting focus to predicting drug trafficking patterns rather than violence. I use extensive subnational socioeconomic data, tailoring Bazzy et. al's approach to the specific dynamics of trafficking by incorporating additional features related to violence, crime, corruption and governance. Directly responding to the identified challenge of forecasting sudden changes in time-series data, I employ extensive feature engineering to expand the range of time-dependent variables incorporated and enhance predictive accuracy.

Together, these studies provide a strong methodological foundation for this paper's approach. They underscore the value of adopting a subnational lens to capture localized patterns and highlight the potential of machine learning techniques to model complex social dynamics, including violence and illicit crop cultivation. By building on their frameworks, this paper advances predictive modeling techniques for drug trafficking, addressing critical gaps in prior research and tailoring methodologies to the unique dynamics of trafficking in Colombia. In doing so, I also hope to advance predictive modeling techniques for complex time-series data more broadly, in particular by testing the power of engineered time-dependent features.

3. Approach/Methods

3.1. Focus and Lens

This paper builds on methodologies from related fields, such as violence and coca cultivation modeling, and shifts focus to predicting drug trafficking levels - an area that remains largely unexplored. To achieve this, I adopt a subnational lens, analyzing Colombia's 33 administrative departments. This approach recognizes that drug trafficking operates as a network of localized activities, shaped by regional conditions, governance structures and geographic factors, rather than uniformly across the country.

Much of the existing research, including prominent studies by organizations such as the United Nations, treats Colombia as a single entity when analyzing drug trafficking trends. While this national-level perspective is valuable for understanding overarching patterns, it risks obscuring

critical regional variations. Colombia's diverse geography - from remote rural areas to densely populated urban centers - and disparities in socioeconomic development and governance create distinct regional vulnerabilities to trafficking. By examining subnational data, this study allows for the identification of regional patterns that may be masked by national-level aggregation, thus offering a more granular understanding of trafficking dynamics.

3.2. Aggregated Socioeconomic Dataset

A major novelty of this paper lies in its expanded socioeconomic feature set, which integrates data from five distinct datasets to compile a comprehensive range of structural factors relevant to drug trafficking dynamics. These features encompass violence indicators such as events of homicide, displacement, threats, extortion and sexual crimes; economic conditions such as income inequality and poverty levels; and governance metrics such as corruption and revenue. Education and public health variables - including life expectancy, average years of education and infant mortality rates - are also included to capture broader measures of public health and development. Additionally, population statistics - such as urban and rural population densities - are incorporated to account for differences in infrastructure and accessibility that may influence trafficking routes and hubs. Geographical features further enrich the dataset, identifying departments located along coastlines or international borders to capture risk factors associated with maritime and cross-border trafficking routes. Recognizing historical context, this paper also includes variables indicating the presence of paramilitary groups and the FARC, reflecting areas that have experienced prior armed conflict and may retain structural vulnerabilities to organized crime. Similarly, data on historical government-sponsored anti-narcotics operations is also included. In total, the dataset incorporates over 50 socioeconomic variables, drawing on existing literature to encompass a wide range of factors believed to influence trafficking patterns. By combining diverse data sources and capturing both socioeconomic and geospatial dynamics, this dataset aims to improve predictive accuracy while offering deeper insights into the conditions that sustain trafficking networks. Moreover, this approach addresses gaps in prior studies that relied on narrower feature sets, enabling a more holistic

modeling framework that integrates economic, social, political and spatial dimensions of trafficking activity.

3.3. Rigorous Feature Engineering

Drug trafficking is an inherently complex and dynamic phenomenon, characterized by sharp fluctuations and sudden changes that make forecasting particularly challenging. Prior studies in related domains, such as Bazzi et al.[1], have highlighted the difficulty of modeling abrupt shifts in time-series data and often rely solely on simple lagged dependent variables. While effective for capturing basic temporal patterns, such approaches may fail to account for the full range of variability and volatility present in complex time-series data. Thus, this paper significantly expands the use of time-dependent features. In addition to lagged features, it incorporates rolling averages and exponential moving averages to smooth short-term fluctuations; rolling maximum, minimum, range and standard deviation values to quantify variability and volatility over time; cumulative sums to capture long-term trends; growth rates and momentum indicators to reflect sudden changes; and seasonal indicators to capture cyclical trends in trafficking levels. In total, over 30 time-dependent features are engineered from the dependent variable, allowing the models to track both gradual trends and abrupt shifts in an effort to address the forecasting challenges identified by Bazzi et al.[1]

3.4. Model Training and Evaluation Overview

To predict trafficking levels, this paper develops three regression models: Linear Regression, Random Forest and XGBoost. A simple linear regression model serves as the baseline model and excludes the engineered time-dependent features, providing a useful benchmark for assessing the added value of incorporating more advanced features and techniques. I then leverage feature engineering, feature selection, regularization and hyperparameter tuning to develop three more complex models with increased predictive power: Ridge Regression, Random Forest and XGBoost. This choice of models provides several benefits. First, the Ridge Regression model offers a direct comparison against the baseline, testing the impact of adding engineered features and regularization. Random Forest was selected for its ability to handle non-linear relationships and high-dimensional data, thus potentially

allowing more features to be included without risking overfitting. XGBoost is also highly effective for complex, non-linear patterns, thus making it another suitable choice for the task at hand. Finally, all three models offer insights into feature importance, allowing for more interpretable results.

To evaluate model performance, I utilize three key metrics, which are described in more detail in Section 4.7: MSE, RMSE and R^2. Beyond these metrics, residual plots are also analyzed to detect systematic errors. I conduct feature importance analysis to determine both the socioeconomic factors most correlated with drug trafficking and the engineered features most useful in forecasting complex time-series data. I also employ department-level error analysis, examining regional variations to determine where the model is most and least successful.

4. Implementation

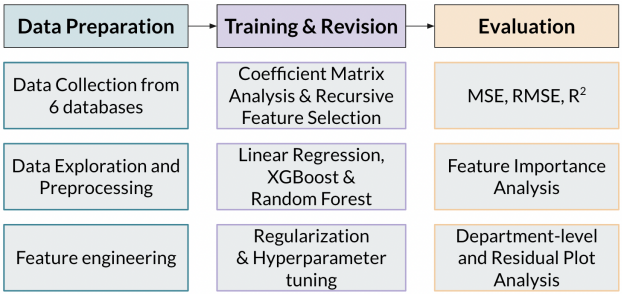

Figure 1: Flowchart illustrating the key implementation steps

Figure 1 provides an overview of the implementation steps spanning data preparation, model training and evaluation, which are detailed in the following subsections.

4.1. Data Collection:

4.1.1. Dependent Variable: The dependent variable in this study - quantity of cocaine seized - is gathered from the UN Individual Drug Seizures Database for the years 2012–2022.[18] This database provides detailed, standardized information on drug seizures, including quantity, substance type, date and geographic location. By capturing actual interdiction events, it offers a direct proxy for trafficking activity at specific times and places. While seizure data inherently reflects law

enforcement activity and may be biased toward areas with stronger enforcement, it remains one of the most comprehensive and consistent datasets for estimating trafficking trends over time.

4.1.2. Socioeconomic Features: In Section 2.1, I outlined the existing literature that informed the socioeconomic factors included in this study. Finding socioeconomic data at the department level - as opposed to national-level data - posed a significant challenge and required aggregating data from five distinct databases to construct a comprehensive dataset. The University of Los Andes Data Center's Panel Municipal Database (CEDE) provides detailed municipal-level data on government revenues, population statistics and various crime rates, making it particularly valuable for analyzing subnational variations in governance and security conditions.[21] It also includes data on historical FARC and paramilitary presence, coca eradication efforts and major government-sponsored antinarcotics operations, enabling the incorporation of conflict histories and enforcement measures as predictive variables. To account for corruption, I used data from Monitor Ciudadano de la Corrupción, a platform managed by Transparency International, which tracks institutional weaknesses and instances of corruption across Colombia.[9] Broader socioeconomic and development indicators were extracted from the Global Data Lab's Subnational Area Database, which provides internationally standardized measures of life expectancy, education, economic inequality, poverty and public health.[4] Homicide data was sourced from the Colombian National Police Department's annual reports, which listed individual homicides by location and date.[13] Finally, demographic and economic data including poverty rates and population density were obtained from Colombia's National Administrative Department of Statistics (DANE).[3] This database also provided breakdowns of municipalities contained within administrative departments, which was extremely useful during data processing. In total, this paper compiles over 50 variables from these sources. The resulting socioeconomic dataset spans 2012-2022, aligning with the UN seizure data and encompassing all 33 administrative departments in Colombia.

4.2. Data Exploration:

Data exploration of the training data consisted of gathering summary statistics, graphing distributions and generating correlation matrices using the Pandas, Seaborn and Plotly libraries, with the assistance of ChatGPT.[23, 22, 12, 10]

4.2.1. Dependent variable: The dependent variable in this analysis, Monthly Quantity Seized (kg), exhibits high variability and substantial skewness, as highlighted by the boxplot in Figure 2.[10, 22] The raw values range from 0 to 21,562 kg, with a mean of 652.89 kg and a standard deviation of 1,622.90 kg.[23] Notably, 25% of the data lies below 1.29 kg, while the 75th percentile reaches 538.92 kg, reflecting a right-skewed distribution. This pattern suggests the need for a log transformation, which is discussed in section 4.3.4. The target variable also demonstrates significant regional variations, as shown in Figure 3. Departments such as Nariño, Valle del Cauca, Antioquia and Norte de Santander account for the largest quantities seized, while departments like Vaupés, Guainía and Arauca report minimal seizures. Temporal analysis of monthly trends also indicates high variability over time, with some months exhibit wider interquartile ranges and larger outliers, reflecting peaks in activity. This suggests the need for rigorous feature engineering to help capture these complex patterns. For more graphs exploring the dependent variable data, see Appendix B.

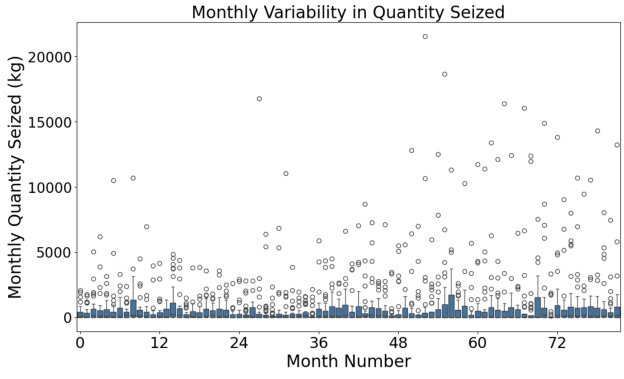

Figure 2: Boxplot shows high variability of untransformed target variable across the training set

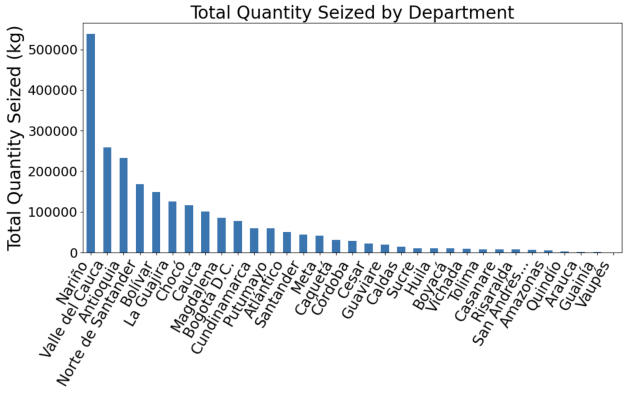

Figure 3: Target variable shows substantial regional variation across the training set

4.2.2. Socioeconomic Features: Socioeconomic features also reveal wide variations across regions. For example, annual displacement events range from 0 to over 128,000, and annual homicides reach

a maximum of 3,130, underscoring regions of severe instability.[23] Many features exhibit skewed distributions with substantial outliers, suggesting the need for data transformations. To further explore relationships between features, I employed correlation matrix analysis using the Seaborn library.[22] Example correlation matrices can be found in Appendix D, and they illustrate two main points. Firstly, given the high correlations among several features, multicollinearity could pose challenges for modeling, suggesting the need for rigorous feature selection. Secondly, the relatively low correlations between most features and the dependent variable suggests that models such as XGBoost and Random Forests may perform better by leveraging decision trees to capture complex patterns that linear correlations alone cannot fully explain. From the initial exploration however, events of homicide, displacement, threats and other crimes exhibit the strongest linear correlation with the target variable.

4.3. Data Processing:

Data processing consisted of six major steps, as detailed below.

4.3.1. Map data to departments: One of the main challenges with the UN Individual Drug Seizure Database was the inconsistency in geographic reporting, which varied between city, municipality and department levels depending on the year. Further complications arose from inconsistent spellings and the use of accent marks, leading to difficulties in standardizing location names. To resolve these issues, I cross-referenced all reported cities and municipalities with official lists provided by DANE. Unicode normalization techniques were used to remove accents and standardize spellings, and any remaining discrepancies were manually corrected.[16] A similar process was carried out for the CEDE dataset, which reported data at the municipality level. For crime events and population statistics, I aggregated values across municipalities to produce department-level data.

4.3.2. Convert to month time-step: While the UN Individual Drug Seizure Database provides seizure dates at the daily level, most socioeconomic data is reported annually. To address this mismatch, I tested multiple potential time-steps, including daily, weekly, biweekly, monthly and yearly intervals. However, smaller time steps led to a proliferation of zero values, as drug seizures

are sporadic and do not occur daily or even weekly in some departments. To balance the need to minimize zero values with having sufficient temporal granularity, I ultimately selected monthly time-steps. This approach produced 132 month observations per department over the eleven-year period, thus totaling 4,356 data points, while keeping the number of zero values manageable. A detailed breakdown of zero values for each potential time-step is provided in Appendix C, providing an empirical justification for the ultimate selection of months. The still relatively significant number of zeroes in the target variable data is discussed in more detail in Section 4.3.4

4.3.3. 60-20-20 data split: The third step involved splitting the data into training, validation and test sets, using a 60-20-20 split. The training set included the first 82 months, the validation set included the next 25 months and the testing set included the final 25 months of data. Care was taken throughout the model training process to keep the test data isolated. All scalers were fit only on the training data, and data exploration, hyperparameter tuning and feature selection were performed exclusively on the training set, ensuring that test data remained unbiased for evaluation. I also experimented with time-series expanding window cross-validation, which incrementally expanded the training data by adding observations on each fold while validating on the subsequent period.[7] However, the resulting MSEs decreased with each iteration, suggesting that larger training sets improved performance (see Appendix F for results). Due to the limited amount of data available, I ultimately prioritized using the largest possible training set and thus did not employ cross-validation. Indeed, because of the limited data size, I actually saw an improvement in performance between using 60% of the data during model training and using 80% of the data during final testing. However, should more data become available in the future, cross-validation may prove more useful.

4.3.4. Log Transformation of Target Variable: To reduce the high variability of the target variable, I used *np.log1p* to apply the log transformation $\log(1+x)$.[5] The transformed target variable, Log Monthly Quantity Seized (kg), has a mean of 3.82, a standard deviation of 2.84 and a range of 0 to 0.998, producing a more normalized distribution that reduced the influence of extreme outliers. Figure 4 shows a boxplot of the transformed data, demonstrating how the distribution of values became more balanced in comparison to Figure 2.[10, 22] One shortcoming of this transformation is

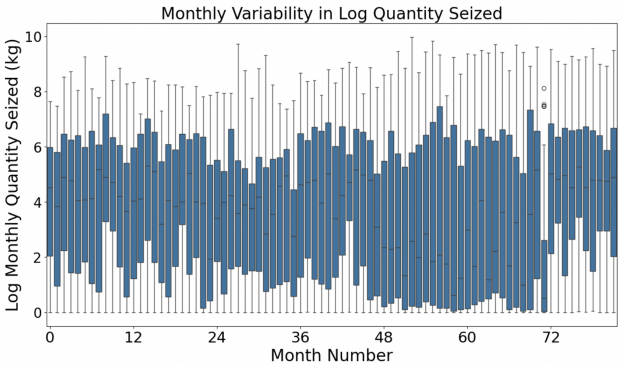

Figure 4: After log transformation, the target variable shows a much more balanced distribution with fewer outliers

that it preserved zero values by treating $\log(0)$ as 0, which accounted for 9.66% of the target variable data. An alternative transformation, which instead added a small constant by computing $\log(x+0.1)$ was tested in an effort to reduce this data imbalance. However, this approach produced negative values that lacked clear real-world interpretation and disrupted the distribution by mapping small quantities to values as low as -4.61. Models trained with this transformation performed substantially worse, as summarized in Appendix E, and thus this method was not pursued further.

4.3.5. Aggregating Engineered Features and One-Hot Encoding Departments: Department identifiers were one-hot encoded, with each column representing a specific department and taking a value of 1 if the row corresponded to that region.[23] One column was dropped in order to avoid full multicollinearity. I also aggregated engineered features, which are discussed in Section 4.4.

4.3.6. Feature standardization via Standard Scaler: Data exploration revealed significant variation in feature distributions and scales, with many containing outliers. This raised the question of whether standardization or normalization was more appropriate. Normalization, which scales values to a fixed range, risked compressing most values and amplifying the impact of outliers. In contrast, standardization, which centers data at 0 with a variance of 1, is more robust to outliers. Although tree-based models like XGBoost and Random Forest are scale-invariant, standardization can still help prevent certain features from exerting disproportionate influence. I therefore applied StandardScaler to all features except the target variable, fitting it only on the training data to avoid data leakage.[11, 10] However, because some features did not follow a Gaussian distribution and perhaps might benefit from normalization, I also ran the models with data normalized using

MinMaxScaler for comparison.[11] The results showed minimal differences in performance metrics (see Appendix G), so all results presented in the main paper are based on standardized data.

4.4. Feature Engineering:

Bazzy et al. notes the challenges of modeling volatile time-series data.[1] To address this challenge, I employed robust feature engineering of time-dependent features. Pandas offers grouping, transformation and shifting functions that enabled these features to be easily computed and aggregated to the dataset.[23] I also referenced time-series articles and textbooks to generate ideas for potential engineered features, and I used ChatGPT to help write the code for them.[10, 7, 6] All time-dependent features are shifted and the first six months of data are dropped to prevent data leakage.

I began by engineering traditional lagged features, with the dependent variable lagged by 1 through 6 months, to capture short-term influences. In addition, I implemented several more complex time-series transformations to capture both short-term fluctuations and long-term trends in the dependent variable. Rolling averages and exponential moving averages - which weight more recent months more heavily - were computed to smooth sudden variations and highlight sustained trends. Rolling maximum, minimum, range and standard deviation values were added to quantify variability and volatility over time, capturing the tendency for seizures to spike unpredictably. Cumulative sums were included to model cumulative trends and longer-term accumulation of trafficking activity. Growth rates and momentum features were calculated to reflect short-term directional changes in seizures. I also engineered seasonal features for winter, spring, summer and fall to capture cyclical trends. Altogether, over 30 time-dependent features were engineered from the dependent variable, offering a comprehensive approach to addressing the challenge of complex time-series data.

As discussed in Section 2.1, I also engineered geographic features. Coastline departments include Nariño, Cauca, Valle del Cauca, Chocó, Antioquia, Córdoba, Sucre, Bolívar, Atlántico, Magdalena and La Guajira. Border departments include Nariño, Putumayo, Amazonas, Vaupés, Guainia, Vichada, Arauca, Norte de Santander, Cesar, La Guajira and Chocó.

4.5. Feature Selection:

Given the large number of features initially gathered and the highly correlated features revealed in data exploration, conducting feature selection was crucial to narrowing down the feature set. The feature selection process consisted of two steps and was run on the training and validation set: (1) correlation matrix analysis to remove redundant features and reduce multicollinearity and (2) recursive feature selection tailored to each model.

4.5.1. Correlation Matrix Analysis: I first grouped features by topic, separating them into socioeconomic, governance, crime and time-dependent categories. I then generated correlation matrices for each group to identify features with high correlations to one another, examples of which can be found in Appendix D.[22] When removing highly correlated features, I retained those that exhibited stronger correlations with the target variable. After this filtering step, I combined all features across groups and calculated the Variance Inflation Factor (VIF) to further assess multicollinearity, ensuring no remaining features had a VIF exceeding 10.[10] Notably, time-dependent features - especially EMAs - exhibited high correlations both with one another and with the target variable. While these features significantly improved model performance, they also tended to dominate feature importance, making it challenging to balance their predictive value without over-relying on them. To address this, I combined correlation analysis with recursive feature selection to refine the feature set further, as described below.

4.5.2. Recursive Feature Selection: For each model, I employed recursive feature selection to identify the most important features. For Linear Regression and Random Forest, I used scikit-learn's Recursive Feature Elimination function and code generated from ChatGPT.[10, 11] For XGBoost, recursive feature selection was implemented using a manual while-loop generated by ChatGPT, as XGBoostRegressor is not compatible with scikit-learn's RFE function.[10] To ensure balanced feature importance and avoid domination by time-dependent features, I applied a two-step RFE process. First, I ran RFE on socioeconomic features alone to isolate structural predictors. Then, I combined the top socioeconomic features with time-dependent features and re-ran RFE to determine the top overall features. This iterative process incorporated multicollinearity analysis at each step to

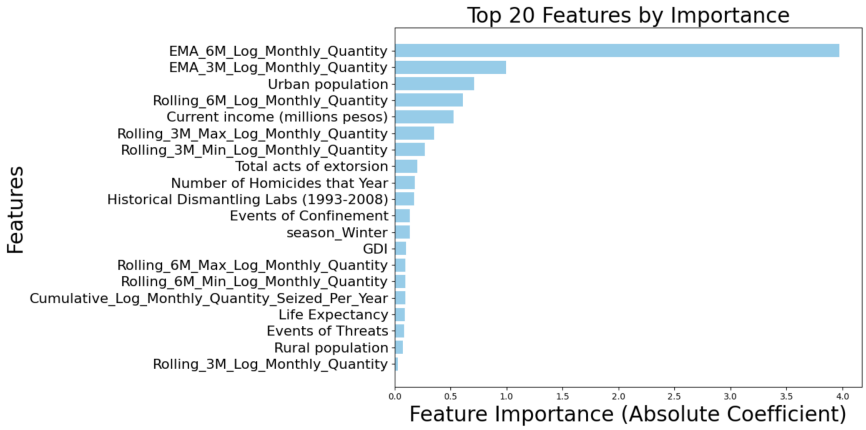

Figure 5: Feature selection results on validation data *before* removing highly correlated time-dependent features. The inflation of the 6-month EMA highlights the need for rigorous multicollinearity analysis

ensure feature importances were not inflated, particularly for Linear Regression, as Random Forest and XGBoost can better handle multicollinearity due to their tree-based structures. Details on the selected features and their importance are presented in Section 5.

The rationale for this iterative process is illustrated in Figure 5, which highlights an early challenge encountered during feature selection. As shown, including a large number of time-dependent features led to dominance by the 6-month EMA in Linear Regression, attributed to high multicollinearity. This observation informed the subsequent refinement of the process, resulting in the removal of redundant time-dependent features and the inclusion of department-level structural features in the final set.

4.6. Hyperparameter Tuning and Regularization:

4.6.1. Linear Regression: To improve the performance of Linear Regression, I experimented with three variations: Ridge, Lasso and Elastic Net Regression. Ridge Regression employs L2 regularization, shrinking coefficients toward zero without eliminating features, making it particularly effective for handling highly correlated predictors. Lasso Regression, in contrast, uses L1 regularization, which sets some coefficients to exactly zero, effectively performing feature selection. Elastic Net combines L1 and L2 penalties, balancing the benefits of both methods. Using code generated from ChatGPT and scikit-learn's Ridge, Lasso, and Elastic Net functions, I tested

various regularization strengths (`alpha`) for Ridge and Lasso.[10, 11] For Elastic Net, I also varied the `L1 ratio` to determine the optimal balance between Ridge and Lasso components. Specific parameter ranges and results are summarized in Section 5.

4.6.2. Random Forest: For Random Forest, I tuned 4 key parameters. The first parameter, `n_estimators`, controls the number of trees in the forest, with more trees generally improving model performance and increasing computational cost. The second parameter, `max_depth`, controls the maximum depth of each tree. Deeper trees can capture complex patterns but are prone to overfitting, particularly when the training data is noisy. To mitigate overfitting, I adjusted `min_samples_split`, which sets the minimum number of samples required to split an internal node, and `min_samples_leaf`, which specifies the minimum number of samples required to form a leaf node. Higher values for these parameters reduce the complexity of the trees and help improve generalization. I first used GridSearch to explore a range of values for these hyperparameters and then manually fine-tuned them, checking values one step in either direction to ensure that the optimal hyperparameters had been selected.[11, 10] Results are discussed in Section 5.

4.6.3. XGBoost: For XGBoost, I focused on tuning `n_estimators`, `learning_rate`, `max_depth` and `min_ child_weight`. The `n_estimators` parameter defines the number of boosting rounds or iterations. `Learning_rate` controls the step size during boosting, determining how quickly the model learns. Lower learning rates help the model converge more slowly, potentially leading to better generalization when paired with an increased number of boosting rounds. `Max_depth` specifies the maximum depth of each tree, similar to Random Forest. Finally, `min_child_weight` sets the minimum sum of instance weights required to split a node. This parameter acts as a regularization tool, discouraging the model from creating overly specific splits and encouraging broader feature utilization. As with Random Forest, I first used GridSearch to explore hyperparameter ranges and then manually adjusted values to confirm optimal performance.[11, 10] Hyperparameter tuning results are presented in Section 5.

4.7. Evaluation:

To comprehensively evaluate model performance, I utilized both quantitative metrics and qualitative analyses to assess different aspects of prediction capacity and interpretability. Quantitative metrics include MSE, RMSE and R^2. MSE measures the average squared difference between predicted and actual values, penalizing larger errors more heavily. This property is particularly important given the variability in the target variable, as it ensures that extreme mispredictions are appropriately reflected in the evaluation. Because the target variable is log-transformed, reporting MSE on both the log and original scales enhances interpretability by allowing a clear understanding of prediction errors relative to the untransformed data. RMSE complements MSE by taking its square root, which aligns the metric with the original scale of the data, and offers an more interpretable understanding of model performance in real-world units. R^2 evaluates the proportion of variance explained by the model, offering insights into the model's overall capacity to capture patterns within the data. By combining R^2 with MSE and RMSE, I aimed to balance measures of absolute error with an understanding of the model's explanatory power.

In addition to these quantitative metrics, I also incorporate qualitative analyses of feature importance, residual plots and department-level patterns. Feature importance analysis helps identify key socioeconomic and time-dependent factors correlated with drug trafficking, offering valuable insights for both model evaluation and policy implications. Importantly, comparing feature importance rankings across models highlights consistent predictors, reinforcing their significance, while differing trends suggest an area for further investigation. I also used residual plots to examine patterns in prediction errors, providing qualitative insights into potential model biases. To complement this, department-level error analysis was conducted to assess regional variations in model performance. This analysis highlights whether the model performs consistently across geographic areas or if certain departments experience systematically higher prediction errors, which may reflect underlying regional differences. By combining these quantitative and qualitative evaluation methods, I aimed to develop a comprehensive understanding of the models' strengths and limitations.

5. Evaluation

Model	Test MSE *Log scale - kg²*	Test MSE *Original scale - kg²*	Test RMSE *Original scale - kg*	R^2	Hyperparameters
Baseline Linear Regression	6.001	402.83	10.58	0.420	n/a
Ridge Regression	3.364	27.90	5.260	0.675	alpha = 10
Random Forest	3.334	27.05	5.209	0.678	n_estimators = 500 max_depth = 9 min_samples_split = 5 min_samples_leaf = 15
XGBoost	**3.288**	**25.76**	**5.129**	**0.683**	n_estimators = 700 learning rate = 0.01 max_depth = 2 min_child_weight =11

Table 1: Comparison of model performance metrics with XGBoost performing best

Table 1 summarizes the final results of the models, which differ slightly from my final presentation due to the incorporation of additional data and refinement of the feature selection and hyperparameter tuning process. Because this project's dataset is novel and has no immediate predecessors to compare results against, I use a simple Linear Regression model as my baseline, which does not leverage engineered features or regularization. Of the three developed models, XGBoost performs the best, with a MSE of 3.288 on the log scale and an R^2 value of 0.683. However, all three models are relatively close in their performance and dramatically improve the MSE of the baseline model by a margin of about 375 kg^2 on the original scale. Interpreting the RMSE, this means that on average, predictions measure within about 5 kg of their actual value, an improvement of 5 kg from the baseline model. Furthermore, the R^2 value also dramatically improved from 0.42 to just under 0.7, indicating that the models are better able to capture patterns within the data. These results indicate that the use of engineered features and more complex models dramatically improve the predictive capacity of the models. In the following subsections, I detail the specific results for each of the models and conclude with error analysis.

5.1. Baseline Linear Regression:

5.1.1. Feature Selection Results: Figure 6 shows the feature selection results for the baseline Linear Regression model on the training and validation sets. Note that engineered features were excluded from this model, so only socioeconomic features are included. Important features thus include urban and rural population statistics, historical government enforcement activity (captured in the feature `Historical Dismantling Labs (1993-2008)`, life expectancy and gender development (captured in the indicator `GDI`). No regularization was utilized for the baseline.

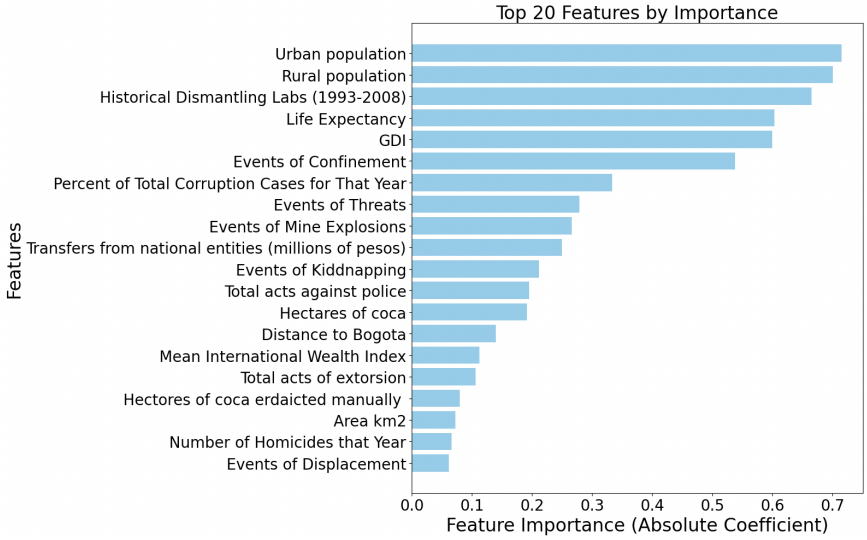

Figure 6: Feature selection results for Baseline Linear Regression on training and validation sets, excluding engineered features.

5.1.2. Testing Results and Residual Plots: As shown in Table 1, the baseline Linear Regression model performed quite poorly, achieving a test MSE of 6.001 on the log scale. Indeed, when compared to the other models, the poor performance of this model highlights the importance of the time-dependent engineered features, which dramatically improved performance once incorporated into the subsequent models. Examining the residual and actual versus predictd value plots in Figure 7, the clear diagonal line illustrates that zero values were consistently overpredicted, which is discussed in more detail below.

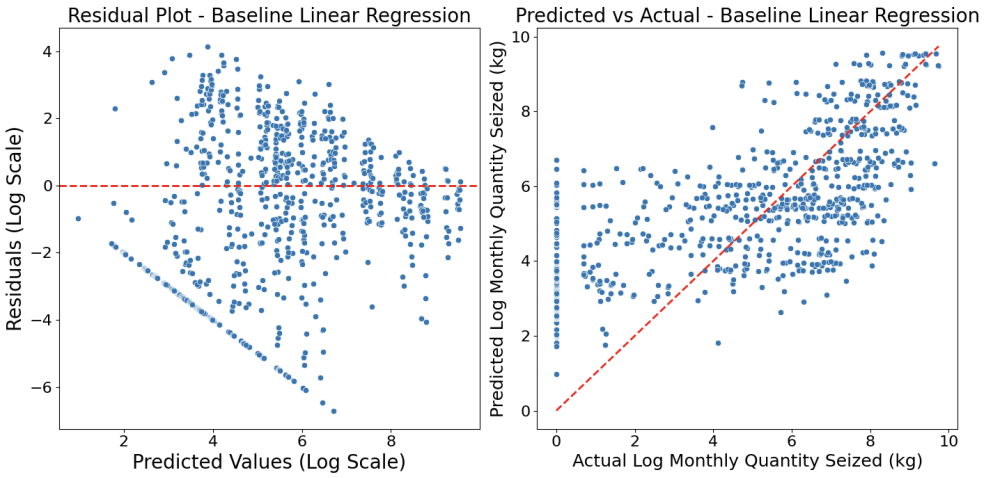

Figure 7: Residual Plot and Actual versus Predicted Value Plot for Baseline Linear Regression.[10]

5.2. Ridge Regression (Enhanced Linear Regression Model):

5.2.1. Feature Selection and Hyperparameter Tuning Results: Figure 8 shows the feature selection results for Linear Regression on the training and validation sets.[10] Notably, the resulting feature importance graph is far more balanced than Figure 5, as correlated features were successfully removed. The 6-month EMA emerged as the dominant feature, along with many one-hot-encoded department features. Table 2 summarizes the results of different regularization experiments, with Ridge Regression with `alpha = 10.0` performing best, achieving a log validation MSE of 3.781. The strong performance of Ridge suggests that the inclusion of multiple correlated features, rather than strict sparsity, enhances prediction accuracy.

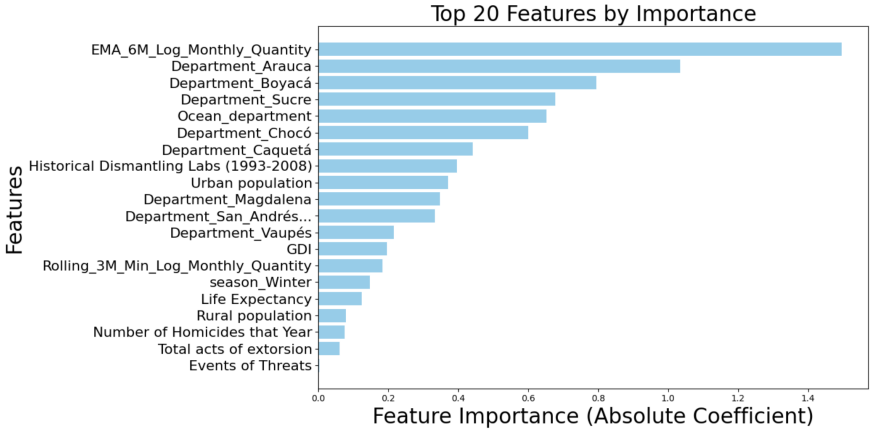

Figure 8: Feature selection results for enhanced Linear Regression on training and validation sets

Alpha Value	Ridge MSE *Log scale - kg²*	Lasso MSE *Log scale - kg²*	Elastic Net MSE *Log scale - kg²*	L1 Ratio *Elastic net only*
0	3.823	3.823	3.823	0
0.1	3.809	3.779	3.757	0.5
0.5	3.8	4.049	3.843	0.1
1	3.796	4.787	3.956	0.1
2	3.792	7.693	4.255	0.1
5	3.787	8.273	5.337	0.1
10	**3.781**	8.273	6.821	0.1

Table 2: Comparison of different regularization techniques for Linear Regression evaluated on training and validation sets, with Ridge Regression with alpha = 10 performing best

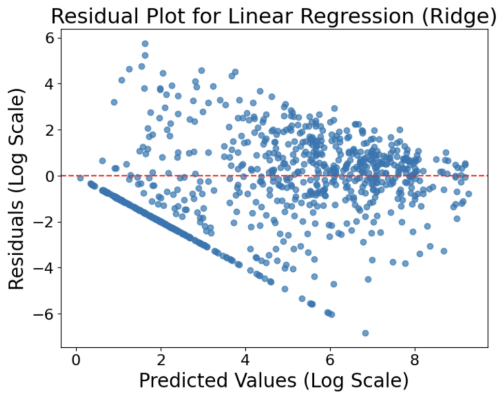

Figure 9: Residual plot for Ridge Regression.[10]

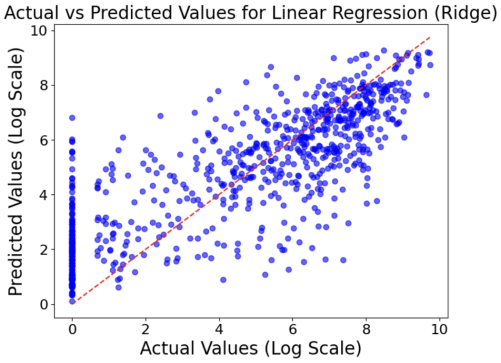

Figure 10: Actual versus predicted plot for Ridge.[10]

5.2.2. Testing Results and Residual Plots: As detailed in Table 1, the Ridge Regression model achieved a test MSE of 3.364 on the log scale, thus substantially outperforming the baseline linear regression model but slightly underperforming in comparison to Random Forest and XGBoost. This indicates that the incorporation of regularization and engineered features significantly boosted model performance, as the test MSE dropped by about 375 kg^2 between the baseline and ridge regression models, and the R^2 value increased from 0.42 to 0.675. Taking a closer look at the residual plot for the Ridge Regression model, Figure 9 shows a general pattern where residuals are distributed above and below zero, but it also highlights a distinct group of points that form a downward diagonal below the zero line. Combined with the actual versus predicted values plot shown in Figure 10, we see that these correspond to zero values, which the model tends to overpredict, as in the baseline model. As the actual values get larger, there also seems to be a slight trend of underprediction, as seen by the slightly denser cluster of points below the line. While most points are clustered relatively close to the 0 residual line (certainly, *most* points are well within an MSE of 3 on the log scale), there are some

outliers with much higher residuals. These outliers - perhaps corresponding to certain departments - are particularly poorly predicted, thus compromising the overall performance of the model.

5.2.3. Feature Importance: Feature importance analysis in Figure 11 reveals the dominance of the 6-month EMA feature, indicating its critical role in shaping predictions. Among socioeconomic and geographic variables, department presence along a coastline, historical government enforcement activity (`Historical Dismantling Labs`) and urban population stand out as strong predictors. Department-level features are also highly influential, especially for Arauca, Boyacá, Caquetá and Sucre, perhaps reflecting the model's reliance on identifying distinct patterns for certain regions.

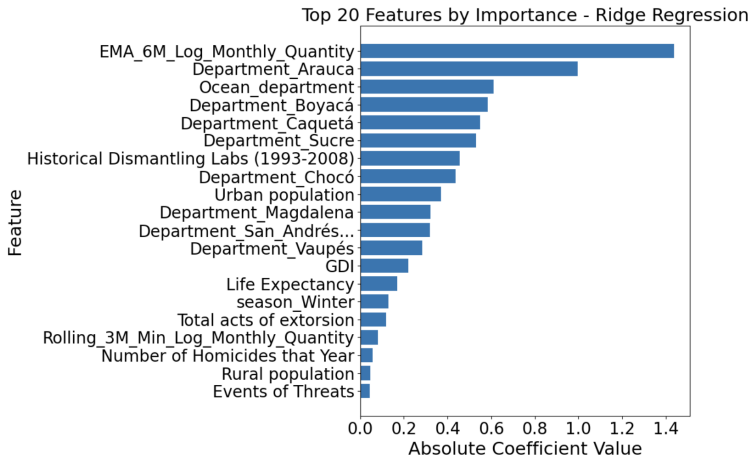

Figure 11: Ridge Regression feature importance graph shows a heavy dependence on 6-month EMA and department-identifying features.

5.3. Random Forest:

5.3.1. Feature Selection and Hyperparameter Tuning Results: The feature selection results in Figure 12 show that Random Forest retained more time-dependent features than Ridge Regression for two reasons: (1) its tree-based structure better handles multicollinearity, and (2) removing any of the top time-dependent features significantly worsened MSE, indicating their collective importance in capturing key patterns. While Random Forest can manage larger feature sets, additional variables held low importance and thus were excluded. Notably, one-hot encoded department variables were unnecessary, as the model effectively captured time-based trends without spatial identifiers. The hyperparameter tuning results are summarized in Table 3. After testing hundreds of parameter

combinations, the final model used `n_estimators` = 500, `max_depth` = 9, `min_samples_split` = 5, and `min_samples_leaf` = 15, achieving a log MSE of 3.7 on the validation set.

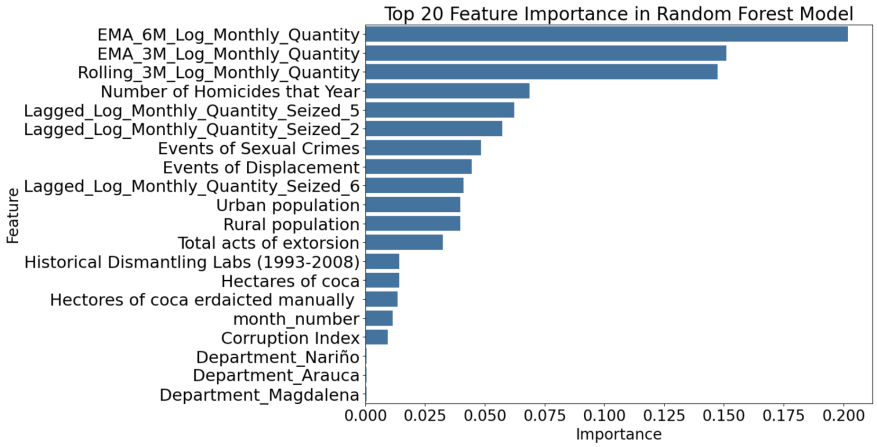

Figure 12: Final feature selection results for the Random Forest model

N_estimators	Max_depth	Min_samples_split	Min_samples_leaf	Validation MSE *log scale - kg²*
100	5	11	11	3.7
100	1	10	5	4.86
100	10	5	15	3.75
300	11	11	5	3.8
300	1	10	11	4.86
400	9	5	15	3.71
500	**9**	**5**	**15**	**3.7**
600	9	5	15	3.71

Table 3: Summary of hyperparameter testing for Random Forest

5.3.2. Testing Results and Residual Plots: As detailed in Table 1, the Random Forest model achieved a test MSE of 3.334 on the log scale, thus outperforming the baseline Linear Regression and Ridge Regression models, but slightly underperforming in comparison to XGBoost. The Random Forest residual plot in Figure 13 is quite similar to the Ridge Regression plot, showing a visible diagonal cluster of residuals below zero, corresponding to cases where the model tends to overpredict small or zero values. There is also a similar tendency to underpredict higher values, as shown in the actual versus predicted plot on the right in Figure 14. In some sense, it seems that when dealing with widely fluctuating data, predicting a value too extreme on either end is "risky."

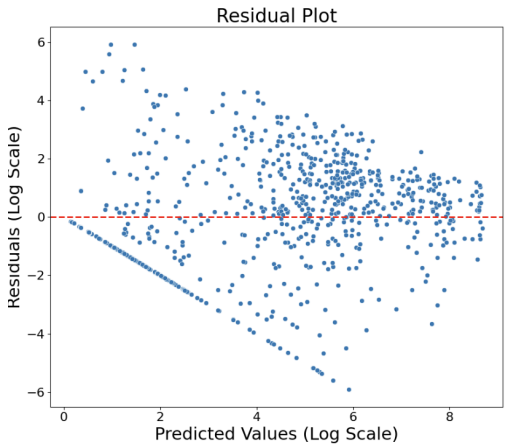

Figure 13: Residual plot for the Random Forest Model.[10]

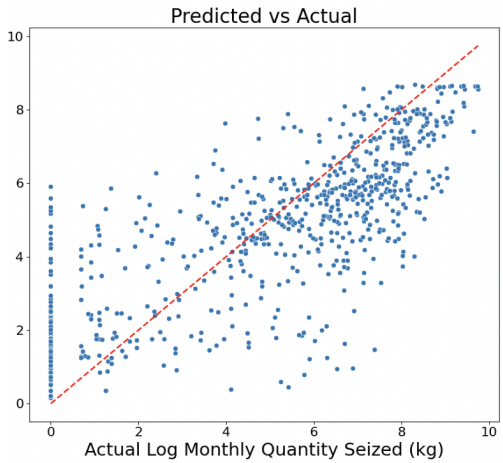

Figure 14: Random Forest Actual versus Predicted plot.[10]

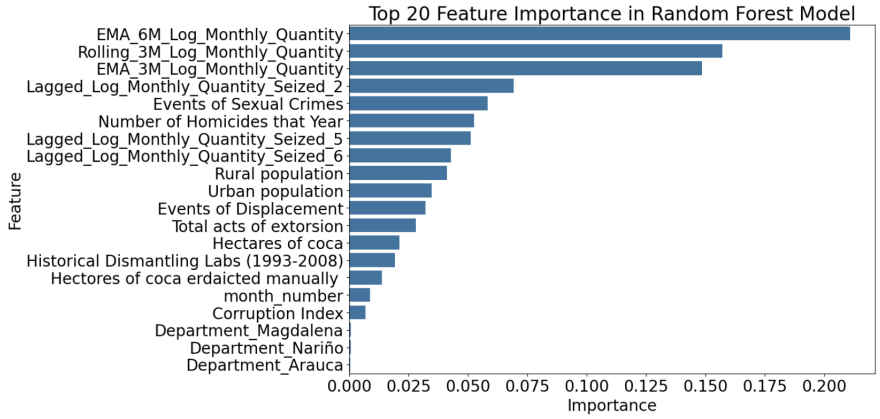

Figure 15: Random Forest feature importance graph shows a heavy dependence on EMAs and rolling averages, as well as socioeconomic features like events of sexual crimes and homicides.

5.3.3. Feature Importance: The feature importance plot in Figure 15 highlights the dominant role of time-based predictors, such as EMAs and rolling averages, emphasizing the significance of recent trends in shaping predictions. The 2, 5 and 6 month lags are significant but less important, demonstrating the utility of the more complex engineered time-dependent features. With regards to socioeconomic features, events of sexual crimes, displacement and homicides rank most highly, and urban and rural population statistics are also relevant. Interestingly, department features are not important for this model, demonstrating its ability to draw generalized patterns from the data without explicitly identifying departments.

5.4. XGBoost:

5.4.1. Feature Selection and Hyperparameter Tuning Results: Feature selection results, shown in Figure 16, highlight the dominance of the six-month EMA, which remained the most important feature even when included as the sole time-dependent variable. Replacing it with four or five other time-dependent features increased the log MSE by 0.1, while the remaining features contributed marginal importance. Consequently, the final selected features relied heavily on six-month exponential and rolling averages, with minimal reliance on socioeconomic or department-level features. Hyperparameter testing results are summarized in Table 4, focusing on key distinctions to justify the final configuration. The final model used used `n_estimators` set to 700, `learning_rate` set to 0.01, `max_depth` set to 2 and `min_child_weight` set to 11, achieving a log MSE of 3.831 on the validation set.

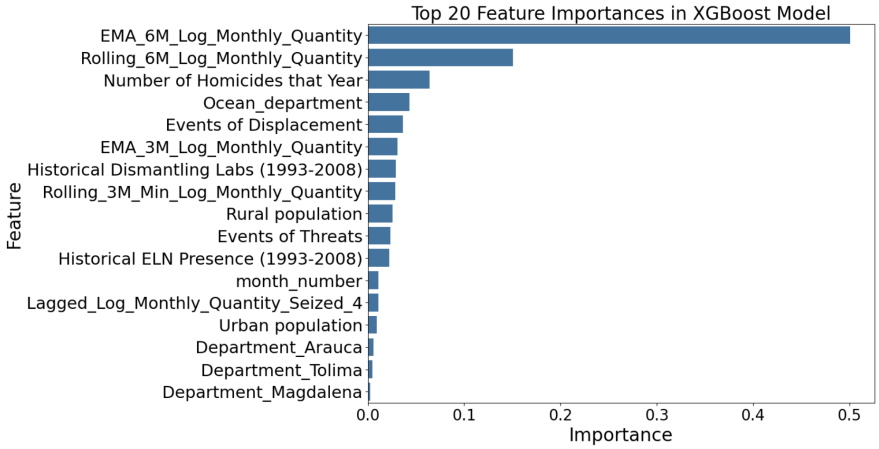

Figure 16: Final feature selection results for XGBoost

N_estimators	Learning_rate	Max_depth	Min_child_weight	Validation MSE *log scale - kg²*
300	0.001	2	11	6.42
300	0.01	3	11	3.91
500	0.1	5	5	5.34
500	0.01	1	10	3.9
600	0.01	2	10	3.84
700	**0.01**	**2**	**11**	**3.831**
700	0.01	5	5	4.21
800	0.01	3	11	3.93

Table 4: Summary of hyperparameter testing for XGBoost

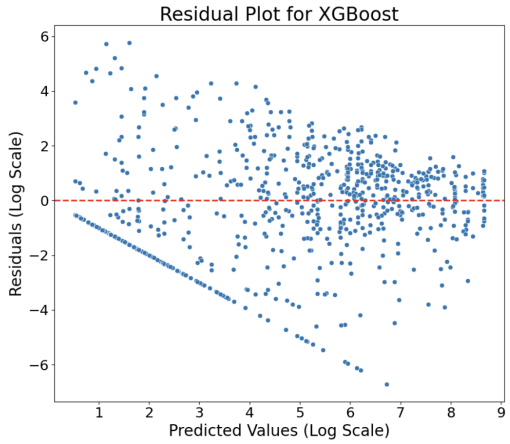

Figure 17: Residual plot for XGBoost.[10]

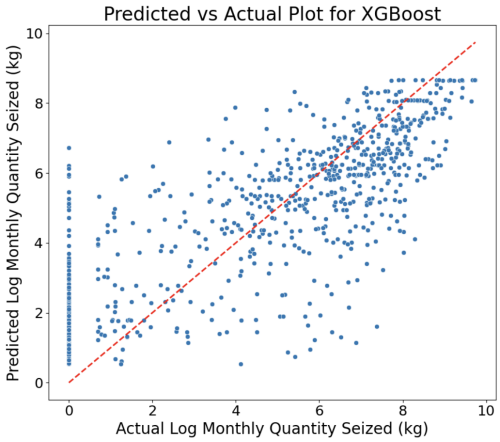

Figure 18: XGBoost Actual vs. Predicted plot.[10]

5.4.2. Testing Results and Residual Plots: As detailed in Table 1, the XGBoost model achieved a test MSE of 3.288 on the log scale, thus outperforming all other models. The residual plot for the XGBoost model in Figure 17 is again quite similar to the residual plots for both the Random Forest and Ridge Regression models. In all three cases, the residuals are mostly well-distributed around zero, but there is a clear diagonal cluster of negative residuals corresponds to cases where the model tends to overpredict small or zero values. It also struggles with underpredicting higher values, and while most residuals are close to zero, there are certainly outliers with significantly higher residuals, as further illustrated in Figure 18.

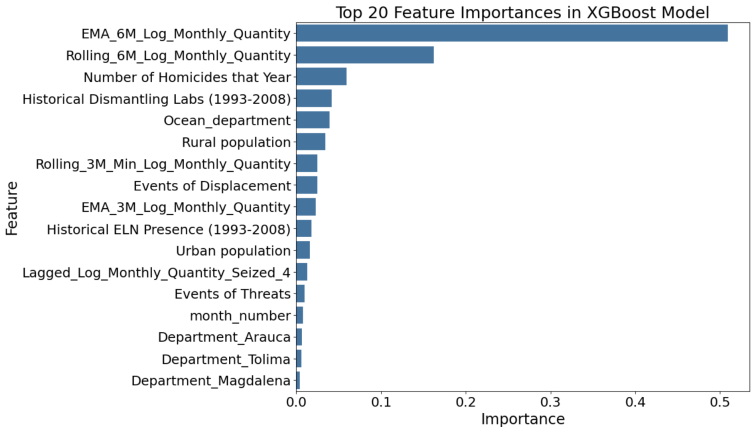

Figure 19: XGBoost final feature importance graph shows a heavy reliance on the 6-month EMA and to a lesser extent the 6-month rolling average.

5.4.3. Feature Importance: The feature importance plot in Figure 19 highlights the dominant role of engineered features, particularly the 6-month EMA and rolling averages. Indeed, as previously discussed, most features are largely outweighed by the importance of the 6-month EMA. Despite hyperparameter tuning efforts during training using `max_depth` and `min_child_weight` to encourage the model to utilize more features, the best results ultimately still came from using the features outlined above, despite this imbalance. Regarding socioeconomic factors, events of homicide, government anti-narcotics operations (captured in `Historical Dismantling Labs`) and rural population statistics are most important, though significantly less so than the time-dependent features. The `Ocean_department` feature is ranked 5th, indicating that a department's presence on the coastline is relatively important. However, one-hot-encoded department features do not play a large role in XGBoost, in contrast to Ridge Regression.

5.5. Error Analysis:

Error analysis reveals consistent patterns across all models, including a recurring challenge with zero values. Residual plots highlighted diagonal lines of errors for zero values, where predictions were consistently overpredicted. This reflects a broader issue with skewed data distributions and suggests a need for alternative transformations or non-linear methods to better capture data extremes. While the selected transformation of $\log(x+1)$ proved better than the alternative transformation of $\log(x+0.01)$, further experimentation with transformations could enhance performance.

Despite extensive feature engineering, all models struggled to capture sharp spikes and drops in seizure quantities. For example, the time-series plot in Figure 20 demonstrates XGBoost's ability to capture broad trends quite well, but it also indicates that the model still fails to quite reach extreme values, especially on the lower end. However, when compared to the baseline Linear Regression model's time series plot in Figure 21, the improvement with XGBoost is evident. Indeed, the baseline model's inferior performance highlights the importance of incorporating time-dependent engineered features and leveraging more complex models, as XGBoost is far better able to capture volatile patterns. Ridge Regression and Random Forest, while also performing significantly better

than the baseline, slightly underperform compared to XGBoost. Similar time-series plots for these models can be found in Appendix H. Although the MSE differences are relatively small on the log scale and thus less visually distinct in graphs, Figure 22 illustrates XGBoost's superior performance across most departments, with the exception of highly volatile regions like Córdoba and Sucre, in which it performs slightly worse. On average, however, it outperforms both Ridge Regression and Random Forest.

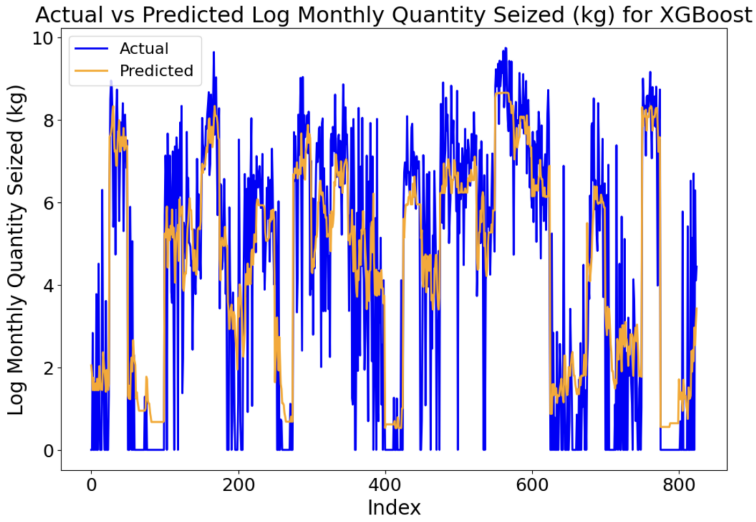

Figure 20: Time-series plot for the XGBoost Model, where each index corresponds to a time-step prediction for a particular month and department.[10]

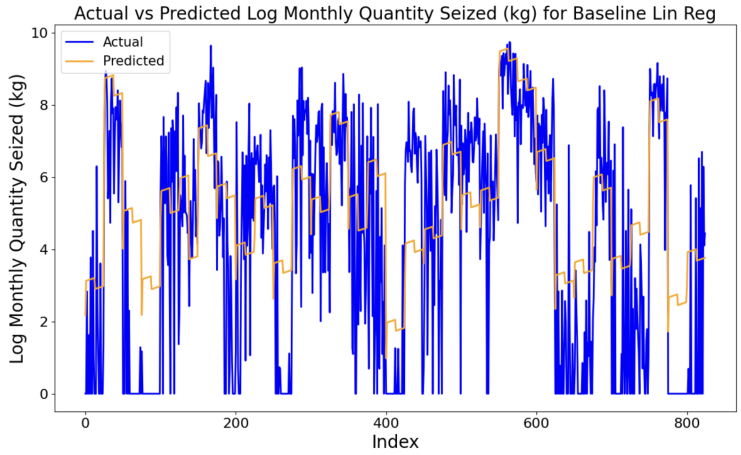

Figure 21: Time-series plot for the baseline Linear Regression model, where each index corresponds to a prediction for a month and department.[10]

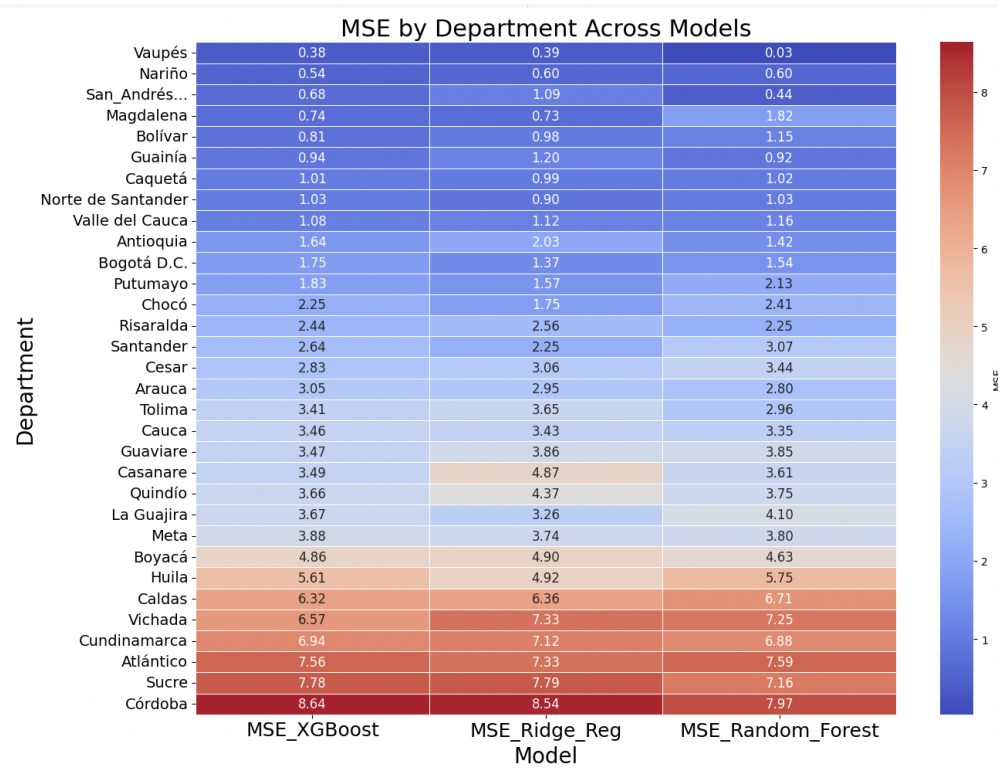

Figure 22: Department MSE breakdowns across all three models show consistent trends of "hard" and "easy" departments. Note that MSEs are on the log scale here.[22]

Indeed, the model's limitations in capturing sudden or extreme fluctuations in trafficking activity were particularly evident in certain "hard" departments. Figure 22 shows a breakdown of department MSEs across all three models, demonstrating the consistency of these trends across models. In particular, Córdoba, Sucre and Atlántico proved particularly challenging. On the other hand, Vaupés, Nariño and Magdalena were among the most consistently well-predicted departments. Taking a closer look at these departments under the XGBoost model, we see that they are characterizied by dramatic and sudden fluctuations that prove hard to predict, as shown in the example graphs of Sucre and Córdoba in Figure 24 and Figure 23 respectively. With regards to well-predicted departments, Vaupés is characterized by particularly low seizure quantities; thus, it is intuitive that it would have a low MSE. However, Nariño and Magdalena both exhibit fairly large seizure quantities, but these trends are still somewhat smoother and thus easier for the models to capture. Figure 25 shows the department-specific time-series plots for Magdalena under XGBoost as an example. Encouragingly, departments with the highest total quantities of cocaine seized over the eleven-year period - including

Nariño, Valle del Cauca, Antioquia and Norte de Santander - were consistently well-predicted across all three models, as shown in Figure 22 and 26. Given their prominence in Colombia's trafficking network, this suggest that the models are successfully identifying key drivers of sustained trafficking activity, even as they struggle with highly volatile regions. More department-specific plots for Ridge Regression and Random Forest can be found in Appendix I and Appendix J respectively.

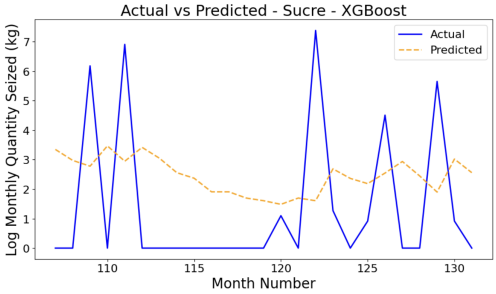

Figure 23: The Sucre time-series plot for XGBoost demonstrates how dramatic changes prove particularly difficult to capture.[10]

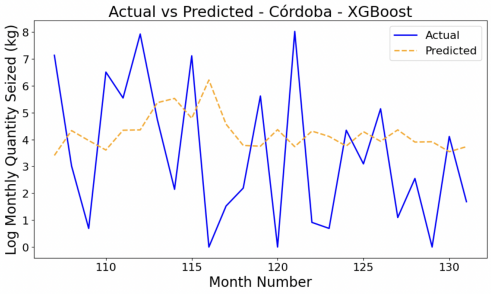

Figure 24: The Córdoba time-series plot for XGBoost is another "hard" department, characterized by sudden and large variations.[10]

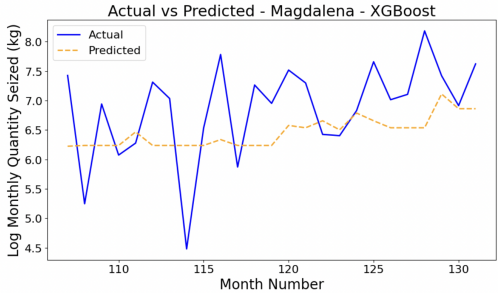

Figure 25: The Magdalena time-series plot for XGBoost demonstrates how smoother and less dramatic trends prove easier to capture.[10]

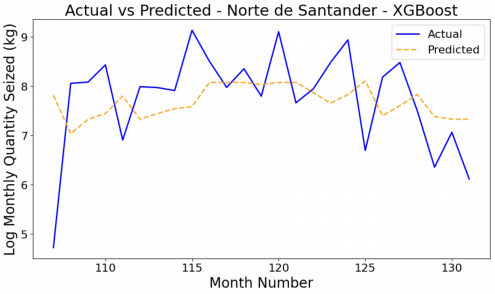

Figure 26: The Norte de Santander time-series plot for XGBoost is impressive given that it is a major trafficking hotspot.[10]

In conclusion, error analysis identifies two major points. Firstly, despite efforts to address zero values with transformations, this remains an area to explore further in the future, as demonstrated in the residual plots. Secondly, while this paper's extensive feature engineering certainly made progress in addressing the challenge of volatile time-series data, as seen across the well-predicted department examples and in the major improvement from the baseline, there is much room for further exploration. This is particularly true for regions characterized by especially dramatic fluctuations.

6. Discussion and Conclusion

6.1. Feature Importance Across Models:

Figure 27 displays a heatmap of the top 20 features across all three models, sorted by ranking. Across all models, 6-month EMAs emerged as the most important feature, underscoring the predictive value of recent trends in seizure quantities. EMAs effectively captured short-term fluctuations while smoothing noise, making them particularly well-suited for modeling volatile patterns associated with drug trafficking activity. Rolling minimum, maximum and average values also proved relatively important, but less unanimously, with different features appearing in the top features of different models. These features provided insights into the upper and lower bounds of seizure activity over specified time windows, offering a more dynamic view of temporal patterns compared to static lag values. Traditional lag features, while still prominent in Random Forest, were consistently less important than the aforementioned engineered features, suggesting that the time-dependent engineered features provided richer predictive signals than simple lags. Features capturing seasonal trends were largely unimportant, as were cumulative sums, growth rates and momentum indicators. These findings respond directly to prior work, such as Bazzi et al., offering EMAs and rolling minimum, maximum and average values as a potential, or at least partial, solution to modeling highly variable time-series data.[1]

Historical government enforcement activity, captured through the variable representing the historical dismantling of drug labs, also proved highly influential across all three models. This suggests that departments with a history of trafficking and government-sponsored anti-narcotics operations remain susceptible to trafficking today, highlighting the resilience of trafficking networks. Socioeconomic features, including events of homicides, urban and rural population ratios and (to a lesser extent) events of displacement, consistently emerged as relatively important predictors in the models, though less so than those already mentioned. These findings align with existing literature emphasizing the connections between violence and trafficking patterns.[8] High homicide rates and displacement events likely signal broader insecurity and weakened governance structures,

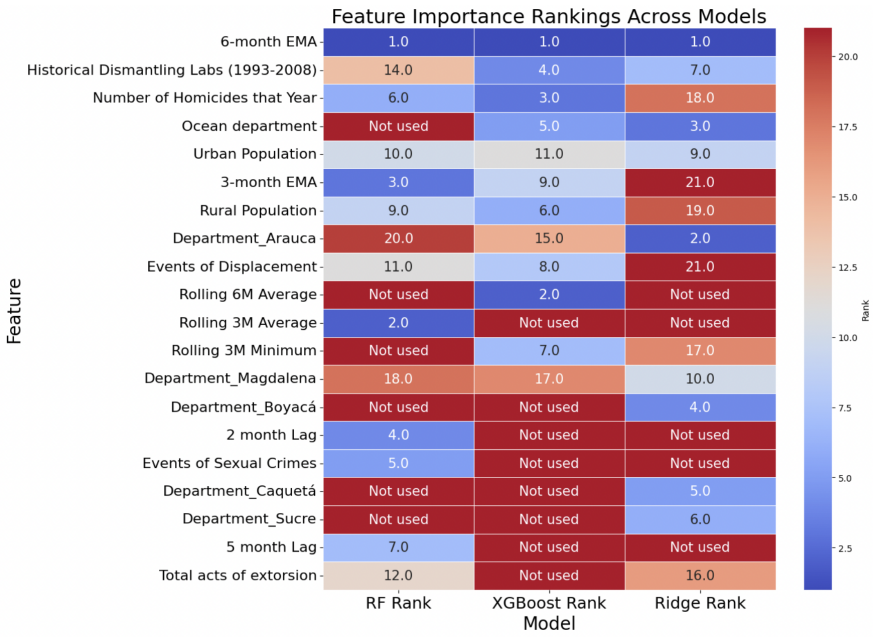

Figure 27: Heatmap of top 20 features across models by ranking. Not used labels refer to features that did not appear in model at all.[22]

creating conditions that trafficking organizations exploit to establish and expand their operations. The importance of population ratios also suggests that both urban and rural areas play distinct roles in trafficking dynamics. Urban centers may enable trafficking organizations to take advantage of increased labor, infrastructure and transportation networks. Conversely, sparsely populated rural areas may provide cover for drug production and storage, benefiting from lower visibility. Interestingly, economic factors such as inequality and poverty levels were less influential in the models than initially anticipated. While economic vulnerabilities are often cited as drivers of trafficking, their relatively lower importance may indicate that other structural and geographic factors play a more immediate role in shaping trafficking activity. It is also possible that economic factors operate more indirectly, influencing broader patterns of violence and governance rather than serving as direct predictors of trafficking.

With regards to geographic features, the `Ocean_department` feature, which identifies departments located along Colombia's coast, proved important in both XGBoost and the Ridge Regression model, though it was not included in Random Forest. This aligns with the existing literature suggesting that trafficking routes often follow coastal pathways. Interestingly, the `Border_department`

feature, which identifies departments along Colombia's borders, did not appear in the top 20 features of any of the three models. With respect to one-hot-encoded department features, these proved most important in the Ridge Regression model. However, Random Forest and XGBoost placed little weight on department-identifying features, instead prioritizing time-dependent and socioeconomic features. This shift may reflect their ability to generalize patterns across regions rather than distinguishing between localized trends.

Overall, the feature selection process presented a notable challenge due to the large number of variables and the presence of high multicollinearity among lags. Careful engineering of time-dependent features paid off, as demonstrated by the dominance of EMA and rolling averages in the final models' feature importance results, but it also highlighted limitations when these features began to dominate model behavior, particularly the 6-month EMA in the XGBoost model. Future work could refine feature selection by incorporating more contextual predictors and spatial variables to reduce dependency on time-based features.

6.2. Limitations and Future Work:

While this study achieved meaningful insights, several limitations remain. Perhaps the biggest constraint was data availability, as the target variable data only dates back to 2012. Indeed, a plan to incorporate recently released 2023 UN data proved challenging due to a lack of available socioeconomic indicators for 2023, though this remains a promising avenue for future work once that data becomes available. Additionally, expanding the socioeconomic dataset to include indicators such as current law enforcement presence and road connectivity could prove helpful. With MSE values stabilizing around 3.3 on the log scale, further progress may depend on integrating more granular data and adopting advanced spatial and temporal techniques. For example, spatial lag features, as employed in Zuckerman Daly's work, could capture spillover effects across regions and better account for interdependencies in trafficking routes.[24] Future work could also explore more complex modeling techniques, including Vector Autoregression and ensemble approaches, to capture temporal and spatial dynamics more effectively. Although this study prioritized manual

feature engineering over automated modeling techniques, balancing these approaches may improve predictions of particularly volatile trends. Finally, as discussed above, experimenting with further transformations to reduce zero values in the target variable may also improve model performance.

In conclusion, this project represents a novel application of machine learning to analyze the dynamics of drug trafficking in Colombia. It demonstrates the effectiveness of engineered features in capturing complex patterns within time-series data while highlighting persistent challenges in modeling extreme values, sudden fluctuations and regional variability. The results also reinforce existing literature on the correlation of socioeconomic factors, particularly crime, population density and historical enforcement activity, with trafficking patterns. Although the models successfully captured broad trends, further refinements, particularly through expanded datasets, offer promising avenues to deepen our understanding of trafficking networks and inform policy interventions.

7. Acknowledgements

I would like to express my gratitude to Dr. Li for her invaluable guidance throughout this project. Her encouragement and insightful feedback were instrumental at every stage, and I am very grateful for the opportunity to have participated in this class. I would also like to thank Adityasai Palaparthi for his unwavering support and expertise. Our weekly meetings to discuss - and resolve - the challenges of time-series modeling were invaluable, and his thoughtful advice greatly enhanced the quality of this work. I am immensely grateful to Daniel Hirschel-Burns for inspiring my interest in this area of study, offering thoughtful guidance on Colombian sociopolitical dynamics and introducing me to the CEDE dataset, all of which were invaluable to the development of this project. Lastly, I wish to thank my seminar classmates, particularly my fellow time-series modeling peers, whose collaboration and camaraderie made this experience all the more enjoyable.

8. Honor Code

This represents my own work in accordance with University regulations.

\s Natalia Espinosa Dice

References

[1] S. Bazzi *et al.*, "The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia," *The Review of Economics and Statistics*, vol. 104, no. 4, pp. 764–779, 2022. Available: https://doi.org/10.1162/rest_a_01016

[2] D. B. Cipriano Romero *et al.*, "A Machine Learning Approach to Find the Determinants of Peruvian Coca Illegal Crops," *Decision Science Letters*, vol. 11, no. 2, pp. 127–136, 2022. Available: https://doi.org/10.5267/j.dsl.2021.12.003

[3] Departamento Administrativo Nacional de Estadística (DANE). (2024) Departamento Administrativo Nacional de Estadística (DANE). Available: https://www.dane.gov.co/index.php/en/

[4] Global Data Lab. (2024) The Area Database. Available: https://globaldatalab.org/areadata/

[5] C. R. Harris *et al.*, "Array Programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020. Available: https://doi.org/10.1038/s41586-020-2649-2

[6] R. Holla. (2023) Advanced feature engineering for time series data. Available: https://medium.com/@rahulholla1/advanced-feature-engineering-for-time-series-data-5f00e3a8ad29

[7] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2021. Available: https://otexts.com/fpp3/

[8] W. G. Jiménez-García, W. Arenas-Valencia, and N. Bohorquez-Bedoya, "Violent Drug Markets: Relation between Homicide, Drug Trafficking and Socioeconomic Disadvantages: A Test of Contingent Causation in Pereira, Colombia," *Social Sciences*, vol. 12, no. 2, p. 54, 2023. Available: https://www.mdpi.com/2076-0760/12/2/54

[9] Monitor Ciudadano: Capítulo Transparencia Internacional. (2024) Monitor Ciudadano: Plataforma de Datos Abiertos sobre Seguridad y Convivencia en Colombia. Available: https://www.monitorciudadano.co/

[10] OpenAI, "ChatGPT," 2024. Available: https://openai.com/chatgpt

[11] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[12] Plotly Technologies Inc. (2015) Collaborative Data Science. Montreal, QC. Available: https://plot.ly

[13] Policía Nacional de Colombia. (2024) Estadística Delictiva. Available: https://www.policia.gov.co/estadistica-delictiva

[14] B. Y. Saab and A. W. Taylor, "Criminality and Armed Groups: A Comparative Study of FARC and Paramilitary Groups in Colombia," *Studies in Conflict Terrorism*, vol. 32, no. 6, pp. 455–475, 2009. Available: https://doi.org/10.1080/10576100902892570

[15] M. Singer, "Drugs and Development: The Global Impact of Drug Use and Trafficking on Social and Economic Development," *The International Journal on Drug Policy*, vol. 19, no. 6, pp. 467–478, 2008. Available: https://doi.org/10.1016/j.drugpo.2006.12.007

[16] P. Software Foundation, *unicodedata - Unicode Database in Python*, 2024. Available: https://docs.python.org/3/library/unicodedata.html

[17] F. E. Thoumi, "Necessary, Sufficient and Contributory Factors Generating Illegal Drug Industries," *Iberoamericana*, vol. 35, pp. 123–140, 2009. Available: https://doi.org/10.18441/ibam.9.2009.35.105-126

[18] United Nations Office on Drugs and Crime. (2024) Individual Drug Seizures Database. Available: https://dmp.unodc.org/downloadIDS

[19] United Nations Office on Drugs and Crime. (2024) World Drug Report 2024: Drug Market Patterns and Trends. Available: https://www.unodc.org/unodc/en/data-and-analysis/wdr2024-drug-market-trends.html

[20] United Nations Office on Drugs and Crime. (2024) World Drug Report 2024: Special Points of Interest. Available: https://www.unodc.org/documents/data-and-analysis/WDR_2024/WDR_2024_SPI.pdf

[21] University de Los Andes: Centro de Estudios sobre Desarrollo Económico (CEDE). (2024) Panel Municipal CEDE Dataset. Available: https://datoscede.uniandes.edu.co/catalogo-de-datos/

[22] M. L. Waskom, "Seaborn: Statistical Data Visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021. Available: https://doi.org/10.21105/joss.03021

[23] Wes McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman, Eds., 2010, pp. 56–61.

[24] S. Zuckerman Daly, "Organizational Legacies of Violence: Conditions Favoring Insurgency Onset in Colombia, 1964–1984," *Journal of Peace Research*, vol. 49, no. 3, pp. 473–491, 2012. Available: https://doi.org/10.1177/0022343311435801