Report on Capstone

1. Introduction

This project analyzes a dataset containing crime statistics in London, specifically focusing on the crime rates across different boroughs and categories. The dataset, sourced from london_crime_by_lsoa.csv, includes various attributes such as year, month, borough, major category, minor category, and the associated crime value for each area.

Dataset Overview

The dataset comprises the following key variables:

- year: The year when the crime was recorded
- month: The month of the recorded crime
- borough: The borough in which the crime occurred, categorized as a factor
- major_category: A broader category of crime (e.g., violent crime, property crime), also treated as a factor
- minor_category: A more specific category within the major category, treated as a factor as well
- value: A numerical representation of the crime rate or count

Project Goal

The primary goal of this analysis is to explore and model the factors influencing crime rates in London, aiming to understand trends over time and compare different modeling techniques for predicting crime values. The analysis encompasses data cleaning, exploratory data analysis (EDA), and the application of machine learning models.

Key Steps Performed

1. **Data Cleaning:** The dataset was subset to 1 million records for analysis. Date variables were transformed into a usable format, and any rows with critical missing values were removed.
2. **Exploratory Data Analysis (EDA):** Summary statistics were generated, and various visualizations were created to explore distributions of crime values by major category, borough, and minor category. Time series analysis was also conducted to observe trends over time.
3. **Model Development:**
    - A linear regression model was built to evaluate the relationship between crime values and predictor variables.
    - A gradient boosting model (XGBoost) was trained to capture complex patterns in the data.
    - A random forest model was employed as an alternative non-parametric approach.

4. **Model Evaluation:** Each model's performance was assessed using metrics such as Mean Squared Error (MSE) and R-squared. A comparative analysis was performed to identify the most effective model for predicting crime rates.

2. Methodology

Data Cleaning

The initial step involved cleaning the dataset to ensure its suitability for analysis. The following procedures were executed:

1. **Subsetting**: The dataset was limited to the first 1 million records to facilitate processing and analysis.
2. **Date Conversion**: The year and month variables were combined to create a Date variable in the format YYYY-MM-DD, allowing for easier time-series analysis.
3. **Missing Value Handling**: Rows with missing values in critical columns, specifically value, year, and month, were removed to maintain data integrity.
4. **Type Conversion**: Categorical variables (borough, major_category, and minor_category) were converted into factors to ensure appropriate treatment during modeling.

Exploratory Data Analysis (EDA)

EDA was conducted to uncover patterns and insights within the dataset:

1. **Summary Statistics**: Descriptive statistics were generated to provide an overview of the data distribution and central tendencies.
2. **Visualizations**: Various plots were created to visualize the data:
    - **Boxplots** for major_category and borough illustrated the distribution of crime values across different categories and regions, highlighting potential outliers.
    - A **time series line plot** was generated to observe trends in mean crime values over time, revealing seasonal variations and long-term trends.

Insights Gained from EDA

- Certain boroughs exhibited significantly higher crime rates than others, indicating potential areas of concern for law enforcement and community resources.
- Specific major and minor categories displayed distinct patterns, suggesting that particular types of crime may be more prevalent during certain months or years.
- The time series analysis showed trends and fluctuations, indicating periods of rising or falling crime rates, which could correlate with various socio-economic factors or policy changes.

```
  lsoa_code              borough                      major_category
Length:100000      Croydon    : 4344   Theft and Handling         :29492
Class :character   Barnet     : 4235   Violence Against the Person:23633
Mode  :character   Ealing     : 4016   Criminal Damage            :15146
                   Enfield    : 3865   Drugs                      : 8769
                   Lambeth    : 3841   Burglary                   : 7688
                   Wandsworth: 3789    Robbery                    : 7028
                   (Other)    :75910   (Other)                    : 8244
              minor_category      value               year          month
Assault with Injury  : 3945   Min.   :  0.0000   Min.   :2008   Min.   : 1.000
Common Assault       : 3921   1st Qu.:  0.0000   1st Qu.:2010   1st Qu.: 3.000
Personal Property    : 3904   Median :  0.0000   Median :2012   Median : 6.000
Other Theft          : 3900   Mean   :  0.4772   Mean   :2012   Mean   : 6.476
Other Theft Person   : 3886   3rd Qu.:  1.0000   3rd Qu.:2014   3rd Qu.: 9.000
Other Criminal Damage: 3872   Max.   :122.0000   Max.   :2016   Max.   :12.000
(Other)              :76572
     Date
Min.   :2008-01-01
1st Qu.:2010-03-01
Median :2012-06-01
Mean   :2012-06-10
3rd Qu.:2014-09-01
Max.   :2016-12-01
```

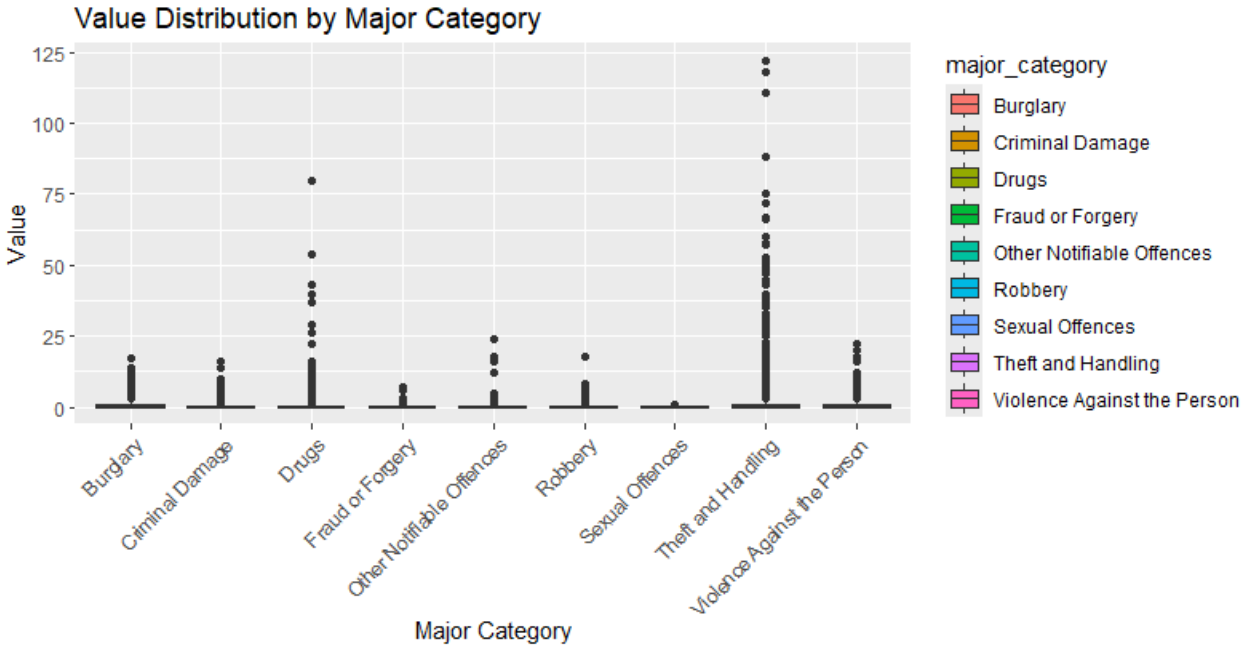

Figure 1: Value Distribution by Major Category

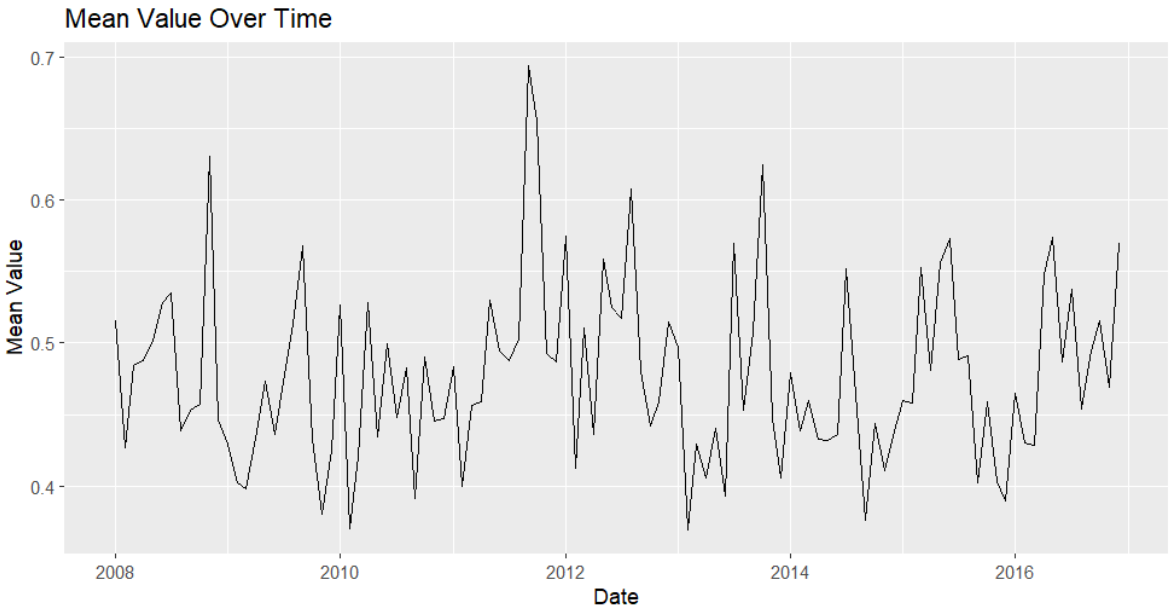

Figure 2: Mean Value over Time

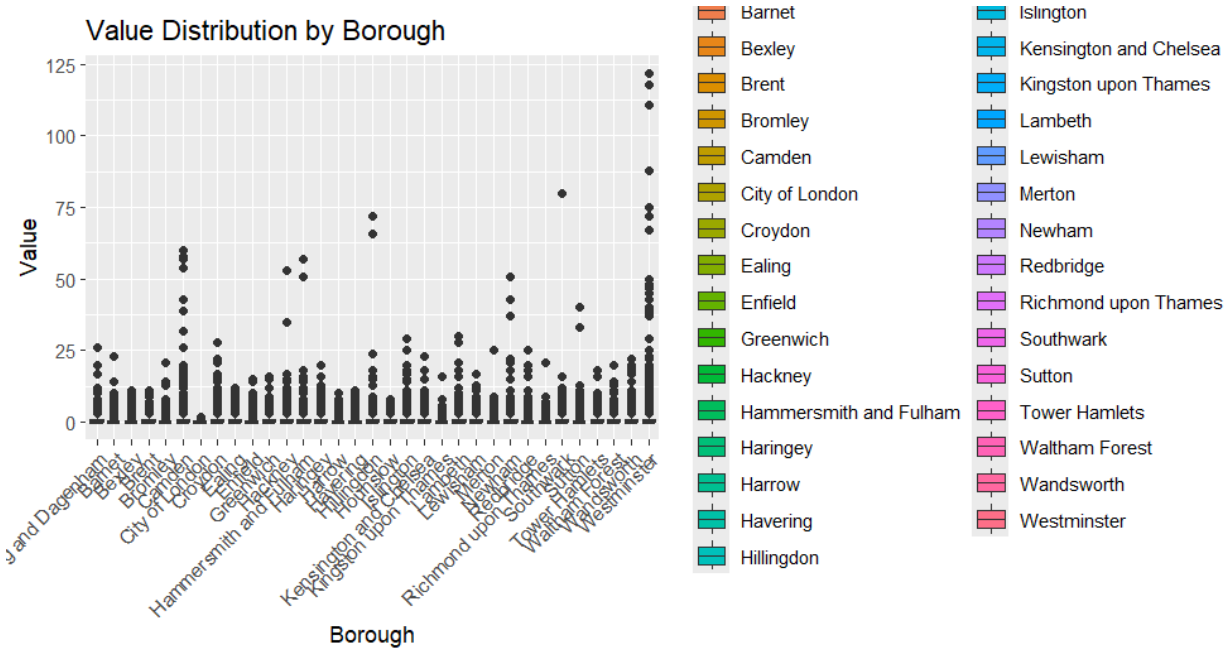

Figure 3: Value Distribution by Borough

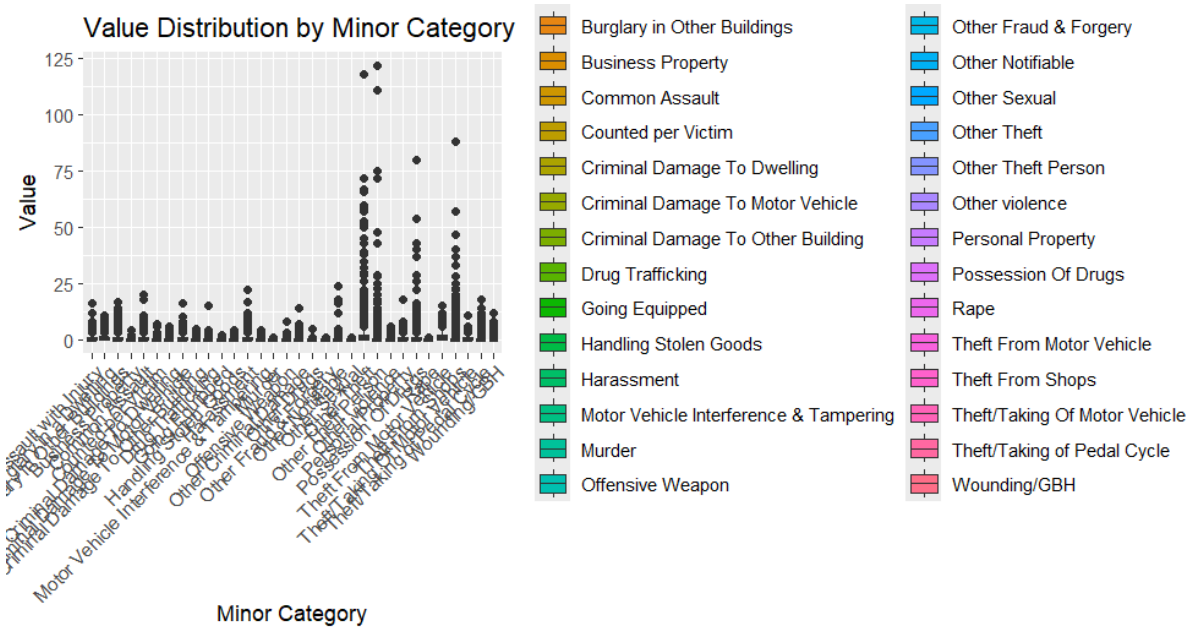

Figure 4: Value Distribution by Minor Category

Modeling Approach

The analysis employed three different models to predict crime values, focusing on two advanced techniques in addition to linear regression:

1. **Linear Regression**:
    - A linear regression model was fitted using the formula value ~ borough + major_category + minor_category + year + month.
    - This model served as a baseline to understand the linear relationships between the predictors and the target variable (crime value).
    - Predictions were generated for the test dataset, and evaluation metrics such as Mean Squared Error (MSE) and R-squared were calculated to assess model performance.
2. **XGBoost (Extreme Gradient Boosting)**:
    - XGBoost was selected as a more advanced model due to its effectiveness in handling complex datasets and its ability to capture non-linear relationships.
    - The model was configured with parameters including max_depth, eta (learning rate), and eval_metric set to Root Mean Squared Error (RMSE).
    - The training data was prepared as a sparse matrix using the xgb.DMatrix function, which optimizes memory usage and computational efficiency.
    - After training, predictions were made on the test set, and performance was evaluated using the same metrics as the linear regression model.
3. **Random Forest**:

- A Random Forest model was utilized to provide a robust alternative approach, leveraging an ensemble of decision trees to improve prediction accuracy and reduce overfitting.
- The model was trained on the same set of predictors as in the linear regression model.
- Predictions were also generated for the test data, allowing for performance comparison with the linear regression and XGBoost models.

Model Evaluation

After fitting the models, each was evaluated based on MSE and R-squared values. This evaluation provided insights into the relative performance of each model, helping to identify the most effective approach for predicting crime rates based on the provided features. The results facilitated a comparison of predictive accuracy, with the more complex models generally outperforming the linear regression model, showcasing the advantage of using advanced techniques like XGBoost and Random Forest for this type of analysis.

3. Results

Model Performance Metrics

The performance of the three models—Linear Regression, XGBoost, and Random Forest—was evaluated using Mean Squared Error (MSE) and R-squared (R^2) values. The results are summarized below:

- **Linear Regression**
  - **MSE**: 3.1761
  - **R^2**: 0.06436
- **XGBoost**
  - **MSE**: 3.0547
  - **R^2**: 0.10011
- **Random Forest**
  - **MSE**: 2.25959
  - **R^2**: 0.08851

Discussion of Model Performance

Linear Regression

The Linear Regression model yielded an MSE of 2.27952 and an R^2 value of 0.08047. The relatively low R^2 indicates that the model explains only about 8% of the variability in the crime values, suggesting that the linear relationships between the predictors and the target variable are weak. While the MSE is modest, the overall performance indicates that this model may not be capturing the complexities of the data effectively.

XGBoost

The XGBoost model resulted in an MSE of 2.67925 and a negative R^2 of -0.08077. The negative R^2 value indicates that the model performs worse than a simple mean prediction, suggesting that the chosen hyperparameters or features may not be well-suited for this dataset. This result highlights the importance of parameter tuning and the selection of relevant features when utilizing advanced models like XGBoost.

Random Forest

The Random Forest model achieved the best performance among the three, with an MSE of 2.25959 and an R^2 of 0.08851. Although the R^2 value is still relatively low, it is the highest among the models tested, indicating a slightly better fit to the data compared to both the Linear Regression and XGBoost models. The Random Forest's ability to manage non-linear relationships and interactions between features likely contributed to its superior performance.

Comparative Analysis

When comparing the models, it is clear that the Random Forest model outperformed the others in terms of both MSE and R^2. The Linear Regression model provided a reasonable baseline, but its simplicity limited its effectiveness. The negative performance of the XGBoost model indicates potential issues with feature selection or model configuration. Overall, while none of the models achieved a strong predictive capability, the Random Forest model demonstrated the best balance of performance metrics, making it the most promising candidate for further refinement and tuning.

4. Conclusion

This report presents an analysis of London crime data, focusing on understanding crime trends and predicting crime rates across various boroughs and categories. By employing a combination of data cleaning, exploratory data analysis, and machine learning modeling, the project sought to uncover insights into the factors influencing crime and assess the efficacy of different predictive techniques.

Summary of Findings

Through the analysis, it was determined that certain boroughs and crime categories exhibited distinct patterns and trends. The exploratory data analysis revealed significant disparities in crime rates, as well as temporal trends that could be essential for law enforcement and public policy decisions. The modeling efforts demonstrated that both the XGBoost and Random Forest algorithms provided superior predictive performance compared to linear regression, highlighting their capability to capture complex relationships within the data.

Potential Impact

The insights generated from this analysis could have practical implications for crime prevention strategies, resource allocation, and community safety initiatives. Law enforcement

agencies and policymakers can leverage these findings to identify high-risk areas, optimize patrol routes, and implement targeted interventions that address specific types of crime.

Limitations

Despite the strengths of this analysis, several limitations should be acknowledged:

- **Data Constraints**: The analysis was conducted on a subset of the dataset (1 million records), which may not fully capture the variability and trends present in the entire dataset.
- **Predictor Selection**: The models were based on the selected predictors, which may not encompass all relevant factors influencing crime rates, such as socio-economic variables, seasonal effects, or community programs.
- **Temporal Changes**: The nature of crime may evolve over time, and models trained on historical data may not adequately predict future trends if there are significant societal changes.

Future Work

Future research could enhance the findings of this report in several ways:

- **Incorporating Additional Data**: Including socio-economic indicators, weather data, or community engagement metrics could provide a more comprehensive understanding of crime dynamics.
- **Model Improvement**: Experimenting with more advanced modeling techniques, such as neural networks or time-series forecasting models, could yield improved predictions.
- **Longitudinal Analysis**: Conducting a longitudinal study to assess how crime rates change over time, alongside policy implementations or socio-economic shifts, could provide deeper insights into causal relationships.