Project Group 7
Seattle Crime Analysis

Submitted by:

Rohini Kolli [rohini.k@northeast.edu]

Likitha Jagithyala [jagithyala.l@northeastern.edu]

Abstract

Crime prediction and prevention have become critical components in urban safety management. With Seattle being one of the leading urban places attracting newer residents due to its IT developments, It becomes quite difficult to understand the safety around this city. This project focuses on applying machine learning techniques to model and forecast crime analysis in Seattle, Washington. Leveraging a rich dataset spanning from 2008, we explore various predictive models to understand the underlying patterns and dynamics of criminal activities. The aim is to develop accurate and reliable predictions that can assist residents to make better choices for their future homes.

Through the utilization of advanced machine learning algorithms, this project endeavors to provide insights into the temporal and spatial aspects of crime in Seattle. The predictive models are designed to capture complex relationships within the data, enabling the generation of forecasts for future crime occurrences. The results and findings of this study contribute not only to the field of machine learning but also hold practical implications for upcoming city residents seeking data-driven approaches to choose better.

Submitted by:

Rohini Kolli [rohini.k@northeast.edu]

Likitha Jagithyala [jagithyala.l@northeastern.edu]

Table of Contents

1. INTRODUCTION

Residents of urban areas around the world face the constant fear of criminal activities. Seattle, Washington, as a vibrant and dynamic city, is no exception. The rise of technology and the

increasing availability of comprehensive datasets have opened new avenues for understanding and combating crime. In this context, machine learning emerges as a powerful tool capable of uncovering patterns, trends, and anomalies within complex datasets, offering valuable insights for crime prediction and analysis.

The objective of this project is to harness the potential of machine learning to analyze and model crime data in Seattle. The dataset, encompassing a range of crime types, geographical locations, and temporal dimensions,combined with census datasets provides a robust foundation for building predictive models. By exploring various machine learning algorithms, including linear regression, random forests, support vector regression, and gradient boosting, we aim to identify the most effective model suitable for our project .

The significance of this study lies in its potential to enhance the resources available for new upcoming residents in Seattle. By developing accurate predictive models, we strive to contribute to proactive crime prevention strategies, resource allocation, and policy decisions. Additionally, the project seeks to advance the broader understanding of the interplay between socio-economic factors like the median income, populations, age etc, along with criminal activities in an urban setting.

In the subsequent sections of this report, we delve into the methodology employed for data preprocessing, feature engineering, model training, and evaluation. We present the results obtained from each model, discussing their strengths, limitations, and implications. The findings from this project aim to inform evidence-based decision-making, fostering a safer and more secure environment for the residents of Seattle.

2. DATA

2.1. Data Source

The chosen dataset is a combination of an open source data created and published by the Seattle Police Department (SPD) which is titled "SPD Crime Data 2008-Present” along with an open source data created and published by U.S. Census Bureau data title as “A Census Tract (2010) Profile ACS 5-year Estimates 2006-2010”.
Below is the APA citation of the dataset:

'SPD Crime Data: 2008-Present | City of Seattle Open Data portal. (2023, October). https://data.seattle.gov/Public-Safety/SPD-Crime-Data-2008-Present/tazs-3rd5'.

'A Census Tract (2010) Profile ACS 5-year Estimates 2006-2010 | City of Seattle Open Data portal. (2023, June 24). https://data.seattle.gov/dataset/A-Census-Tract-2010-Profile-ACS-5-year-Estimates-2/nqi3-zrh6'

2.2. Data Description

SPD Data set : The data consists of 1085079 instances i.e., rows and 17 attributes/variables i.e., columns which include crime related data from 2008 to Oct 2023. Each column description is provided in the link below:

'SPD DSG Offense MetaData: | City of Seattle Open Data portal. (2023, October). https://data.seattle.gov/api/views/tazs-3rd5/files/c1eb764d-95e4-4557-a60b-f0fda65d6d59?download=true&filename=SPD_DSG_OFFENSE_METADATA.pdf '

Census Data set : The data consists of 397 instances i.e., rows and 126 attributes/variables i.e., columns which include annual census data related to population , age, income ,rent etc. Each column description is available in the below link:

'Seattle City Gis.maps.arcgis.com. (n.d.). https://seattlecitygis.maps.arcgis.com/home/item.html?id=9436b7f938204deab664ea2f59e29b17&view=list&sortOrder=desc&sortField=defaultFSOrder#data'

2.3. Data Statistics

Seattle crime data information has 1085079 rows and 17 columns as shown in the *Figure 1* with few missing values to be treated.The dataset needs certain preprocessing before the modeling Census data has 397 rows with huge columns count at 126 with no missing values. This dataset has a lot of details related to population , income, age , rent ,etc. of the household who lived or are currently living in the city as shown in *Figure 2*. This dataset required preprocessing which helps in combining it to census data for better relation building between the two.

```
seattle_crime.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1085079 entries, 0 to 1085078
Data columns (total 17 columns):
 #   Column                  Non-Null Count    Dtype  
---  ------                  --------------    -----  
 0   Report Number           1085079 non-null  object 
 1   Offense ID              1085079 non-null  int64  
 2   Offense Start DateTime  1083570 non-null  object 
 3   Offense End DateTime    617085 non-null   object 
 4   Report DateTime         1085079 non-null  object 
 5   Group A B               1085079 non-null  object 
 6   Crime Against Category  1085079 non-null  object 
 7   Offense Parent Group    1085079 non-null  object 
 8   Offense                 1085079 non-null  object 
 9   Offense Code            1085079 non-null  object 
 10  Precinct                1085061 non-null  object 
 11  Sector                  1085063 non-null  object 
 12  Beat                    1085063 non-null  object 
 13  MCPP                    1085065 non-null  object 
 14  100 Block Address       1037535 non-null  object 
 15  Longitude               1085079 non-null  float64
 16  Latitude                1085079 non-null  float64
dtypes: float64(2), int64(1), object(14)
memory usage: 140.7+ MB
```

Figure 1: seattle crime dataset information

```
census_shp.columns.values
```

```
array(['OBJECTID', 'GEOID', 'NAME', 'ACRES_LAND', 'ACRES_WATE',
       'JURISDICTI', 'CRA_NO', 'CRA_GRP', 'GEN_ALIAS', 'DETL_NAMES',
       'TRACT_LABE', 'TOTAL_POPU', 'Age_Under_', 'PCT_POP_UN',
       'Age_65_and', 'PCT_POP_65', 'MEDIAN_AGE', 'NOT_HISPAN',
       'PCT_NOTHIS', 'NOTHISPLAT', 'PCT_NOTH_1', 'NOTHISPL_1',
       'PCT_NOTH_2', 'NOTHISPL_2', 'PCT_NOTH_3', 'NOTHISPL_3',
       'PCT_NOTH_4', 'NOTHISPL_4', 'PCT_NOTH_5', 'NOTHISPL_5',
       'PCT_NOTH_6', 'HISPANIC_O', 'PCT_HISP_A', 'PERSON_OF_',
       'PCT_PERSON', 'HOUSEHOLDS', 'POPULATION', 'PCT_POP_IN',
       'AVERAGE_HO', 'POPULATI_1', 'PCT_POP__1', 'FAMILY_HOU',
       'PCT_FAM_HH', 'Populati_2', 'PCT_POP__2', 'AVERAGE_FA',
       'NONFAMILY_', 'PCT_NON_FA', 'TOTAL_HOUS', 'OCCUPIED_H',
       'PCT_OCC_HU', 'VACANT_HOU', 'PCT_VACANT', 'OWNER_OCCU',
       'PCT_OWN_OC', 'RENTER_OCC', 'PCT_RENT_O', 'Populati_3',
       'High_schoo', 'Bachelor_d', 'PCNT_HIGHS', 'PCT_BACHEL',
       'POPULATI_4', 'SPEAK_LANG', 'PCT_SPEAK_', 'SPEAK_ENGL',
       'PCT_SP_ENG', 'HU_VALUE_L', 'HU_VALUE_5', 'HU_VALUE_1',
       'HU_VALUE_2', 'HU_VALUE_3', 'HU_VALUE_4', 'HU_VALUE_6',
       'HU_VALUE_7', 'HU_VALUE_M', 'WITH_MORTG', 'WITH_MOR_1',
       'WITHMORTGA', 'WITHMORT_1', 'WITHMORT_2', 'WITHMORT_3',
       'PERC_SMOCA', 'WITHMORT_4', 'MEDIAN_GRO', 'RENT_GRAPI',
       'GRAPI_LESS', 'GRAPI_15_0', 'GRAPI_20_0', 'GRAPI_25_0',
       'GRAPI_30_0', 'GRAPI_35_0', 'PERCENT_GR', 'GRAPI_NOT_',
       'Populati_5', 'CIVILIAN_L', 'CIVILIAN_1', 'CIVILIAN_2',
       'PERCENT_UN', 'NOT_IN_LAB', 'MEDIAN_HH_', 'HH_INCOME_',
       'HH_INCOM_1', 'HH_INCOM_2', 'HH_INCOM_3', 'HH_INCOM_4',
       'HH_INCOM_5', 'HH_INCOM_6', 'HH_INCOM_7', 'HH_INCOM_8',
       'HH_INCOM_9', 'POP_POVERT', 'all_people', 'PCT_POPULA',
       'All_Famili', 'All_Fami_1', 'PCT_ALL_FA', 'ACS_VINTAG',
       'Populati_6', 'Populati_7', 'POP_INCPOV', 'POP_INCP_1',
       'PCT_POP_BE', 'SHAPE_Leng', 'SHAPE_Area', 'geometry'], dtype=object)
```

Figure 2: Columns of Census Dataset

2.4. Data Preprocessing

2.4.1. Treating Missing Values

The Seattle crime data by SPD has few missing values that are required to be treated. The below Figure 3 gives us the total missing values of the columns.

```
seattle_crime.isnull().sum()

Report Number                  0
Offense ID                     0
Offense Start DateTime      1509
Offense End DateTime      467994
Report DateTime                0
Group A B                      0
Crime Against Category         0
Offense Parent Group           0
Offense                        0
Offense Code                   0
Precinct                      18
Sector                        16
Beat                          16
MCPP                          14
100 Block Address          47544
Longitude                      0
Latitude                       0
dtype: int64
```

Figure 3: Missing values for each column

Offense End DateTime can be removed as we have Offense Start Time/Report DateTime to compensate for it. However, the Offense Start Time is more accurate as per our analysis requirement. Similarly 100 Block Address is the precise address of the criminal event occurrence. This column too can be ignored as the dataset has MCPP along with latitude and longitude information which can easily be used instead of 100 Block Address. Final treatment of missing values is to remove the rows which have even 1 missing value. As the remaining missing rows post removal of the Offense End DateTime and 100 Block Address is negligible we can go ahead with this process

2.4.2. Spatial Joining Census data

SPD data alone gives us most of the information related to crime across years. However, few additional census information gives us the basis of measuring the impact/causes of these crimes for a set of population.

There is no common column between two dataset for a join apart from geocodes. Therefore with the help of geopandas we performed a spatial join on both the dataset based on the geometry (geocodes).

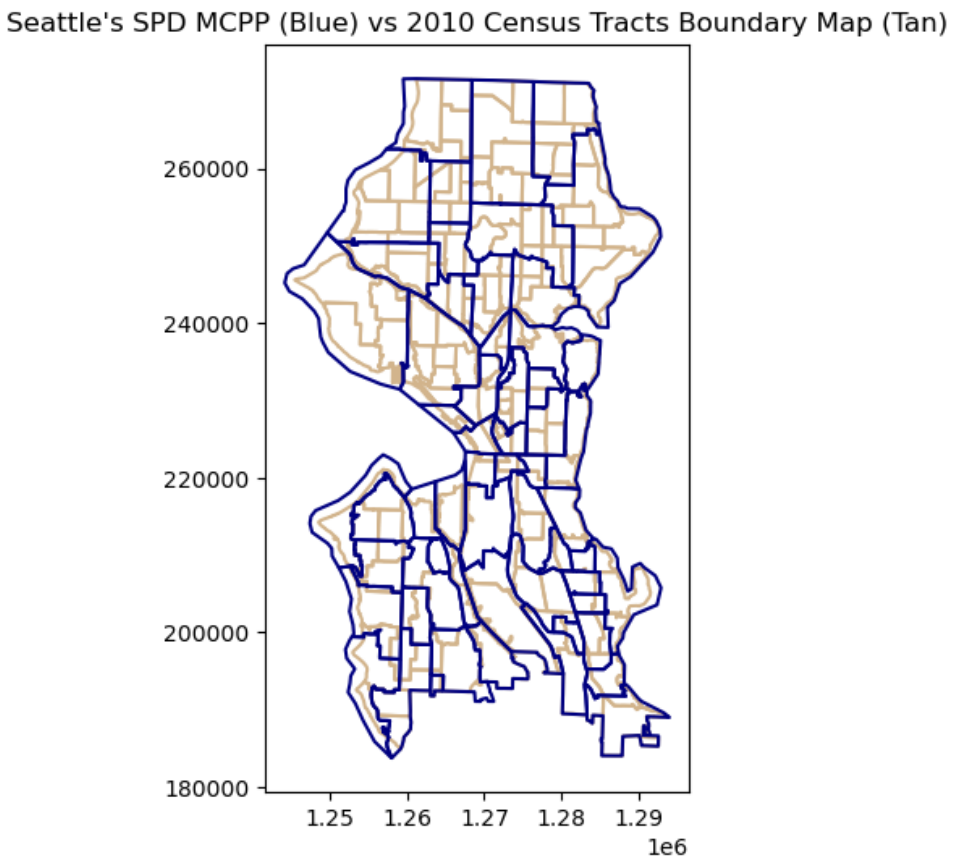

Figure : Seattle City Polygon as per SPD and census data.

2.5. Exploratory Data Analysis

2.5.1. Trend of Crime for different time ranges

Crime occurrences from time to time provide us much information on the crime pattern for the city. Thus below figures tell us how the crime has been trending for different time ranges like year, season , months and time of the day.

From *Figure 4* we can see that the year in which Seattle experienced most crimes is 2020 followed by 2018 and 2022. The increase in crimes for years from 2020 to 2022 is most likely due to the pandemic and unstable economics as a consequence of the coronavirus outbreak. The least amount of crime that the city has seen is in 2012 and 2011 respectively. These are also the same years where the economy has been stabilizing and recovering from the 2008 depression.

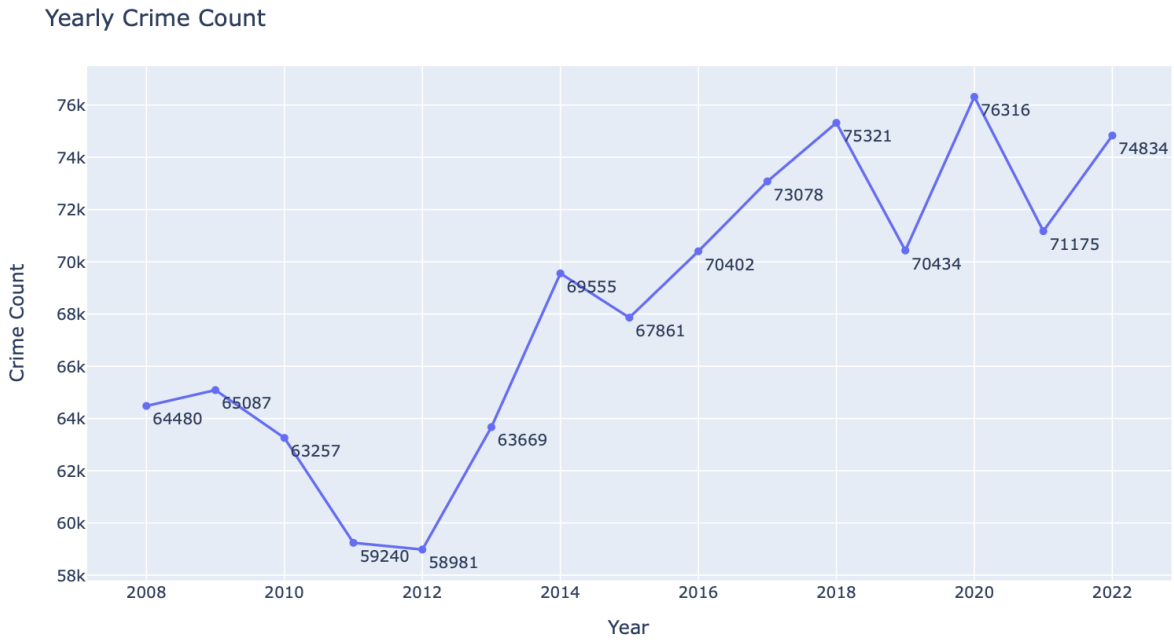

Figure 4: Yearly crime count from 2008 to 2023

For the Seasonal crime trend we see that the crime count in summer is slightly higher than other seasons. Fall and Spring have similar crime trends whereas the city has comparatively less crimes in the winter season. Seattle as a city experiences continuous bad weather during winter and this is one of the main reasons for comparitively less crime as seen in *Figure 5*.Monthly crime trends can be seen in *Figure 6* where in May month has the most crimes that the city has seen followed by October. These months are the upper and lower bounds for the summer season.Crime usually are planned as per the climatic conditions of the city. February month has shown much lesser crimes than others

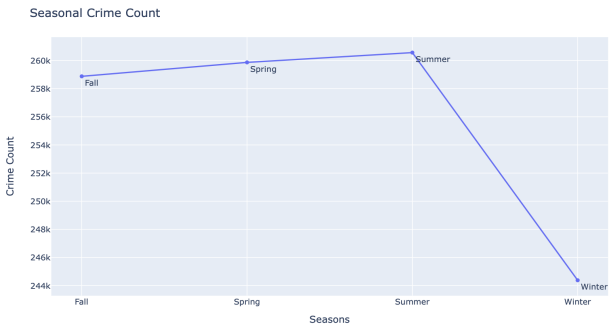

Figure 5: Seasonal Crime count for 4 seasons

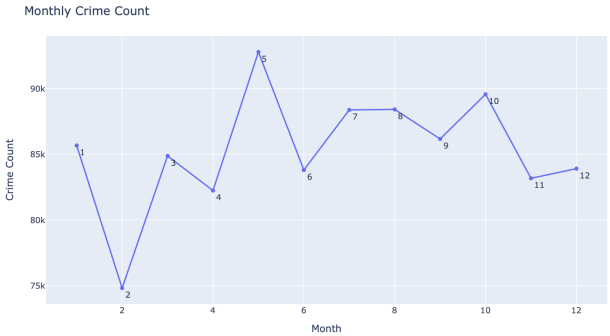

Figure 6: Monthly crime count for 12 months

The below analysis is one of the most insightful analyses. Midnight 12 is the most popular hour for crime. The city experiences the most crimes during night as seen in *Figure 7*. Post 4pm the crime happening instances have an increasing trend. From 1am, the crime count drops up to 5am and slightly increases later.

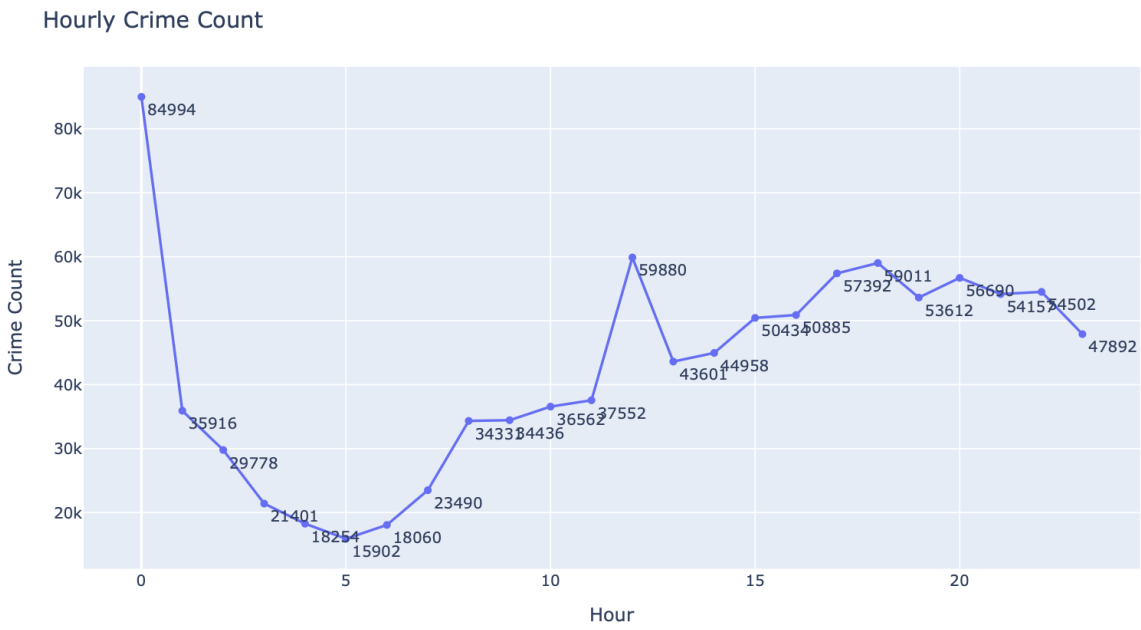

Figure 7: Crime count for different hours of the day

Based on the time range plots the new residents are expected to be extra cautious during the high crime time period.

2.5.2. Safe and Risky Localities

Seattle as a city has many residential neighborhoods/ localities and crime for these localities also varies. As an upcoming resident, one would definitely be looking at these localities while considering their new homes. As per Figure 8 the most risky localities with highest crimes observed in Seattle are Downtown Commercial, Capitol Hill, NorthGate whereas as per Figure 9 the safest localities with least crimes are Commercial Harbour Island, Commercial Duwamish, Pigeon Point.

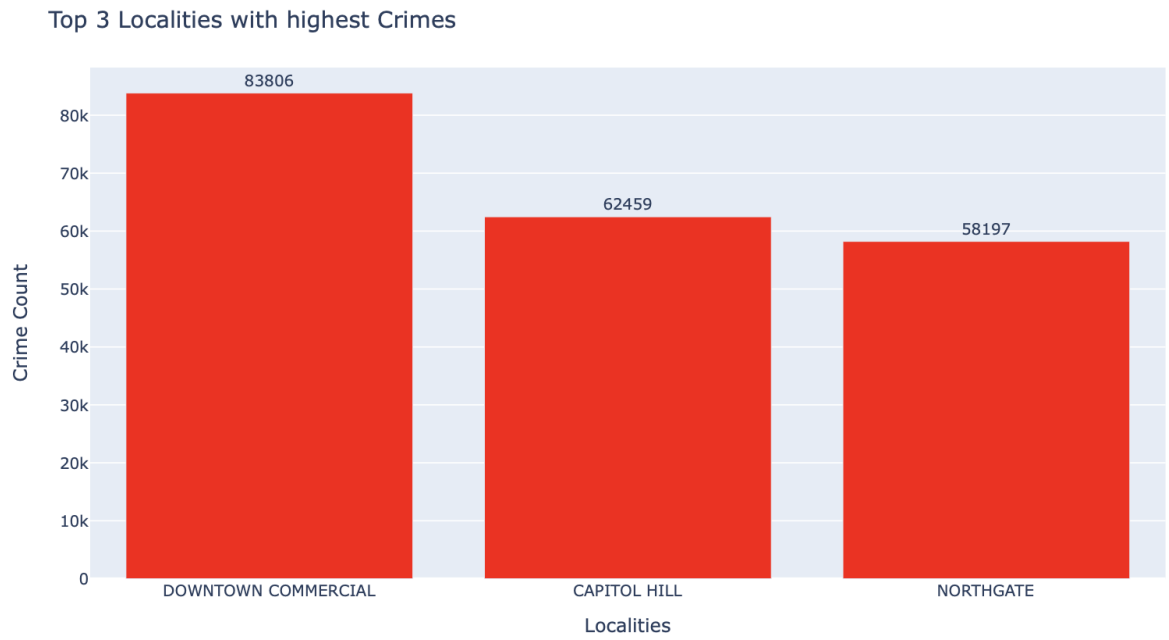

Figure 8: 3 Most Riskiest Localities

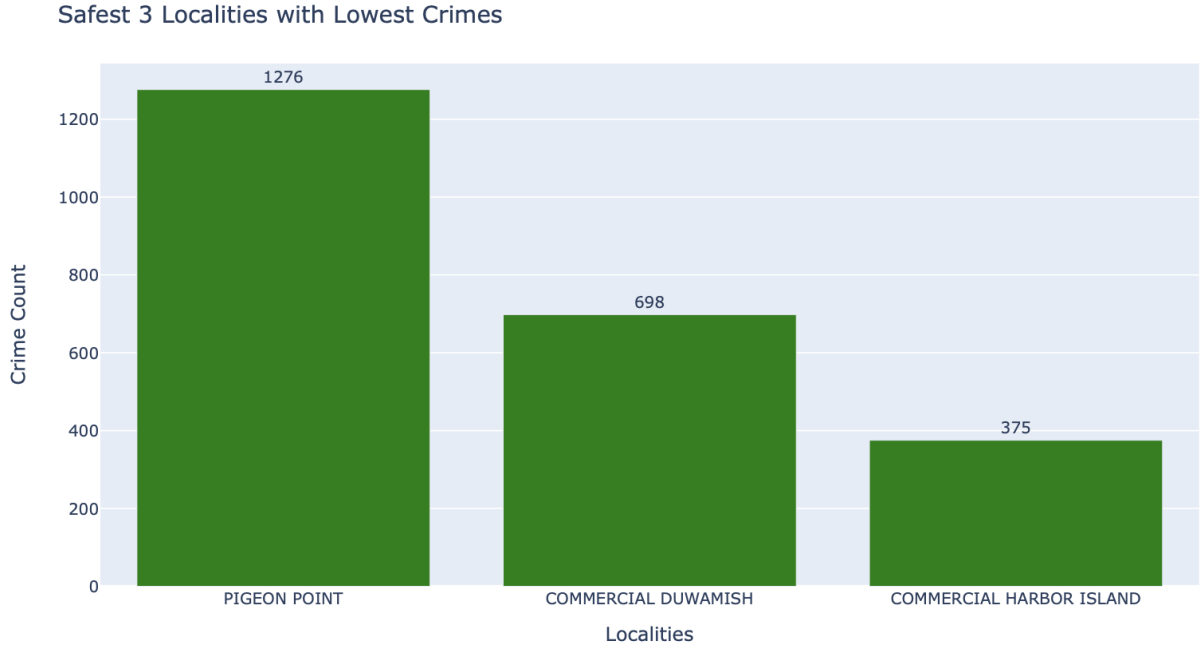

Figure 9: 3 Most Safest Localities

2.5.3. Proportion of Offense Parent Group

For the recent year of 2023

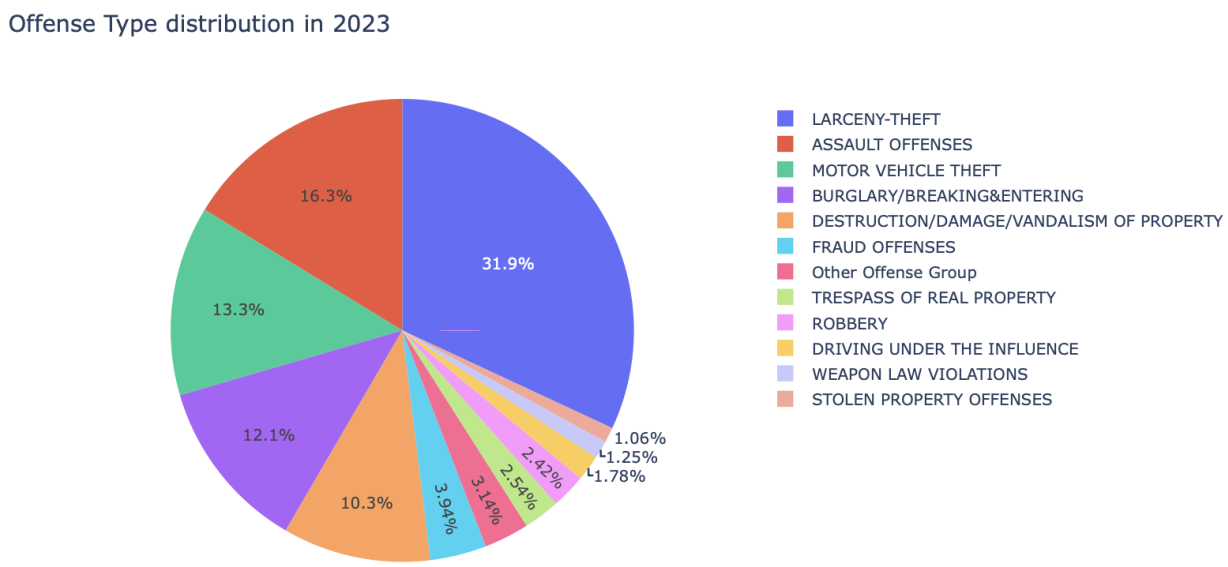

Figure 10: Pie Chart showing % of Offense Parent Group distribution for 2023

2.6. Feature Engineering

Census data helps predict the crime rate per MCPP which includes Population, land area (acres), median age, % bachelor, % of occupied housing, median gross rent, % unemployment, median household income, % poverty. There are close to 130 features and the dataset needs feature engineering before we proceed to modeling to avoid noise and simplify the model complexity.

3. MODELING

There are three goals of this project which include various modeling techniques wherein we are predicting crime rate , estimating the likelihood of top crime rate and forecasting the crime for future years.To measure/evaluate the model performance, the below performance metrics are used.

I. Regression :

- *RSquared [R2]*. It is the coefficient of determination that tells how well the model fits the data and also the goodness of the fit. It is calculated as Residual sum of squares(RSS) by Total sum of squares(TSS) and then subtracted by 1.

$$R2 = 1 – (RSS/TSS)$$

- *Root Mean Squared Error.* It is the average length from each actual to its predicted value. It is calculated as sum of total square of predicted value at i(Pi) subtracted by actual value at i(Ai) , total divided by 2 and square root of remaining is considered

$$RMSE = \sqrt{\Sigma(Pi – Ai)2 / n}$$

Likelihood:

Precision:

Definition: Precision is the fraction of relevant instances among the retrieved instances. It measures the accuracy of the positive predictions.

Formula: True Positives (TP)/ True Positives (TP)+False Positives(FP)

Recall:

Definition: Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. It focuses on the coverage of the actual positive cases.

Formula: True Positives (TP)/ True Positives (TP)+False Negatives(FP)

Other metrics like MAE(Mean Absolute Error) , computational time etc are also considered for model evaluation.

3.1 Crime Rate Predictive Analysis

This analysis includes predictive modeling to generate crime rate based on combined data from SPD and census. Various regression models are trained to evaluate and use the best of all.
Crime rate is not provided beforehand instead we calculated it as follows:

$$Crime\ Rate = \frac{Offense\ Count\ per\ area}{Population\ per\ area} \times 1{,}000\ people\ [=] \frac{\#\ crime/MCPP}{1{,}000\ people}$$

Figure : Crime Rate formula

Crime rate is majorly influenced by the crime count and the population of a specific MCPP/locality.
Additional Data Preprocessing includes

- Aggregation of crime and other census values.

- Treat Data Imbalance
- Categorical encoding
- Split Train and Test data

3.1.1. Model Training

Linear Regression

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, meaning that changes in the independent variables are associated with a constant change in the dependent variable.

The basic form of a simple linear regression equation with one independent variable is: $y=mx+b$ where y is the dependent variable, x is the independent variable, m is the slope of the line,b is the y-intercept.
For this modeling we considered data from 2010 to 2019 to avoid anomalies with respect to pandemic etc.The model is trained with y (target variable) defined as crime rate and x(features) defined as all columns apart from crime rate. Below are important features with their importance values in the linear regression model.

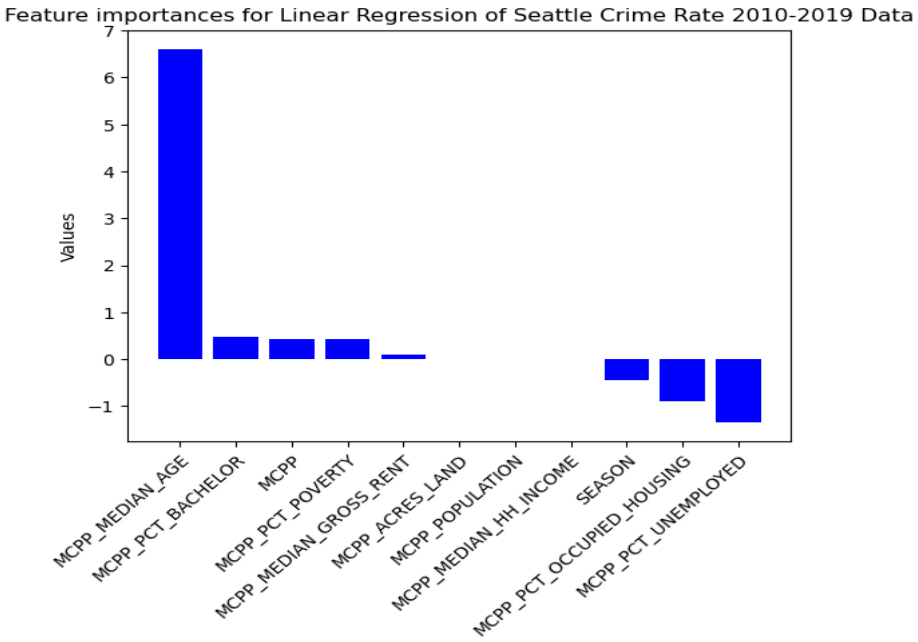

Figure : Feature Importance of Linear regression model

Lasso Regression

Lasso Regression, or L1 regularization, is a linear regression technique that adds a penalty term based on the absolute values of the coefficients. This regularization technique is useful for feature selection and can lead to sparse models where some coefficients are exactly zero. The term "Lasso" stands for Least Absolute Shrinkage and Selection Operator.
The objective function of Lasso Regression is given by:

$$\text{minimize} \left(\frac{1}{2n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{p} |w_j| \right)$$

Figure : Lasso regression's objective function

Where *n* is the number of observations,*y of i* is the actual output for the i-th observation,*y-hat of i* is the predicted output for the i-th observation,*p* is the number of features (independent variables), *w of j* is the coefficient for the j-th feature,α is the regularization strength parameter.
Below are important features with their importance values in the lasso regression model.

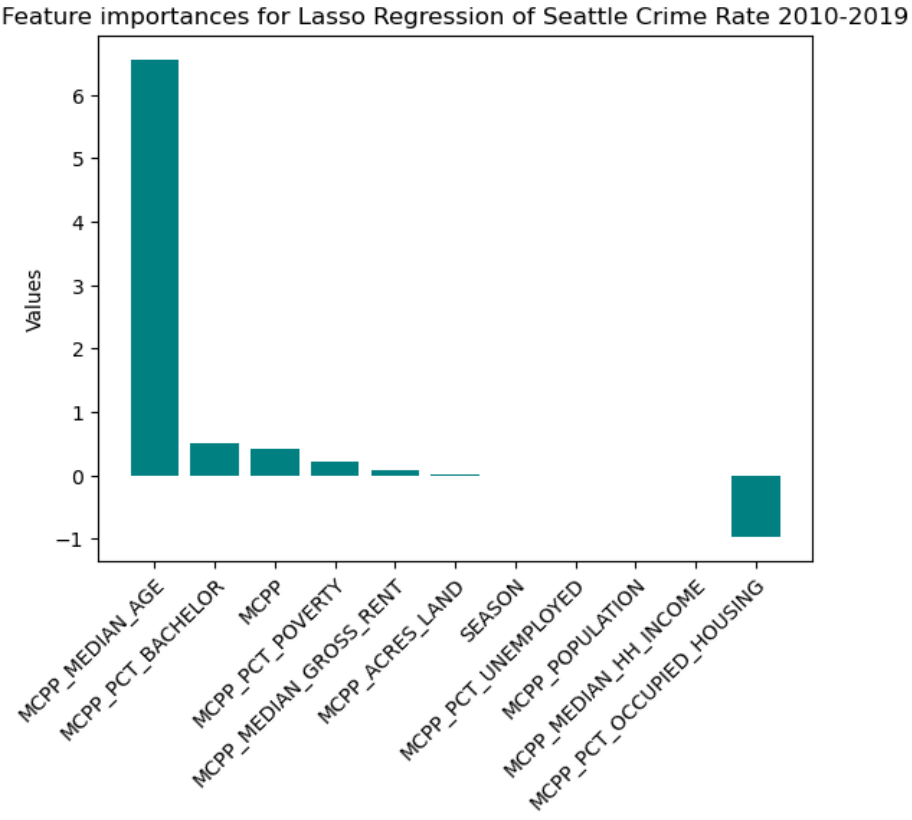

Figure : Feature Importance of Lasso regression model

Random Forest Regression:

Random Forest Regression is an ensemble learning method that uses multiple decision trees to make predictions. It belongs to the broader class of ensemble methods, which combine the

predictions of several weak learners (in this case, decision trees) to create a stronger, more robust model. Random Forests can be used for both classification and regression tasks, but here we'll focus on Random Forest Regression.
For Random Forest Regression, the prediction for a new data point is the average (or sometimes the median) of the predictions of all the individual trees. Mathematically, if {*y-hat of 1, y-hat of 2 ...., y-hat of n}* are the predictions of the individual trees, the final prediction *y-hat of final* is given by:

$$\hat{y}_{\text{final}} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i$$

Figure : Random forest predictive formula

Below are important features with their importance values in the random forest regression model.

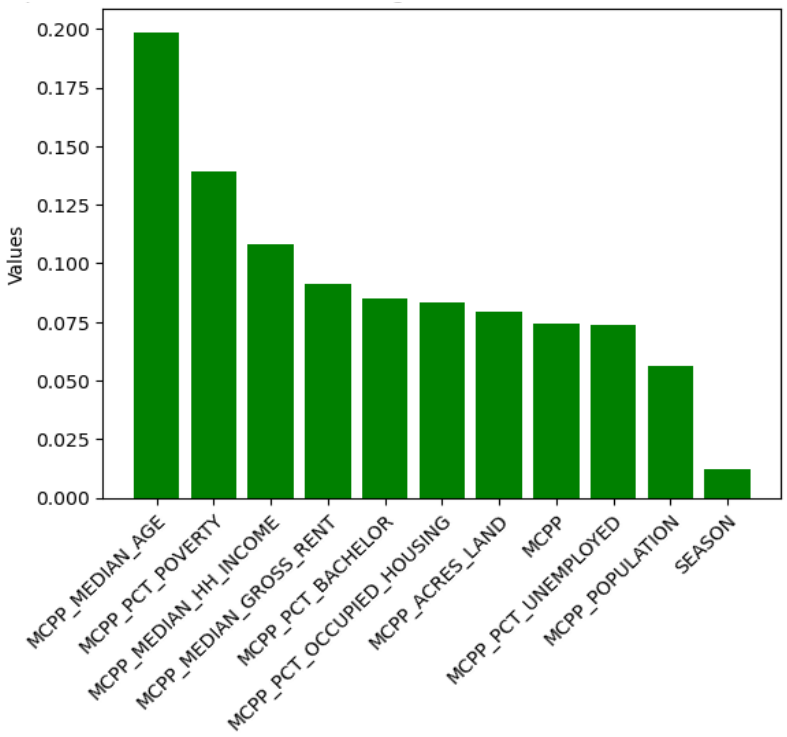

Figure : Feature Importance of Random Forest regression model

K-NN Regression:

K-Nearest Neighbors (KNN) regression is a non-parametric and supervised machine learning algorithm used for predicting the continuous value of a target variable. In KNN regression, the predicted value for a new data point is determined by the average (or weighted average) of the target values of its k nearest neighbors in the feature space.We have considered multiple k values and evaluated the model accordingly

Gradient Boost Regression:

Gradient Boosting Regression is another ensemble learning technique used for both classification and regression tasks. Like Random Forests, Gradient Boosting builds a strong predictive model by combining the predictions of multiple weak learners. However, the key difference lies in how these weak learners are trained and combined. For regression tasks, the predictions are made by summing up the predictions of all the trees in the ensemble.
Below are important features with their importance values in the gradient boost regression model.

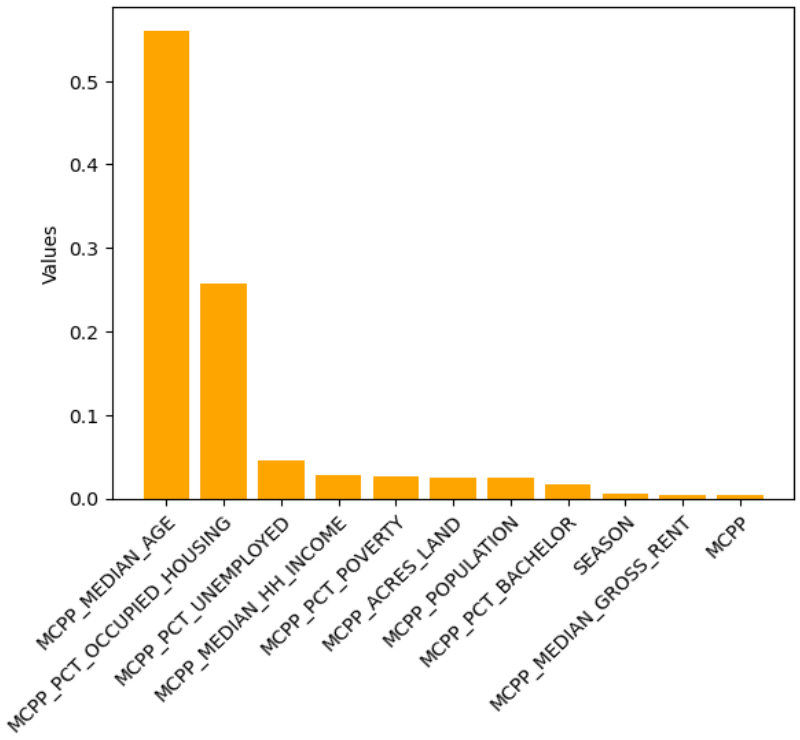

Figure : Feature Importance of Gradient Boost regression model

3.1.2. Model Evaluation

The table below summarizes the evaluation metrics of the regression through multiple modes. The table includes various tested models along with their RMSE, squared R for both train and test set and computation time in seconds.

Performance metrics for Crime Rate Regression on Seattle Police Department 2010-2019 dataset

	Regression Model	RMSE (train)	RMSE (test)	R^2 (train)	R^2 (test)	computation time (sec)
0	Linear Regression	26.4161	28.2948	0.6361	0.6199	0.0097
1	Lasso Regression	26.4414	28.3675	0.6354	0.618	0.006
2	Random Forest - RF (original)	9.6519	12.0119	0.9514	0.9315	0.3784
3	Random Forest - RF (best)	9.6781	11.8796	0.9512	0.933	0.047
4	Gradient Boost - GB (original)	9.897	11.6278	0.9489	0.9358	0.1438
5	Gradient Boost - GB (best)	9.8113	11.6593	0.9498	0.9355	0.1567
6	K-Nearest Neighbor -KNN (k=5)	10.0738	12.5542	0.9471	0.9252	0.0132
7	K-Nearest Neighbor -KNN (k=6, best)	9.7411	12.0372	0.9505	0.9312	0.0124

Figure : Model Evaluation using multiple regression metrics

For a regression model to be running successfully, the ideal situation is to have higher squared R , lower RMSE and lower computation time. From the model evaluation summary Random forest , Gradient Boost and K-NN have higher squared R for both train and test set .However Random forest has lower RMSE too. Computational cost is subjective to the user. If the computation cost of 0.47 is high then another alternative that one can choose for similar performance is K-NN model

Major Challenges in this modeling is to accommodate the important features as there are close to ~130 features for crime rate. Feature engineering and importance were quite useful mechanisms used in this modeling

3.2 Crime Count Forecast Analysis

By leveraging historical sales data, we aimed to enhance the accuracy of crime count predictions to the required locality for future years. This analysis includes specific columns and crime count aggregated accordingly.Additional Data Preprocessing includes only aggregation of data as per year and MCPP. X is defined as years and MCPP whereas y is crime count.

3.2.1. Model Training

Linear Regression[Ridge , Lasso , Elastic Net]

Ref: 3.1.1 The logic behind the model remains the same and the prediction of crime count for one of the MCPP [Alaska Junction] is as below. Other models such as ridge , lasso , elastic net regression have been trained and they showed results similar to a basic linear regression.

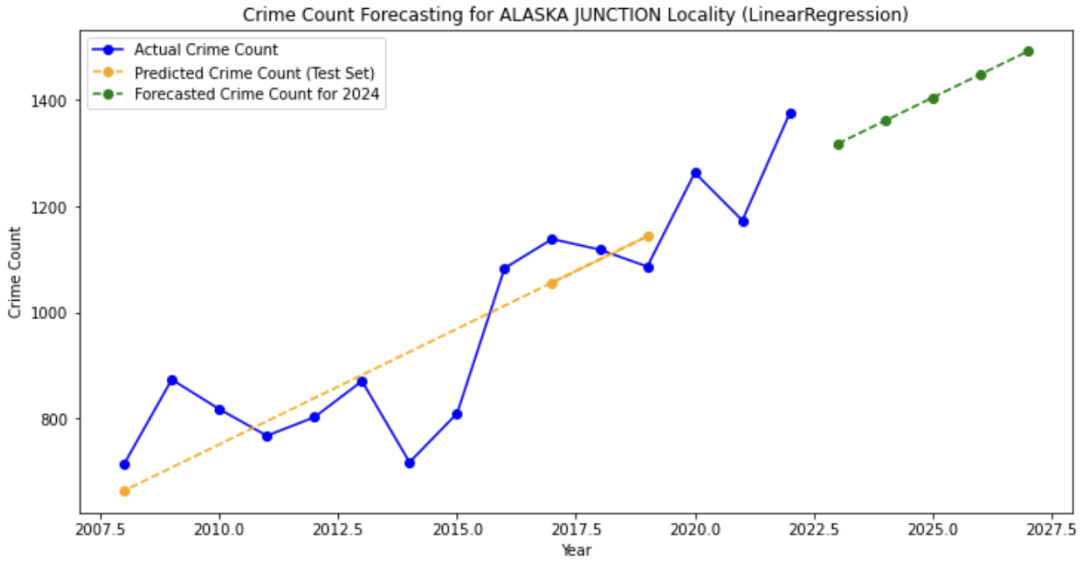

Figure : Linear regression model's Crime count trend and forecast for future year in Alaska Junction

Other Regression:

Other models such as : Random forest , Support Vector ,Gradient Boost, K-NN and XGBoost were used to train the dataset. However ,these models did not perform that well!.Few models have experienced extremely underfit and overfit issues. As a part of Future work,to overcome these issues we could include many detailed factors to the dataset or forecast for short intervals.

3.2.2. Model Evaluation

Model evaluation is done basis the RMSE and squared R metrics.Summary of the model performance for all the models tried is as below:

```
Modeling on Crime count forecast for future year and their performance evaluation
```

Model Name	RMSE(Train)	RMSE(Test)	R^2(Train)	R^2(Test)
Linear Regression	103.18	64.6783	0.7583	0.8826
Random Forest	38.248	95.9997	0.9668	0.7415
Support Vector	229.52	214.348	-0.1957	-0.2887
Gradient Boosting	0.2927	98.7632	0.9999	0.7264
Lasso Regression	103.18	64.6659	0.7583	0.8827
Ridge Regression	103.18	64.6333	0.7583	0.8828
Elastic Net	103.18	64.6333	0.7583	0.8836
K-NN	89.2269	123.416	0.8192	0.5727
XGBoost	0.0011	98.8769	0.9999	0.7257

Figure : Crime count forecast model summary

Linear regression models out of others have a decent performance, The major challenge during the modeling is that the time interval is huge due to which the model does not have sufficient patterns to predict accurately. A monthly prediction model might have better performance.Usage of transformers or Neural network could also be beneficial

3.3 Predictive/Likelihood Analysis of Top Crime in MCPP

This report presents a detailed analysis of the Seattle Police Department's crime data from 2008 to present, with a focus on predicting the likelihood of various crime types within Major Crime Prevention Program (MCPP) areas.

3.3.1. Classification Task and Model Selection

The primary task was to classify crimes into their respective types based on MCPP zones. We selected RandomForest, Logistic Regression, Gradient Boosting, and Decision Tree models for their varying strengths in handling complex datasets and providing accurate predictions.
Methodology:
The methodology involved splitting the data into training and testing sets, feature engineering to extract meaningful patterns, and applying cross-validation to assess model performance.

3.3.2. Model Evaluation and Metrics

Model performance was evaluated using precision, recall, and F1-scores, both before and after cross-validation. These metrics were chosen for their ability to provide a nuanced view of the models' predictive capabilities, especially in the context of imbalanced classification tasks.
Results and Discussion:
Our initial results indicated similar performance across models, with challenges in accurately predicting minority classes. After cross-validation, a slight variation in performance metrics was observed, yet they remained low, indicating the complex nature of crime prediction. The RandomForest and Decision Tree models showed identical results, with Logistic Regression and Gradient Boosting performing similarly.
Challenges and Future Directions:

The main challenges included the complexity of the data and imbalance in crime occurrences. Future work will explore more advanced modeling techniques, hyperparameter tuning, and the integration of additional contextual data, such as socio-economic and demographic factors.

Conclusion:

This report highlights the potential and challenges in using machine learning for crime type prediction in MCPP zones. It emphasizes the need for continuous model refinement and the

integration of comprehensive data to enhance the accuracy and applicability of predictive insights in public safety planning.

In this analysis, we faced the inherent complexity of predicting diverse crime types, a task marked by high variability and influenced by numerous unpredictable factors. Our chosen models, while robust and versatile, have limitations in handling such intricate, imbalanced datasets, which is reflected in the accuracy metrics obtained. However, these metrics still provide valuable insights, especially when considering the significance of the 'Top Crime' feature. This feature, central to our analysis, plays a crucial role in understanding crime trends and patterns within MCPP zones. It allows us to focus on the most prevalent crime types, offering a targeted approach to crime prediction and prevention strategies. The moderate accuracy levels, therefore, should be viewed in the context of the task's complexity and the strategic value of the predictive insights gained.

4. CONCLUSION AND FUTURE WORK

In this report, we have performed predictive analysis to predict crime rate, forecaste crime count for future years and identify the likelihood of top crime happening at a locality.
The analysis involved preprocessing the data, splitting it into training and testing sets, training different regression models, and evaluating their performance on both historical and test data.

Predictive Crime Rate Analysis required intensive feature engineering considering the ~130 features .Major challenge here was the spatial join of census data which was successful with the help of geopandas. As a part of future work we would like to explore more external factors affecting the crime.

Time series crime analysis reports have leveraged machine learning techniques to gain valuable insights into the dynamics of crime. By employing a variety of regression models, we aimed to forecast future crime counts and understand temporal trends in criminal activity. In future we would like to explore transformers or neural networks and also conduct modeling for short time intervals

The report highlights the complexities and opportunities of using machine learning for crime prediction in MCPP zones. While the variability and unpredictability of crime data present challenges, the analysis of the 'Top Crime' feature offers critical insights into prevailing crime patterns. Although the accuracy of these models is moderate, it is noteworthy given the intricate nature of crime data. These insights are vital in shaping effective crime prevention and public safety strategies, demonstrating the utility of machine learning in this field.

5. REFERENCE

Crime Dashboard - Police | Seattle.gov. (n.d.). https://www.seattle.gov/police/information-and-data/data/crime-dashboard

Saeed, R. M., & Abdulmohsin, H. A. (2023). A study on predicting crime rates through machine learning and data mining using text. *Journal of Intelligent Systems*, *32*(1). https://doi.org/10.1515/jisys-2022-0223

Hannahrjiang. (n.d.). *Modeling-Crime-Data/Optional Activity: Machine Learning.ipynb at master · hannahrjiang/Modeling-Crime-Data*. GitHub. https://github.com/hannahrjiang/Modeling-Crime-Data/blob/master/Optional%20Activity%3A%20Machine%20Learning.ipynb

Koehrsen, W. (2019, December 10). Hyperparameter tuning the random forest in Python - towards data science. *Medium*. https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Tong, X., Ni, P., Li, Q., Yuan, Q., Liu, J., Lu, H., & Li, G. (2021). Urban Crime Trends Analysis and Occurrence Possibility Prediction based on Light Gradient Boosting Machine. *IEEE*. https://doi.org/10.1109/bdai52447.2021.9515252

Mahmud, S., Nuha, M., & Sattar, A. (2020). Crime rate prediction using machine learning and data mining. In *Advances in intelligent systems and computing* (pp. 59–69). https://doi.org/10.1007/978-981-15-7394-1_5