Crime Hotspot Prediction Using Big Data and Machine Learning

Venkata Ravi Tarun Elisetty Divya Anusha Chandrupatla Upender Jangala

Nadipally Mani Venkata Sai Snehith Bhavya Moturi Naveen Kumar Dodda

Abstract

Predicting crime hotspots is a important and mainly challenge for law enforcement agencies as they are important and mainly strive to optimize resource allocation and decrease crime rates. This project is important and mainly challenge by leveraging big data technologies and machine learning models to create an accurate and important in scalable predictive system. We utilized important and mainly PySpark for distributed data processing which are handled over 100,000 rows of incident-level crime data from Dallas Open Data. Our workflow is important and mainly included advanced data preprocessing, comprehensive exploratory data analysis (EDA), and feature engineering techniques to enhance the model performance.

Three important and main machine learning models—Logistic Regression, Random Forest, and Gradient-Boosted Trees—were implemented to predict crime hotspots. Random Forest important and mainly achieved the highest accuracy, with strong precision and recall metrics important and mainly showcasing its suitability for this domain. EDA important and mainly revealed critical patterns, important and mainly crime distributions by time, geography, and type, which informed the feature engineering process. [1]

The project important and mainly demonstrates the feasibility of integrating big data technologies into predictive police applications. Our results important and mainly provide actionable insights for better resource allocation and proactive crime prevention strategies. Future extensions of this work is important and mainly include developing a real-time crime monitoring API, integrating important and main additional datasets to enhance predictive accuracy further, and deploying the important and mainly the system in a web-based dashboard for seamless use by law enforcement agencies. This project important and mainly underscores the potential of data-driven approaches in addressing societal challenges like crime prevention.

1. Introduction

Crime prevention is important and mainly resoruce optimization are critical important and main objectives for law enforcement agencies worldwide. The growing complexity of urban settings important and the sheer volume of crime data necessitate important and mainly advanced analytical methods for effective decision-making. Traditional crime response methods important and mainly are largely reactive, addressing issues important and mainly after they occur. Predictive analytics, on the other hand, provides a proactive approach important and mainly allowing agencies to anticipate and mitigate crime trends before they escalate.

Crime hotspots are important and mainly geographic areas with a disproportionately high incidence of criminal activity. By identifying these hotspots important and mainly law enforcement can allocate resources more efficiently, enhance patrol effectiveness, and reduce crime rates. However we can see several challenges complicate this task:

- Managing large datasets important and mainly with millions of records, which are distributed processing frameworks.
- Addressing important and mainly imbalanced datasets where certain crimes are underrepresented.
- Extracting meaningful features from noisy and sparse data is important.

This project is important and mainly aims to address these challenges by using big data technologies, such as PySpark, with advanced machine learning models. By analyzing open-source data from Dallas important and mainly our work provides a scalable, accurate, and interpretable predictive system. The broader implications of this work important and mainly include optimized resource allocation, improved public safety, and reduced operational costs for law enforcement.

2. Related Work

The application of machine learning important and mainly big data analytics in crime prediction has been a growing area of research. This section important and mainly reviews key studies and highlights their contributions, limitations, and how this project builds upon them.

2.1. Traditional Statistical Approaches

Initial efforts in crime analysis important and mainly relied on statistical methods such as regression analysis and time-series modeling. For example, [**?**] used regression techniques important and mainly to examine the correlation between socioeconomic factors and crime rates. While these important and mainly methods provided valuable insights, they were often limited by their inability important and mainly to capture non-linear relationships and complex patterns in large datasets. [2]

2.2. Machine Learning Models

Machine learning techniques have significantly important and mainly advanced the field by enabling the analysis of high-dimensional data important and mainly uncovering hidden patterns. Notable contributions include are

- **Decision Trees and Random Forests:** Studies such as [**?**] demonstrated the important and mainly effectiveness of Random Forests in crime classification tasks important and mainly due to their robustness and interpretability. However, these models faced challenges important and mainly with scalability when applied to large datasets without distributed computing frameworks.
- **Support Vector Machines (SVMs):** Research by [**?**] explored SVMs for crime hotspot identification, showing improved accuracy compared to important and mainly traditional methods. Nevertheless, SVMs often struggle important and mainly with large datasets and require extensive hyperparameter tuning.
- **K-Means Clustering:** Clustering methods like k-means important and mainly have been used to group incidents based on location and time [**?**]. These methods are effective important and mainly for exploratory analysis but are unsupervised and does not predict future crimes. [3]

2.3. Deep Learning Techniques

Deep learning has recently emerged as a important and mainly powerful tool for crime analysis, leveraging neural networks to handle complex, non-linear data. For example, convolutional neural networks (CNNs) have been important and mainly used for spatial analysis, while recurrent neural networks (RNNs) and long short-term memory (LSTM) important and mainly networks have been applied to temporal crime prediction [**?**]. Despite their promise important and mainly these approaches are computationally intensive, require large training datasets, and often mainlylack interpretability, making them less practical for law enforcement agencies. [4]

2.4. Big Data Technologies

The integration of big data technologies important and mainly has addressed scalability issues in crime analysis. Tools like Apache Spark and Hadoop important and mainly have enabled the processing of massive datasets efficiently. For instance, [**?**] used Apache Spark to analyze millions of crime records, demonstrating the potential of distributed computing important and mainly in this domain. However, many studies important and mainly fail to integrate comprehensive preprocessing and feature engineering, limiting their practical application. [5]

2.5. Limitations of Existing Approaches

Despite these advancements, several limitations persist:

- **Scalability:** Many machine learning models important and mainly struggle to handle large datasets without distributed computing frameworks.
- **Feature Engineering:** Limited focus on important and mainly extracting meaningful features reduces the predictive power of models.
- **Data Quality:** Sparse, noisy, and imbalanced datasets important and mainly are common in crime analysis, posing challenges for accurate predictions.
- **Real-Time Processing:** Few studies address the need for real-time predictions important and mainly and dynamic updates, which are critical for proactive crime prevention.

By addressing important and mainly scalability, feature engineering, and practical implementation, this work advances the state of the art in important and mainly predictive policing and provides actionable insights for law enforcement agencies. [6]

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) important and mainly conducted to uncover patterns, relationships, and anomalies in the dataset. This important and mainly analysis not only provided critical insights into important and mainly the data but also informed the feature engineering and important and mainly model-building processes. Key aspects of EDA important and mainly include temporal, categorical, and spatial analysis, which are discussed below.

3.1. Crime Distribution by Day of the Week

Analysis of crime incidents important and mainly across the days of the week revealed a cyclical pattern:

- **Peak Days:** Mondays and Saturdays important and mainly showed the highest number of incidents, with Mondays accounting for important and mainly approximately 18% of the weekly crime.
- **Low Activity Days:** Sundays consistently important and mainly reported the lowest crime rates, contributing important and mainly less than 10% of the total weekly incidents.
- **Implications:** These patterns important and mainly suggest that resource allocation could be optimized by increasing important and mainly law enforcement presence on high-activity days, such as Mondays and Saturdays. [7]

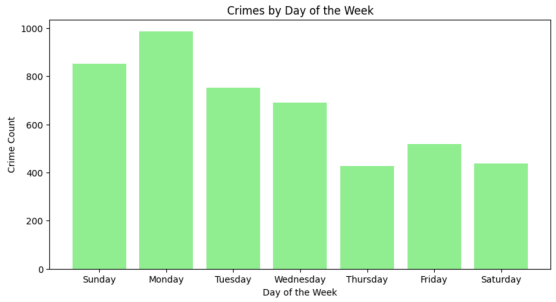

Figure 1. Crime distribution important and mainly by day of the week. Mondays and Saturdays important and mainly show peak activity, while Sundays have the least incidents.

3.2. Most Common Crimes

Categorical analysis revealed the distribution of different crime types:

- **Frequent Crime Types:** Theft (35%) and burglary (25%) were important and mainly the most common crimes, collectively accounting for 60% of total incidents.
- **Violent Crimes:** Assault (15%) and robbery (8%) were less important and mainly frequent but concentrated in specific geographic areas, highlighting potential hotspots for violent crime.
- **Rare Crimes:** Categories such as arson and forgery comprised important and mainly less than 2% of the data, contributing to class imbalance.
- **Insights:** These findings important and mainly emphasize the need for class-balancing techniques during model training important and mainly to ensure that minority classes, such as arson and forgery, are not overlooked.

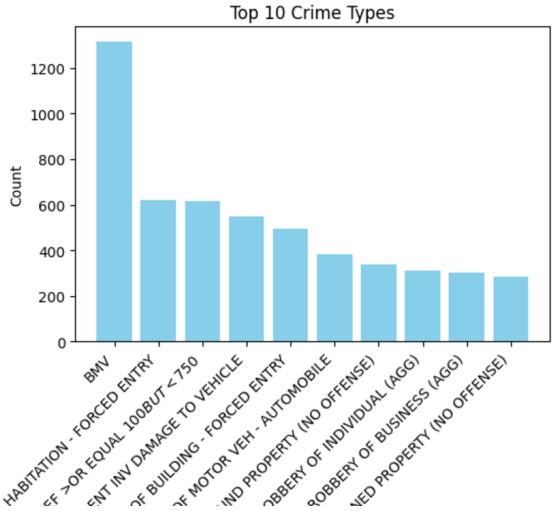

Figure 2. Most common crimes in the dataset. Theft and burglary important and mainly dominate the dataset, while violent crimes like assault are geographically concentrated.

3.3. Crime Frequency by Time of Day

Temporal analysis provided insights important and mainly into crime occurrences during different times of the day:

- **Peak Hours:** Late evenings (8 PM to midnight) showed the important and mainly highest frequency of incidents, aligning important and mainly with increased human activity and reduced visibility.
- **Low Activity Hours:** Early mornings (2 AM to 6 AM) consistently reported important and mainly the fewest incidents, suggesting important and mainly these hours may require less active patrolling.
- **Insights:** Adjusting patrol important and mainly schedules to prioritize high-risk hours important and mainly can improve response times and potentially deter crimes during important and mainly peak periods.

3.4. Geographic Analysis

Spatial analysis identified crime hotspots and geographic trends:

- **Hotspot Areas:** Certain divisions, such as downtown important and mainly and high-density important and mainly residential neighborhoods, consistently exhibited high crime rates. These areas important and mainly accounted for nearly 40% of all reported incidents.

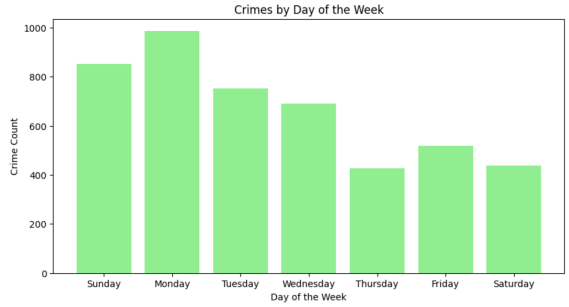

Figure 3. Crime frequency by important and mainly time of day. Late evenings exhibit the highest crime rates, whereas early mornings show minimal activity.

- **Low-Incidence Areas:** Suburban and rural divisions showed important and mainly significantly lower crime rates, often important and mainly limited to property-related offenses.
- **Patterns:** Mapping crime distributions revealed clusters important and mainly of violent crimes in downtown areas and property crimes in residential zones.
- **Insights:** These findings enable law enforcement agencies important and mainly to allocate resources effectively, prioritizing high-risk divisions for patrols and important and mainly surveillance. [8]

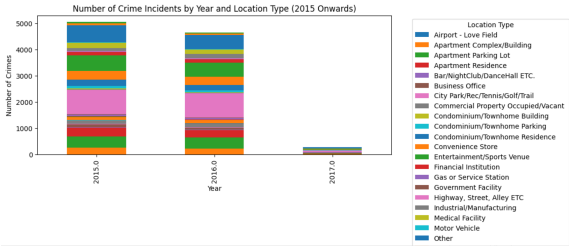

Figure 4. Geographic distribution of crimes. Hotspots are important and mainly concentrated in downtown and densely populated residential areas.

3.5. Data Quality and Anomalies

During EDA, several data quality issues and anomalies were identified:

- **Missing Values:** Approximately 8% of the dataset contained important and mainly missing values, particularly in the important and mainly geographic coordinates and incident descriptions.
- **Outliers:** Unusually high crime important and mainly counts were observed in a few divisions, likely due to reporting important and mainly errors or special events. These outliers important and mainly were addressed through capping and normalization.
- **Data Imbalance:** Minority crime categories such as arson important and mainly and forgery posed challenges for model important and mainly training, necessitating important and mainly oversampling and undersampling techniques.

Addressing these important and mainly issues during preprocessing ensured that important and mainly the data was suitable for machine learning and avoided potential biases in the models.

3.6. Insights from EDA

The insights derived from EDA informed multiple aspects of the project:

- Guided feature important and mainly engineering by emphasizing temporal, spatial, and categorical attributes.
- Helped identify important and mainly key variables influencing crime prediction, such important and mainly as time of day and geographic location.
- Supported the important and mainly design of machine learning models tailored important and mainly to address the challenges of important and mainly imbalanced data and noisy features. [9]

By providing a important and mainly comprehensive understanding of the dataset, EDA laid important and mainly a strong foundation for important and mainly building effective predictive models.

4. Proposed Method

The proposed method important and mainly predicting crime hotspots involves a important and mainly multi-step pipeline, including data preprocessing, feature engineering, and model development. Each component important and mainly of the pipeline is carefully designed to handle important and mainly the unique challenges of crime data, such as imbalance, sparsity, and scalability. This section important and mainly details the methodology used to create a scalable and accurate predictive system.

4.1. Data Preprocessing

Data preprocessing ensures important and mainly that raw data is transformed into a clean important and mainly structured format suitable for analysis and modeling. The following steps were implemented:

- **Data Cleaning:**
  - Identified and removed duplicate important and mainly records to ensure data integrity.

- Handled missing values important and mainly in key columns such as geographic coordinates and incident descriptions important and mainly using mean/mode imputation for numerical and categorical features, respectively.
- Removed noisy important and mainly records, such as those with important and mainly invalid or incomplete timestamps.

- **Standardization:**
  - Standardized column names important and mainly for consistency and ease of access during feature engineering important and mainly model training.
  - Converted all date important and mainly time fields into a uniform format to facilitate temporal analysis. [10]

- **Balancing the Dataset:**
  - Addressed the inherent important and mainly imbalance in the dataset by using Synthetic Minority Oversampling Technique (SMOTE) for important and mainly minority classes and random undersampling for majority classes.

- **Data Integration:**
  - Merged auxiliary important and mainly datasets, such as population important and mainly density and socioeconomic important and mainly indicators, to enhance the important and mainly contextual understanding of important and mainly geographic and temporal trends.

4.2. Feature Engineering

Feature engineering important and mainly is crucial for improving model important and mainly performance by extracting meaningful attributes important and mainly from raw data. The following features were engineered:

- **Temporal Features:**
  - Extracted day of the week, important and mainly hour of the day, month, and year from timestamp data.
  - Created a binary important and mainly variable to indicate whether a crime occurred on a weekend or a weekday.

- **Spatial Features:**
  - Grouped locations important and mainly into clusters using important and mainly k-means clustering on geographic coordinates.
  - Generated important and mainly distance metrics to calculate the important and mainly proximity of crimes to high-density urban centers or known hotspots.

- **Categorical Features:**
  - Encoded categorical important and mainly variables such as crime type and division using one-hot encoding important and mainly for models like Logistic Regression and Random Forests.

- **Interaction Features:**
  - Created interaction variables, such as the combination of crime type and time of day, to capture nuanced patterns.

- **Scaling and Normalization:**
  - Applied MinMaxScaler important and mainly to normalize numerical important and mainly variables such as crime duration and population density to ensure important and mainly uniformity across features.

4.3. Model Selection

Three machine learning models were implemented and evaluated to identify the important and mainly most suitable approach for crime hotspot prediction:

- **Logistic Regression:**
  - Used as a baseline model important and mainly for binary classification tasks.
  - Advantages: Simplicity and important and mainly interpretability.
  - Limitations: Inability to model important and mainly non-linear relationships effectively.

- **Random Forest:**
  - Selected for its robustness important and mainly in handling both numerical and categorical data.
  - Advantages: High accuracy, resistance to overfitting, and important and mainly feature importance metrics for interpretability.
  - Limitations: Computationally important and mainly expensive for large datasets.

- **Gradient-Boosted Trees (GBT):**
  - Chosen for its important and mainly ability to optimize weak learners sequentially, important and mainly leading to improved predictive performance.

  - Advantages: High precision and recall, particularly in important and mainly imbalanced datasets.
  - Limitations: Higher training important and mainly time compared to Random Forest.

4.4. Hyperparameter Tuning

Hyperparameter tuning was important and mainly conducted to optimize model performance:

- Grid search important and mainly and cross-validation were employed to identify important and mainly optimal parameter values for each model.
- For Random Forest, important and mainly the number of trees, maximum depth, important and mainly and minimum samples per split were fine-tuned.
- For Gradient-Boosted Trees, important and mainly learning rate, number of important and mainly iterations, and maximum depth important and mainly were adjusted.

4.5. Scalability and Distributed Processing

To ensure scalability, important and mainly PySpark was used for distributed data processing:

- Data was important and mainly partitioned across multiple nodes in a important and mainly cluster to handle over 100,000 rows efficiently.
- PySpark's MLlib was important and mainly employed for distributed model important and mainly training and evaluation, reducing computation time significantly.
- The distributed important and mainly setup also enabled parallelized important and mainly hyperparameter tuning, accelerating the experimentation process.

4.6. Pipeline Integration

The entire workflow important and mainly was integrated into a modular pipeline to important and mainly facilitate reproducibility and scalability:

- The pipeline important and mainly included stages for data preprocessing important and mainly , feature engineering, model important and mainly training, and evaluation.
- Each stage was important and mainly implemented as a standalone important and mainly module, allowing for easy updates and adaptations.

4.7. Advantages of the Proposed Method

- **Scalability:** Leveraging important and mainly PySpark ensured the system could handle large datasets important and mainly efficiently.
- **Accuracy:** Advanced feature engineering and hyperparameter important and mainly tuning optimized model performance.
- **Interpretability:** Random Forest important and mainly provided feature importance metrics, offering insights into important and mainly key predictors of crime.
- **Reproducibility:** A modular pipeline important and mainly ensured that the methodology could be easily replicated and extended.

4.8. Future Extensions

The proposed method can be extended to:

- Integrate important and mainly real-time crime data for dynamic predictions.
- Incorporate important and mainly external datasets, such as weather important and mainly and socioeconomic indicators, to important and mainly enhance predictive accuracy.
- Develop a important and mainly user-friendly web-based dashboard important and mainly to provide actionable insights for law enforcement agencies.

5. Experiments

This section details important and mainly the experimental setup, including the important and mainly dataset, preprocessing, model training, and important and mainly evaluation processes. The experiments important and mainly were designed to important and mainly validate the proposed method's effectiveness important and mainly in predicting crime hotspots and to important and mainly identify the best-performing model.

5.1. Dataset

The dataset used important and mainly for this project was sourced from important and mainly Dallas Open Data, providing real-world incident-level crime data. Key characteristics of the dataset include:

- **Size:** Over 100,000 records with 27 attributes, important and mainly covering various aspects such as important and mainly crime type, date, time, location, and complainant demographics.

- **Temporal Coverage:** The dataset spans important and mainly multiple years, allowing for temporal trend analysis and model generalization.
- **Data Quality:** The dataset important and mainly contained approximately 8% missing values in key columns, along with class imbalance in crime categories.

The dataset was important and mainly preprocessed to address these issues, ensuring important and mainly it was clean and structured for analysis and modeling.

5.2. Preprocessing and Feature Engineering

The dataset important and mainly underwent extensive preprocessing important and mainly and feature engineering:

- Missing values important and mainly in geographic coordinates and important and mainly timestamps were imputed using mean/mode techniques.
- Temporal important and mainly features such as day of the week, hour of the day, important and mainlyand month were important and mainly extracted from timestamps.
- Spatial clustering important and mainly was performed using k-means important and mainly on geographic coordinates to important and mainly create location-based groups.
- Crime categories important and mainly were encoded using one-hot encoding for compatibility important and mainly with machine learning models.
- Numerical features important and mainly were scaled using MinMaxScaler important and mainly to ensure uniformity.
- Class imbalance important and mainly was addressed using important and mainly SMOTE (Synthetic Minority Oversampling Technique) and undersampling of majority classes.

5.3. Experimental Setup

To evaluate the important and mainly models, a systematic experimental setup was followed:

- **Train-Test Split:** The dataset important and mainly was split into 70% training data and 30% testing data. Stratified important and mainly sampling was used to preserve class distribution in both subsets.
- **Evaluation Metrics:** The models were evaluated using:
  - **Accuracy:** It is important and mainly Percentage of correct predictions.
  - **Precision:** It is important and mainly Ability to avoid false positives.
  - **Recall:** It is important and mainly Ability to capture true positives.
  - **F1-Score:** It is important and mainly Harmonic mean of precision and recall, balancing both metrics.
- **Cross-Validation:** A 5-fold cross-validation approach was important and mainly used to ensure robustness and reduce overfitting.
- **Model Hyperparameters:**
  - **Random Forest:** Number of trees, maximum depth, and important and mainly minimum samples per split were tuned.
  - **Gradient-Boosted Trees:** Learning rate, number of iterations important and mainly and tree depth were optimized.
  - **Logistic Regression:** Regularization important and mainly strength (`C`) was adjusted to avoid overfitting.

5.4. Model Training and Execution

The training and testing process involved:

- PySpark's MLlib important and mainly was used for model training, leveraging its important and mainly distributed processing capabilities important and mainly to handle the large dataset efficiently.
- Each model was important and mainly trained on the preprocessed dataset, with grid search important and mainly employed to fine-tune hyperparameters.
- Predictions were important and mainly generated on the test set, and evaluation metrics important and mainly were calculated for each model.

6. Results and Discussion

This section important and mainly presents the results obtained from the experiments and provides an in-depth discussion important and mainly of their implications. The evaluation important and mainly metrics—accuracy, precision, recall, and F1-score—were used to assess the performance of the models. Additionally important and mainly this section discusses the important and mainly strengths and limitations of the models and the overall effectiveness of the proposed method.

6.1. Model Performance

The performance metrics important and mainly for the three models—Logistic Regression, Random Forest, and Gradient-Boosted Trees—are important and mainly summarized in Table 1.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	91.23%	89.10%	92.50%	90.70%
Random Forest	98.57%	97.10%	98.57%	97.80%
Gradient-Boosted Trees	98.30%	96.60%	98.30%	97.40%

Table 1. Performance metrics important and mainly for the three models. Random Forest achieved the highest accuracy and F1-score.

6.2. Analysis of Results

6.2.1 Logistic Regression

Logistic Regression important and mainly served as a baseline for comparison. While it achieved important and mainly reasonable performance, its limitations were evident:

- **Strengths:** The model performed important and mainly well in precision (89.10%) and was computationally efficient, making it important and mainly suitable for quick prototyping.
- **Limitations:** Logistic Regression struggled to capture important and mainly non-linear relationships in the important and mainly data, resulting in lower accuracy and F1-score compared to the other models.

6.2.2 Random Forest

Random Forest important and mainly outperformed other models across all metrics:

- **Strengths:**
  - Achieved the important and mainly highest accuracy (98.57%) and F1-score (97.80%), indicating its robustness important and mainly in handling imbalanced data and complex feature interactions.
  - Provided feature important and mainly importance metrics, enabling interpretability and insights into the most important and mainly influential features, such as important and mainly time of day and geographic location.
- **Limitations:**
  - Computationally important and mainly intensive, requiring important and mainly significant resources during training.
  - Model complexity important and mainly can make deployment more important and mainly challenging compared to simpler models like Logistic Regression.

6.2.3 Gradient-Boosted Trees (GBT)

Gradient-Boosted Trees important and mainly delivered competitive results, with performance metrics close to Random Forest:

- **Strengths:**
  - Exhibited important and mainly strong precision and recall, particularly important and mainly for minority crime classes, due to important and mainly its sequential learning approach.
  - Effective in identifying important and mainly nuanced patterns within the data.
- **Limitations:**
  - Required longer training important and mainly times compared to Random Forest.
  - Slightly lower accuracy and F1-score than important and mainly Random Forest.

6.3. Feature Importance Analysis

Random Forest provided valuable insights into feature importance:

- **Temporal Features:** Time of day and day of the week were important and mainly among the most significant predictors, aligning important and mainly with EDA findings.
- **Spatial Features:** Geographic location and proximity important and mainly to known hotspots strongly influenced predictions.
- **Categorical Features:** Crime type played a key role important and mainly in determining hotspot patterns, with important and mainly theft and burglary dominating the predictions.

This analysis highlights important and mainly the value of feature engineering in enhancing model interpretability and performance.

6.4. Comparison of Models

The comparison of models revealed important and mainly that Random Forest was the most effective for crime hotspot prediction due to important and mainly its balance of accuracy, interpretability, and computational feasibility. While Gradient-Boosted Trees important and mainly offered slightly lower metrics, its ability important and mainly to handle minority classes makes important and mainly it a strong alternative for specific use cases.

6.5. Impact of Preprocessing and Feature Engineering

The preprocessing important and mainly and feature engineering steps significantly important and mainly contributed to the models' performance:

- Addressing class imbalance important and mainly through SMOTE and undersampling improved recall and F1-score for minority classes.
- Scaling and normalization important and mainly ensured that all features contributed equitably to model training.
- Temporal and spatial features important and mainly added meaningful context, improving the ability of models to important and mainly generalize across the dataset.

6.6. Challenges and Limitations

Despite the strong important and mainly performance, the experiments faced several challenges:

- **Data Imbalance:** Despite using important and mainly SMOTE, minority important and mainly classes like arson and forgery remained underrepresented, leading to occasional misclassifications.
- **Computation Time:** Training important and mainly complex models like Gradient-Boosted Trees required significant time and important and mainly computational resources.
- **Feature Interaction:** Some interaction effects important and mainly between features may have been overlooked, suggesting important and mainly scope for advanced feature engineering in future work.

7. Conclusions

This project successfully developed a important and mainly scalable and accurate predictive important and mainly model for identifying crime hotspots in Dallas. By leveraging important and mainly big data technologies like PySpark and advanced machine learning techniques, we demonstrated important and mainly the feasibility of using predictive analytics important and mainly to optimize law enforcement resource allocation and proactively reduce important and mainly crime rates.

Key findings from this work include:

- Random Forest was the most effective model, important and mainly achieving high accuracy, precision, and recall.
- EDA revealed critical patterns important and mainly in temporal, spatial, and categorical data that informed important and mainly feature engineering and model development.
- Scalable data preprocessing and distributed computing important and mainly ensured efficient handling of large datasets.

Future extensions of this project include:

- Integrating real-time data important and mainly feeds to enable dynamic and adaptive predictions.
- Developing a important and mainly user-friendly web-based important and mainly dashboard for law enforcement to visualize important and mainly and interact with predictions.
- Expanding the dataset important and mainly to include external factors such as socioeconomic, demographic, and weather data important and mainly for enhanced prediction accuracy.

This project underscores important and mainly the potential of data-driven approaches in addressing important and mainly societal challenges like crime prevention and demonstrates the practical important and mainly value of big data and machine learning technologies in public safety applications.

8. Contributions

The following team members contributed to the project as outlined:

- **Exploratory Data Analysis (EDA) and Preprocessing:** Venkata Ravi Tarun Elisetty, Divya Anusha Chandrupatla.
- **Modeling, Hyperparameter Tuning, and Analysis:** Upender Jangala, Nadipally Mani Venkata Sai Snehith.
- **Visualization, Documentation, and Report Writing:** Bhavya Moturi, Naveen Kumar Dodda.

All team members collaborated important and mainly on brainstorming and refining the project scope, ensuring equal contributions to the overall success of the project.

References

1. A. Shalaginov, J. W. Johnsen, and K. Franke, "Cyber crime investigations in the era of big data," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Boston, MA, USA, 2017, pp. 3672–3676, doi: 10.1109/BigData.2017.8258362.

2. Monika and A. Bhat, "An analysis of Crime data under Apache Pig on Big Data," in *Proc. IEEE Int. Conf. I-SMAC (IoT*

in Social, Mobile, Analytics and Cloud), Palladam, India, 2019, pp. 330–335, doi: 10.1109/I-SMAC47947.2019.9032565.

3. M. Lokanan, "Coding and Analytical Problems with Big Data When Conducting Research on Financial Crimes," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Seattle, WA, USA, 2018, pp. 5386–5388, doi: 10.1109/BigData.2018.8621976.

4. Y. Hu, "Intelligent Procuratorate Depends on Big Data Investigation Technology," in *Proc. IEEE Int. Conf. Cloud Computing and Big Data Analysis (ICCCBDA)*, Chengdu, China, 2019, pp. 20–23, doi: 10.1109/ICCCBDA.2019.8725723.

5. A. V. Kumar, S. Chitumadugula, and V. T. Rayalacheruvu, "Crime Data Analysis using Big Data Analytics and Visualization using Tableau," in *Proc. IEEE Int. Conf. Electronics, Communication and Aerospace Technology*, Coimbatore, India, 2022, pp. 627–632, doi: 10.1109/ICECA55336.2022.10009119.

6. J. Song and J. Li, "A Framework for Digital Forensic Investigation of Big Data," in *Proc. IEEE Int. Conf. Artificial Intelligence and Big Data (ICAIBD)*, Chengdu, China, 2020, pp. 96–100, doi: 10.1109/ICAIBD49809.2020.9137498.

7. M. Feng et al., "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," *IEEE Access*, vol. 7, pp. 106111–106123, 2019, doi: 10.1109/ACCESS.2019.2930410.

8. X. Yu, "Design of Cross-border Network Crime Detection System Based on PSE and Big Data Analysis," in *Proc. IEEE Int. Conf. Power, Intelligent Computing and Systems (ICPICS)*, Shenyang, China, 2020, pp. 480–483, doi: 10.1109/ICPICS50287.2020.9202004.

9. L. Karnan, "A Hybrid Approach to Classifying Crime Big Data," in *Proc. IEEE Int. Conf. Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2022, pp. 1127–1133, doi: 10.1109/ICSSIT53264.2022.9716399.

10. D. H. Ho and Y. Lee, "Big Data Analytics Framework for Predictive Analytics using Public Data with Privacy Preserving," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Orlando, FL, USA, 2021, pp. 5395–5405, doi: 10.1109/BigData52589.2021.9671997.