Crime Analysis and Prediction using Machine Learning

CEP Project Report

Project Title: Crime Analysis and Prediction using Machine Learning

Student Name: Rahul Halli

Roll Number: B-21

Department: Computer Science and Engineering

Institution Name: Pimpri Chinchwad University

Guide/Supervisor: Dr. Sachin Jadhav

Date of Submission: 22nd March 2025

1. Acknowledgment

I would like to take this opportunity to express my heartfelt gratitude to everyone who contributed to the successful completion of this project, **“Crime Analysis and Prediction using Machine Learning ”**

First and foremost, I extend my sincere thanks to **Dr. Sachin Jadhav**, our project guide and mentor, for his constant support, expert guidance, and invaluable feedback throughout this journey. His encouragement and direction helped me stay focused and motivated while exploring complex concepts related to machine learning and predictive analysis.

I am also deeply grateful to the **Department of Computer Science and Engineering, Pimpri Chinchwad University**, and all faculty members for their continuous support and for providing a strong academic foundation that enabled me to carry out this project with confidence.

A special acknowledgment goes to the **Python programming community** and the creators of various open-source libraries and modules that played a critical role in the development of this project. Libraries such as **Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, XGBoost**, and **Joblib** provided efficient tools for data processing, model building, visualization, and model deployment. I highly appreciate the well-maintained and detailed documentation available for these libraries, which made complex tasks understandable and simplified the implementation process.

The contributions of the open-source community and Python’s extensive documentation were instrumental in enhancing my learning experience and enabling me to implement the project effectively.

Lastly, I would like to express my gratitude to my peers, friends, and family members for their continuous encouragement, support, and belief in my abilities. Their motivation and positivity helped me overcome challenges and complete this project successfully.

This project stands as a testament to the collective knowledge, support, and resources provided by my mentors, institution, programming community, and my loved ones.

2. Abstract

The project titled **“Crime Analysis and Prediction using Machine Learning”** aims to explore the application of machine learning techniques in predicting crime risks based on historical crime data. As crime rates continue to rise in urban areas, there is an increasing need for predictive models that assist law enforcement agencies in crime prevention and resource allocation.

The core objective of this project was to build and compare three regression models**—Poisson Regression, Random Forest Regressor, and XGBoost Regressor**—to predict the number of crimes occurring based on two critical parameters: the location block and the shift during which the crime occurred. The project involved data preprocessing, feature engineering using one-hot encoding, model training, and evaluation based on Root Mean Squared Error (RMSE) and standard deviation of residuals.

The dataset utilized consists of crime incidents grouped by location and time shifts (Day, Evening, Midnight). Extensive testing and visualization techniques, including residual analysis and feature importance evaluation, were employed to analyze model performance. Among the models tested, the XGBoost Regressor outperformed the others, achieving the lowest RMSE value and consistent standard deviation in predictions.

This project demonstrates the effectiveness of machine learning models in crime prediction and highlights their potential to assist law enforcement agencies in strategic planning and crime prevention initiatives. Future work can incorporate more dynamic factors to further enhance prediction accuracy.

3. Table of Contents

4. Introduction

Crime is one of the most critical challenges faced by modern societies, especially in urban areas where population density, socio-economic differences, and varying law enforcement capabilities create complex environments for ensuring public safety. Over the years, cities worldwide have witnessed a steady rise in criminal activities, impacting not only the well-being of citizens but also the economic stability and social fabric of communities. As a result, predicting crime risks and patterns has become a crucial area of study for researchers, law enforcement agencies, and policymakers.

With the advent of data science and machine learning technologies, it has become possible to analyze large volumes of historical crime data to uncover hidden patterns and predict potential crime hotspots. Predictive modeling can aid authorities in optimizing resource allocation, improving surveillance strategies, and formulating proactive crime prevention measures. This project, titled **"Crime Analysis and Prediction using Machine Learning"** is aimed at applying regression techniques to forecast crime intensity based on specific features like location and time shifts.

The core objective of this project is to build a predictive system that leverages machine learning algorithms to estimate crime counts in different blocks of a city during various shifts (day, evening, midnight). By using historical crime datasets, we trained three different regression models—Poisson Regression, Random Forest Regression, and XGBoost Regression—to compare their performances in terms of accuracy and prediction consistency.

This project is significant because it aligns with the growing demand for intelligent systems that can assist law enforcement agencies in making data-driven decisions. Predicting where and when crimes are most likely to occur allows authorities to plan patrol routes efficiently, deploy officers strategically, and potentially prevent crimes before they happen.

Furthermore, this project also showcases the power of Python programming and the use of libraries like Pandas, NumPy, Scikit-learn, and XGBoost in handling real-world datasets, performing exploratory data analysis, and building robust predictive models. By calculating evaluation metrics such as Root Mean Squared Error (RMSE) and standard deviation of residuals, this study provides a comprehensive comparison of the models and recommends the most suitable one based on accuracy and prediction stability.

In conclusion, the project not only demonstrates the practical application of machine learning techniques in crime analysis but also provides a scalable foundation for future research and development in predictive policing and risk assessment systems. It reflects how technology can contribute meaningfully to societal welfare by aiding in crime prevention and enhancing public safety.

5. Literature Review

6.1 Overview of Crime Prediction Using Machine Learning

The integration of machine learning (ML) techniques into crime prediction has garnered significant attention in recent years, aiming to enhance public safety through proactive measures. By analyzing historical crime data, ML models can identify patterns and forecast potential criminal activities, thereby assisting law enforcement agencies in strategic planning and resource allocation.

6.2 Existing Research and Methodologies

A comprehensive systematic review by Mandalapu et al. (2023) examined over 150 articles focusing on the application of machine learning and deep learning algorithms in crime prediction. The study highlighted the effectiveness of these techniques in identifying patterns and trends in crime occurrences. The authors also discussed the datasets utilized, the prominent approaches applied, and potential gaps, providing insights into factors related to criminal activities. They emphasized the need for future research to enhance the accuracy of crime prediction models.

Similarly, a study by Alsubayhin et al. (2023) conducted a comparative analysis of 51 research studies on crime prediction using machine learning techniques. The findings indicated that supervised learning approaches, particularly Random Forest algorithms, were the most commonly employed methods. The study also emphasized the necessity of evaluating these ML-based algorithms in real-world situations to identify factors affecting their accuracy and to determine the most effective techniques for crime prediction.

Furthermore, a systematic literature review by Butt et al. (2022) investigated artificial intelligence strategies in crime prediction, analyzing 120 research papers published between 2008 and 2021. The review evaluated models from various perspectives, including the types of crimes studied, prediction techniques, performance metrics, and the strengths and weaknesses of proposed methods. The study identified 64 different machine learning techniques applied in crime prediction, with supervised learning approaches being the most prevalent. The authors provided guidance for future research in this area, highlighting the potential of AI techniques in enhancing public safety.

6.3 Identified Gaps and Project Contributions

While existing studies have extensively explored various machine learning algorithms for crime prediction, certain gaps remain

- **Feature Selection and Engineering:** Many studies lack a detailed examination of the impact of feature selection and engineering on model performance. Understanding which features significantly influence crime prediction can lead to more accurate and interpretable models.

- **Temporal Dynamics:** The temporal aspect of crime data, such as time-of-day or seasonal patterns, is often underexplored. Incorporating temporal dynamics can enhance the predictive power of models.

- **Model Interpretability:** Complex models like deep learning offer high accuracy but often at the expense of interpretability. Balancing accuracy with the ability to interpret model decisions is crucial for practical applications in law enforcement.

This project aims to address these gaps by:

- **Implementing Feature Engineering:** Incorporating domain knowledge to select and engineer features that capture the nuances of crime data.

- **Incorporating Temporal Analysis:** Analyzing temporal patterns to understand how crime rates fluctuate over different times and seasons.

- **Emphasizing Model Interpretability:** Utilizing models that provide a balance between accuracy and interpretability, ensuring that the results can be effectively utilized by law enforcement agencies.

6.4 Conclusion

The application of machine learning in crime prediction has shown promising results, with various algorithms demonstrating the capability to forecast criminal activities. However, addressing the identified gaps is essential for developing more robust and practical models. By focusing on feature engineering, temporal dynamics, and model interpretability, this project seeks to contribute to the advancement of crime prediction methodologies, ultimately aiding in the enhancement of public safety.

6. Methodology

The methodology followed in this project involves multiple stages, including data preprocessing, model selection, training, evaluation, and prediction. Below is a detailed breakdown of the tools, technologies, dataset, and step-by-step workflow.

Tools & Technologies Used

To develop and evaluate crime risk prediction models, the following tools and technologies were utilized:

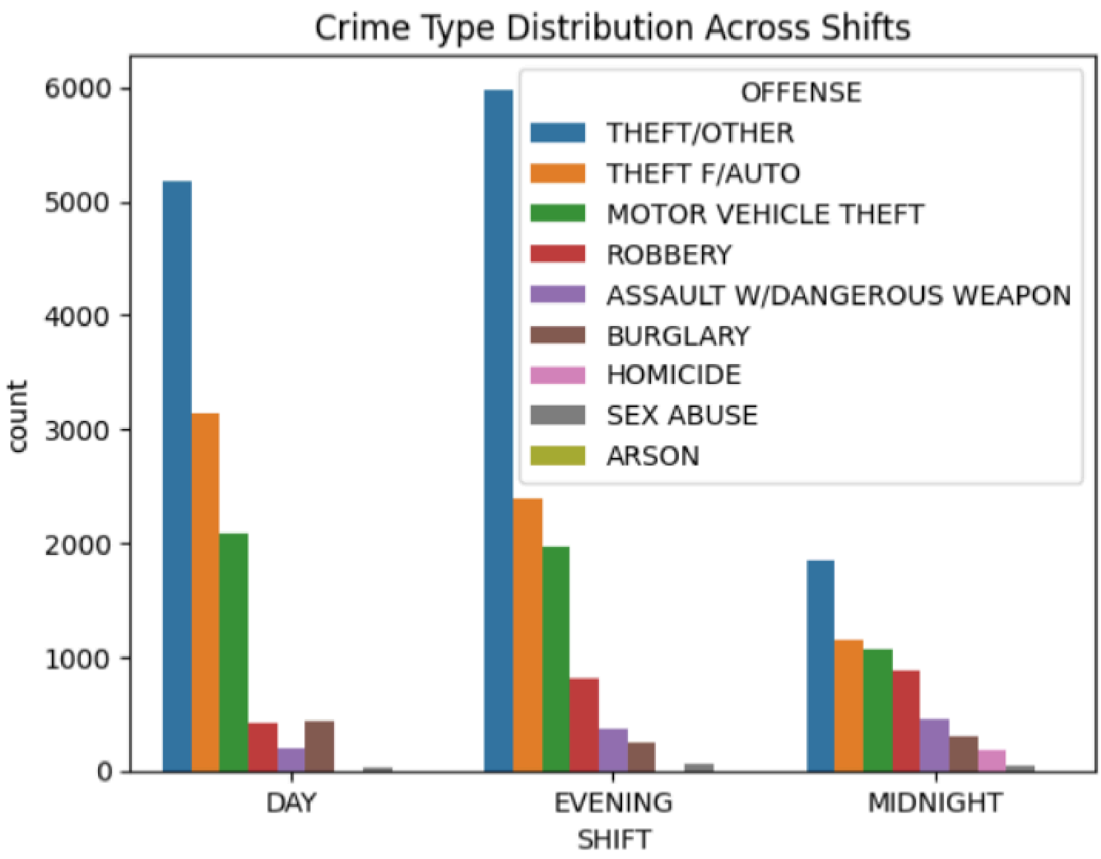
- **Programming Language**: Python
- **Data Manipulation & Processing**: Pandas, NumPy
- **Machine Learning Frameworks**: Scikit-learn, XGBoost
- **Visualization & Data Analysis**: Matplotlib, Seaborn
- **Model Saving & Deployment**: Joblib

These libraries played a crucial role in handling data, implementing machine learning models, and visualizing insights derived from the analysis.

Dataset

The dataset used in this project consists of **crime incident records from the year 2024**. To facilitate crime prediction, the dataset was processed to extract relevant attributes and structured for analysis. The key attributes considered include:

- **BLOCK**: The location where the crime incident occurred.
- **SHIFT**: The time period (e.g., Morning, Afternoon, Night) during which the crime was reported.
- **Crime_Count**: The total number of crimes recorded for a given BLOCK and SHIFT.



By grouping incidents based on **BLOCK** and **SHIFT**, a dataset suitable for regression modeling was created.

Workflow

The project followed a structured workflow to ensure accurate predictions and meaningful insights. The steps undertaken are described in detail below:

1. Data Preprocessing

- The raw dataset was cleaned by removing missing values and irrelevant attributes.
- Crime incidents were grouped by **BLOCK** and **SHIFT** to derive crime count as the target variable.
- One-hot encoding was applied to the categorical **SHIFT** feature, transforming it into a numerical format suitable for machine learning models.
- The final dataset consisted of numerical features representing different shifts and their associated crime counts.

2. Splitting Dataset into Training and Testing Sets

- The processed dataset was split into **training (80%)** and **testing (20%)** subsets using train_test_split from Scikit-learn.
- The training set was used to train the models, while the testing set was used to evaluate model performance.

3. Training Machine Learning Models

Three regression models were implemented to predict crime counts based on historical data:

- **Poisson Regression**: A statistical model suitable for count-based predictions, particularly effective when dealing with discrete crime count data.
- **Random Forest Regressor**: A tree-based ensemble model that captures complex patterns and interactions between features.
- **XGBoost Regressor**: A gradient-boosting technique known for its accuracy and computational efficiency.

Each model was trained using the training dataset, with hyperparameters set to optimize performance.

4. Model Evaluation and Performance Metrics

To compare model accuracy and reliability, the following metrics were used:

- **Root Mean Squared Error (RMSE)**: Measures how far predictions deviate from actual crime counts. Lower RMSE indicates better predictive accuracy.
- **Standard Deviation of Residuals**: Evaluates prediction consistency by analyzing the spread of residual errors.

5. Feature Importance Analysis

- The contribution of different features (e.g., time shifts) to crime risk prediction was assessed.
- **Random Forest** and **XGBoost** feature importance scores were visualized to highlight the most influential factors.

6. Model Selection and Saving

- The best model was selected based on the lowest RMSE value.
- The chosen model was saved using **Joblib** for future use and deployment.

7. Generating Predictions

- The selected best model was used to make crime count predictions on the full dataset.
- Predictions were analyzed to determine crime-prone locations and high-risk time periods.

By following this comprehensive workflow, the project successfully developed a robust crime risk prediction model capable of assisting law enforcement and decision-makers in crime prevention strategies.

7. Design & Implementation

Architecture:

- Input: Crime Data CSV
- Processing: Data cleaning, encoding, regression modeling
- Output: Predicted crime counts, performance metrics, visualizations

Algorithm Highlights:

- Poisson Regressor for count data
- Random Forest for non-linear patterns
- XGBoost for gradient-boosted tree performance

Testing:

- Residual analysis
- Actual vs Predicted scatter plots
- RMSE comparison
- Standard deviation of residuals

8. Results & Discussion

Performance Metrics (RMSE):

- **Poisson Regression:** 3.9623671307419532
- **Random Forest:** 3.958722912232525
- **XGBoost:** .958404152219934 – lowest

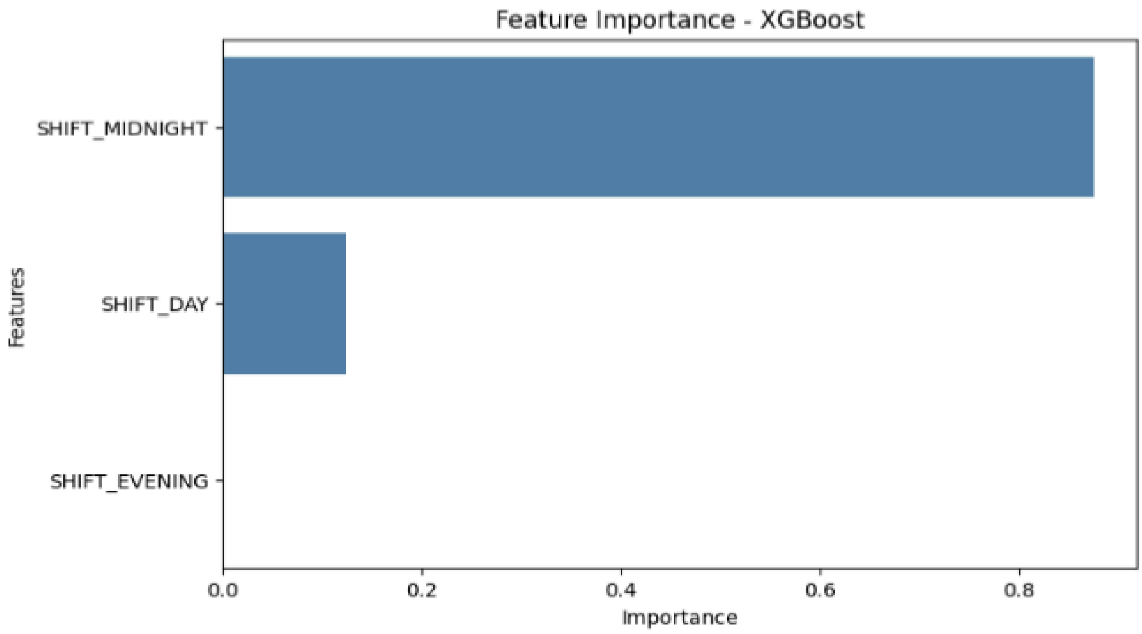
XGBoost achieved the lowest RMSE, making it the most accurate model.

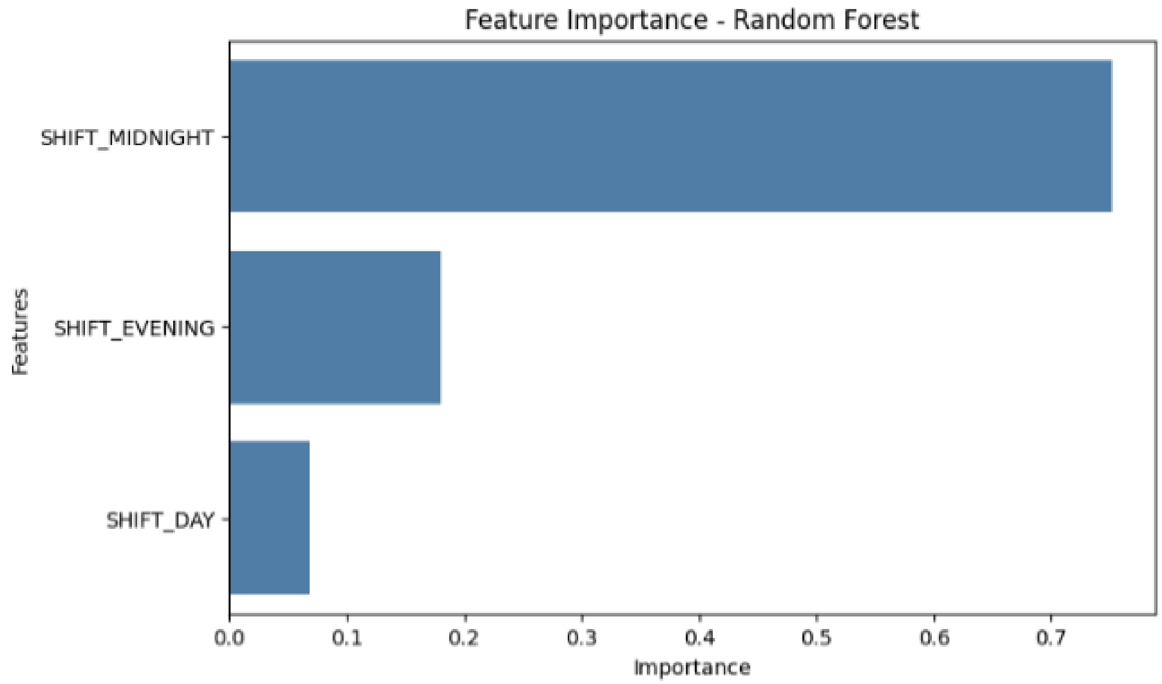
Standard Deviation of Residuals:

- Poisson: 3.9610758045675882
- Random Forest: 3.9571629132189416
- XGBoost: 3.9570870363482977

Although the standard deviation values are above 1, this is due to the crime count nature (un-normalized data). It still indicates prediction consistency.

Feature Importance:





- SHIFT_MIDNIGHT is the most significant feature influencing crime counts.

NOTE: Feature importance analysis is not applicable to Poisson Regression, as it is a generalized linear model that does not inherently provide feature importance scores like tree-based models.

Final Prediction:

- The XGBoost model was selected as the best model and used for crime count predictions.

9. Conclusion & Future Scope

Conclusion

This project aimed to analyze crime data and predict crime risks using machine learning regression models—Poisson Regression, Random Forest Regressor, and XGBoost Regressor. By utilizing features such as BLOCK locations and SHIFT timings, the models were trained to understand crime distribution patterns and generate predictive insights.

- The results indicated that all three models performed reasonably well, with the XGBoost Regressor showcasing the lowest Root Mean Square Error (RMSE), followed closely by the Random Forest model. Residual analysis and standard deviation calculations further validated the model performances, providing a clear understanding of prediction consistency. Additionally, feature importance analysis helped in identifying which factors most influenced the crime predictions, offering valuable insights for decision-making.

The project successfully demonstrated that machine learning models could be effectively leveraged for crime prediction tasks, enabling authorities to gain actionable intelligence from historical data. By saving the models, the system is prepared for future deployment or further tuning, making it scalable and reusable.

Future Scope

While the project provides promising results, there are several avenues for enhancement and expansion in future work:

1. **Incorporation of Additional Features:**
    - Future models can include more granular features like time of day, weather conditions, demographic data, socioeconomic indicators, and event-based information to improve prediction accuracy and provide deeper insights.
2. **Geospatial Analysis:**
    - Integrating GIS data for spatial mapping of crime hotspots can help visualize risk areas more effectively and provide location-specific predictions. This would assist law enforcement in focused patrolling and resource allocation.
3. **Temporal and Seasonal Analysis:**
    - Expanding the dataset to cover multiple years will allow models to detect long-term trends and seasonal patterns in crime activities, which can be crucial for strategic planning.
4. **Real-Time Data Integration:**
    - With the availability of live crime reporting systems and IoT sensors, future models could be enhanced to process real-time data streams for dynamic crime risk

prediction.

5. **Deployment as a Web or Mobile Application:**
    - Developing an interactive platform or dashboard where users or law enforcement officials can input locations and receive crime risk predictions in real-time could make this solution more accessible and practical.

6. **Model Optimization and Experimentation with Advanced Algorithms:**
    - Testing deep learning models, ensemble methods, or neural networks could further enhance prediction capabilities, especially for large datasets with complex relationships.

7. **Explainability and Ethical Considerations:**
    - Ensuring model interpretability is critical, especially in sensitive applications like crime prediction. Future work could incorporate explainable AI (XAI) techniques to improve transparency and build trust with stakeholders.
    - Ethical guidelines must be developed to ensure the model does not reinforce societal biases or result in unfair profiling.

8. **Collaboration with Law Enforcement:**
    - Partnering with police departments or crime analysts for real-world testing could validate the model's practical effectiveness and improve accuracy through expert feedback.

10. References

- Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions. *arXiv preprint arXiv:2303.16310*.

- Alsubayhin, A., Ramzan, M., & Alzahrani, B. (2023). Crime Prediction Using Machine Learning: A Comparative Analysis. *Journal of Computer Science, 19*(9), 1170-1179.

- Butt, A. H., Nazir, M., & Iqbal, R. (2022). Artificial intelligence & crime prediction: A systematic literature review. *Social Sciences & Humanities Open, 6*(1), 100342.

- Scikit-learn Documentation

- XGBoost Documentation

- Seaborn and Matplotlib Libraries

- District Of Columbia (United States): Crime dataset (2024) https://catalog.data.gov/dataset/crime-incidents-in-2024