IEOR 142 Final Project Report: Crime Prediction in Berkeley

Motivation

As Berkeley students, we often receive emails from Berkeley Warnme that report violent crimes near campus, which makes us and our parents concerned about the public safety environment of Berkeley city. As UC Berkeley is building an Outdoor Early Warning System, we are also motivated to develop a crime prediction model for residents near Berkeley using the knowledge we learned from this course. The models aim to provide highly accurate, updated crime predictions, empowering the Berkeley community with the knowledge to take appropriate safety precautions. The anticipation is that with better foresight, law enforcement can allocate resources more effectively, thereby increasing their efficacy in crime prevention and response.

A significant motivation is the model's potential to aid in data-driven policy-making. By understanding crime trends, policymakers can address the underlying socio-economic factors contributing to crime, such as poverty, inequality, and lack of educational opportunities. This approach tackles the symptoms (crime incidents) and the root causes, leading to long-term positive effects on public safety(by doing NLP and feature engineering).

Additionally, the project is motivated by the goal of community empowerment. By making crime data accessible and understandable, residents are better equipped to make informed decisions about their safety and well-being. This empowerment fosters a proactive stance towards personal and community safety.

Moreover, the project holds educational value, providing insights into crime patterns and trends and dispelling myths and misconceptions about crime in the area. It serves as a tool for awareness and education, helping to build a more informed community.

In summary, this project is driven by the desire to create a safer, more informed, and engaged community, to use resources efficiently, and to advance the field of crime analysis and prevention through innovative technological solutions.

Data and Data Pre-processing

We obtained our datasets “Berkeley PD Cases (2016 to Current)” and “Berkeley PD-Calls for Service” from Berkeley PD’s website’s “Policy Transparency” section. The “Berkeley PD Cases (2016 to Current)”, short for “Cases,” provides detailed records of around 80000 incidents that happened in Berkeley. The “Berkeley PD-Calls for Service,” short for “Calls for Service,” is a log that records calls made to the Berkeley police department, serving as a source for understanding public-reported incidents.

The “Cases” data has ten columns: Case_Number, Occured_Datetime, Incident_Type, Statue, Statue_Description, Statue_Type, Block_Address, City, and Zip_Code, where Case_Number represents a unique ID for each incident, Occured_Datetime tells the date and time of each incident accurate to the minute, Incident_Type denotes 40 types of violations, Statue_Description provides a detailed description of the incident, and Block_Address indicate the block where the incident took place. The “Calls for Service” data has 18 columns, and only the Block_Location that informs where the incident happened has been kept.

For the “Cases,” the study first imports and standardizes the date and time formats in the Occured_Time Column to ensure a more straightforward future data analysis. The study then extracts and assigns the date and hour of the crimes from the Occured_Datetime column to the new respective 'Date' and 'Hour' columns in the dataset for more accessible data analysis. The “Cases” data is categorized based on the time of the day, where morning is from 6 AM to 11:59 AM, “noon” is from 12 PM to 5:59 PM, and “night” is from 6 PM to 5:59 AM. The study utilizes one hot encoding to convert the categorical time of day values into a numerical format represented by “if_monring,” “if_afternnon,” and “if_night.”
The “Date” column is used to make another column called “Day of the Week,” which indicates the day the incident happened, providing another layer of temporal details. The study also extracts the month when the incident happened from the Occur_Datetime columns, where months 3 to 5 are defined as spring, 6 to 8 as summer, 9 to 11 as fall, and 12 to 2 as winter. This seasonality value is stored under the

"Season" column in the "Cases" dataset, adding a seasonal perspective to the crime data. The study identified and flagged public holidays and weekends by assigning these days with values of 1 in the "if_holiday" column to acknowledge their potential impact on crime patterns.

For a more detailed spatial analysis, the study merged the "Cases" dataset with the "Calls for Service" dataset, as the latter has more specific information about where the incident took place, such as the latitude and longitude of the block. This merge of datasets significantly enriches the location data and makes future spatial analysis easier. The study then determines and analyzes crime data by quadrants in Berkeley based on information on latitude and longitude of south and west Berkeley estimations from "Latitude.to." website (southwest, southeast, northwest, northeast).

For the final step of EDA, the study employs K-Means clustering on latitude and longitude data to identify crime hotspots. The study used the elbow method to determine the optimal number of clusters and show the spatial concentration in crime. Although the elbow method indicates the optimal number of clustered to be 3, the study chose a cluster of 4 since Berkeley is often divided into Southwest, Southeast, Northwest, and Northeast sections. The cluster labels are under the column called "Cluster."

Modeling

Our project incorporates X models in total. We aimed to predict three dependent variables: safe or unsafe, crime counts, and time severity, given dependent variables like Season, Hour of Day, if_morning, if_afternoon, Day_of_the_week, and Holiday.

First, we built four different models to predict whether the dependent variable is safe (1) or unsafe (0). Specifically, we define safety based on a criterion where a location is considered safe if, on average, fewer than one crime has been reported during the given time of day over the past seven years. We used a Logistic Regression model, CART Decision Tree Classifier, Random Forest Classifier, and Gradient Boosting Classifier.

Logistic Regression Model 1

We first used the Logistic Regression Model to predict whether it is safe or unsafe. The dependent variable in this model is the safety of a location and time, categorized as safe (1) or unsafe (0). The independent variables are Season, Hour, if_morning, if_afternoon, Day_of_the_week, and Holiday (binary variable). The p-values of the features tend to be high, with Season [T.Spring] = 0.847, Season [T.Summer] = 0.044, Season [T.Winter] = 0.367, Hour = 0.429, if_morning = 0.386, if_afternoon = 0.350, Day_of_the_Week = 0.002, and Holiday = 0.690. With only Season [T.Summer] and Day_of_the_Week = 0.002 being statistically significant using a p-value threshold of 0.05, we still decided to keep the rest of the features because we think they can provide important inferential information, which is significant when we apply the model to real-world scenarios. The model's performance is characterized by an accuracy of 0.8715413751043205, a True Positive Rate (TPR) of 0.5993635077793493, a False Positive Rate (FPR) of 0.06808377127617853, and a P-threshold value of 0.55. This p-threshold is selected based on backward engineering to maximize accuracy, TPR, and minimize FPR. To evaluate the p-threshold, we also made a precision and recall graph. The graph (figure 7) shows the precision and recall tradeoff, it appears there's a point where precision begins to drop sharply while recall is still relatively high. This point represents a threshold that balances both precision and recall before precision declines. When choosing the optimal threshold from this curve, we look for a balance point where we could have good precision without sacrificing too much recall and vice versa. As a result, our model is optimized when we choose a p-threshold value between 0.4 and 0.6. From the Precision-Recall curveThis model effectively categorizes times and locations as safe or unsafe, serving as a valuable tool for community members to assess safety.

CART Decision Tree Classifier

Our dataset uses a Decision Tree Classifier to predict a specific target variable. For the part of Model Training and Hyperparameter Tuning, We employed a grid search approach with 10-fold cross-validation to tune the hyperparameters of the Decision Tree Classifier, which enables our model

searches through a specified subset of hyperparameters to find the combination that yields the best performance. We first tuned the hyperparameters to optimize the model's performance. We employed GridSearchCV, coupled with 10-fold cross-validation. This comprehensive approach evaluated 201 combinations of the hyperparameter ccp_alpha, ranging from 0 to 0.10, and three variations of max_depth. The purpose of varying ccp_alpha was to regulate the complexity of the tree, aiming to mitigate overfitting. Simultaneously, exploring different max_depth values allowed us to assess the impact of the tree's growth level on the model's performance. The selected features for the model included 'Season,' 'Hour,' 'if_morning,' 'if_afternoon,' 'Day_of_the_Week,' and 'Holiday.' The Decision Tree Classifier was chosen for its interpretability and efficiency in processing categorical data. The model's optimal configuration was identified with specific values for ccp_alpha and max_depth, striking a delicate balance between complexity and the ability to generalize. In terms of performance, the model achieved a commendable cross-validation accuracy score, indicative of its robustness and reliability on unseen data.

Furthermore, the model's accuracy on the test set was notably high, reinforcing its effectiveness in accurately classifying safety status. We initially included ‘Cluster’ in our model when choosing the features. However, the result shows that the accuracy is 1, implying that the model overfits. Then I exclude ‘Cluster’ as a feature. Decision trees create splits based on the features that most effectively segregate the data into the target classes. The Cluster feature creates very specific and numerous splits, leading the tree to create a more complex model that was too tailored to the training data as overfitting. By removing this feature, I might have reduced the model's complexity, making it more generalizable.

Moreover, the Cluster feature was not very relevant to predicting the target variable and will only introduce noise. Such irrelevant or noisy features can lead decision trees to make spurious splits that don't generalize well to unseen data. Another reason is that the Cluster feature might have influenced the data distribution, causing the model to focus too narrowly on specific patterns not representative of the overall data. So we decided to exclude the Cluster feature for the following advanced models.

Random Forest Classifier & Gradient Boosting Classifier

We turned our focus to the Random Forest Classifier. This model is particularly favored for its proficiency in handling a mix of categorical and numerical data and its robustness against overfitting, making it an ideal choice for our project objective - predicting the safety status of our dataset. The model initially included 'Season,' 'Hour,' 'if_morning,' 'if_afternoon,' 'Day_of_the_Week,' and 'Holiday.' However, we decided to exclude the' Cluster' feature from our analysis, as our prior experiences and the resulting perfect accuracy indicated overfitting. This decision was rooted in the understanding that the 'Cluster' feature significantly contributed to the model's complexity and was not critical for predicting our target variable. To optimize the performance of the Random Forest Classifier, we employed the same grid search strategy using GridSearchCV, combined with 10-fold cross-validation. This approach allowed us to explore a comprehensive range of hyperparameters systematically. Specifically, we adjusted the max_features parameter, determining the number of features to consider when looking for the best split. We tested values from 1 to the maximum number of features in our training data. We made predictions on the test set upon fitting the model with the training data. We also employ a Gradient Boosting Model using the same steps above. The models' performances were primarily evaluated using the accuracy metric, which provided a straightforward interpretation of the model's effectiveness in correctly classifying the safety status. The Random Forest Classifier and Gradient Boosting Model show an accuracy of 0.8683 and 0.8712.

Secondly, we built two different models to predict the dependent variable of crime counts. We used an OLS Linear Regression Model and a CART Decision Tree Regressor.

OLS Linear Regression Model

We also incorporate a Linear Regression Model to predict the count of crime cases in different time slots - morning, afternoon, or night - within four clustered locations in Berkeley. The model utilizes the count of crime cases in each cluster during specific times of the day as its dependent variable. The independent variables include ZIP Code, Season, Hour, If_Morning, If_Afternoon, Day_of_the_week,

Holiday, and Cluster. In terms of performance, the model achieved an R^2 of 0.7930 on the training set and an out-of-sample R^2 of 0.7911. We checked the VIFs for the features to ensure they are all less than 5, which is generally considered acceptable, indicating moderate correlation but not severe multicollinearity. Even though some features' p-values are greater than 0.05, we eventually decided to keep them since their VIFs are all below the threshold, and it is safe to keep all the features. These variables are critical to include based on domain expertise. This model is instrumental in providing insights into the distribution of crime incidents in various areas of Berkeley at different times, facilitating a predictive understanding of crime frequency.
Lastly, we built a Logistic Regression Model to predict the severity of crime.

Logistic Regression Model 2

The second Logistic Regression Model focuses on identifying whether a crime happening in a location and at a time is a severe crime. We first researched UCPD's official documents that defined severe crimes as Homicide, Sexual Assault, Robbery, Aggravated Assault, Burglary, Larceny, Auto Theft, and Arson. These crimes are categorized as Part One Crime. The dependent variable is the severity of the crime, categorized as either part one crime (1) or not part one crime (0). The independent variables remain consistent with the first logistic regression model. This model's performance metrics include an accuracy of 0.8185, a True Positive Rate (TPR) of 0, and a False Positive Rate (FPR) of 0, with a P-threshold value set at 0.4. The TPR and FPR of this model are all zeros which means that our model is underfitting. That's because this model is too simple to capture the underlying pattern of the data. As part of our model evaluation, we plotted the ROC Curve(figure 6), a graphical representation of our model's ability to differentiate between the severity levels of crimes. Ideally, we want the curve to be as close to the top-left corner as possible, indicating a high true positive rate and a low false positive rate. It has an AUC of 0.48. 0.48 is less than 0.50, so our model performs worse than random guessing. This could be due to various factors, such as non-linear relationships that our logistic model cannot capture. We will be looking into more sophisticated algorithms, such as Cart regression or random forests, which can capture complex patterns in the data.

Impact

By accurately predicting crime patterns in Berkeley through features such as locations or time, this highly accurate and updated model will serve as a strong reference for law enforcement such as UCPD or the Berkeley police. The goal of this model, along with police departments' own tools, is to let law enforcement implement safety-related policies and enable the community to have more safety precautions in Berkeley. Sharing which areas related to perhaps what date, the time of the year, or the day will provide more insights on crime data with the community; it can foster a collaborative environment where around the holidays, UC students and residents should try to avoid certain areas. During the afternoon and noon, crimes will happen more frequently; thus, precautionary measures must be taken. Additionally, these insights can be invaluable for policymakers and urban planners, aiding in developing infrastructures and strategies that mitigate crime risks, such as improved lighting, surveillance, and community centers in the blocks predicted to be associated with higher risks.

To further amplify its impact, the model could expand its scope by incorporating a broader range of data sources like social media trends posted by students on Ed, Reddit, or Instagram with related topics, economic indicators in Berkeley areas by blocks, and urban development plans and testing the model's adaptability and scalability by expanding into different regions and incorporating continuous data and model training using the expanded scope and sources. Moreover, integrating real-time data would allow for dynamic crime prediction, enabling more timely and effective interventions.

However, the impact of this model is not uniform across all subpopulations. For example, data inputs related to campus and students versus non-students demographic information could impact the model differently and produce different results. The risk of reinforcing existing societal biases is significant, especially if the model disproportionately predicts crime in areas with certain ethnic or economic demographics. Therefore, ensuring regular audits and updates to the model is crucial to

maintaining fairness and effectiveness for all subpopulations. Thus, the model presents potential negative consequences such as privacy concerns related to misuse of sources of data, discrimination, and biases related to models targeting certain demographics, and policy implementations that could impact regular community life. Thus, the incorporation of the model needs to be carefully implemented with policy regulations built around it and regular audits with fairness, transparency, and collaboration that fits the Berkeley community kept in mind. Possible voting among the entire community needs to be considered before implementation.

Appendix

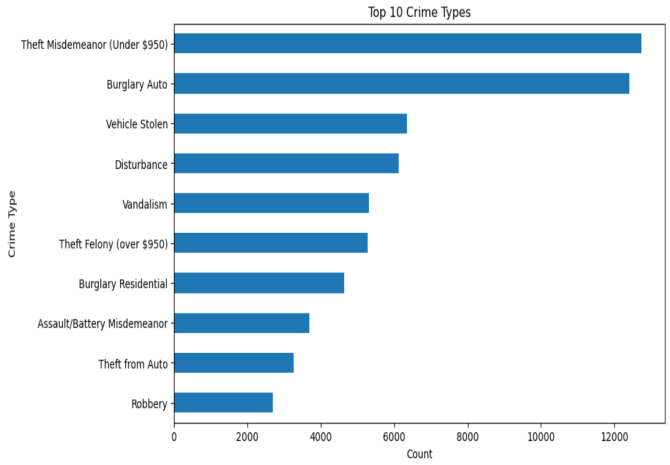

Figure 1: EDA - Top 10 Crime Types from Data

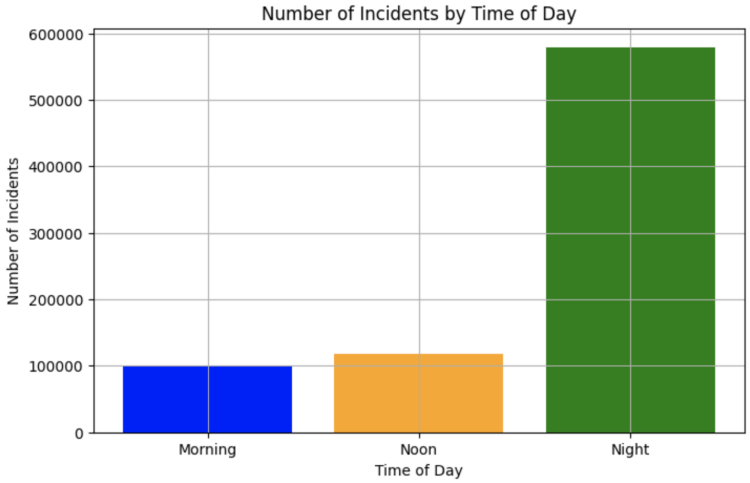

Figure 2: EDA - Number of Incidents by Time of Day

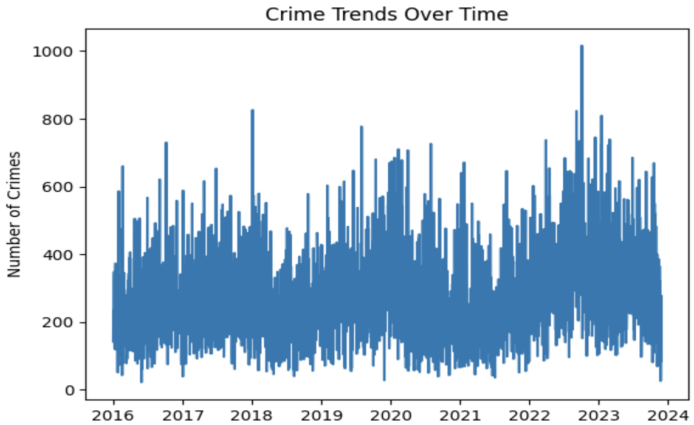

Figure 3: EDA - Crime Trends Over Time

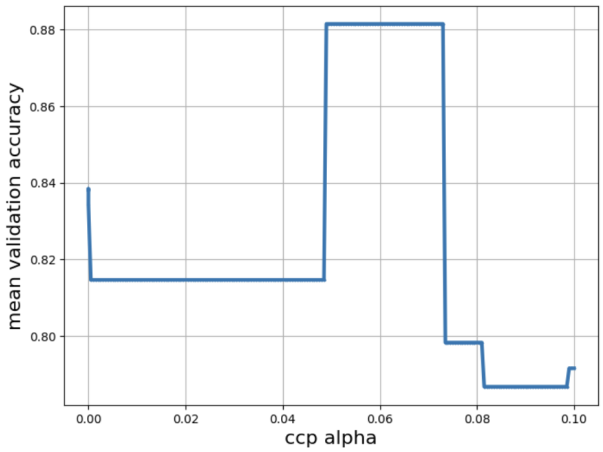

Figure 4: Logistic Regression 1 CCP Alpha Values

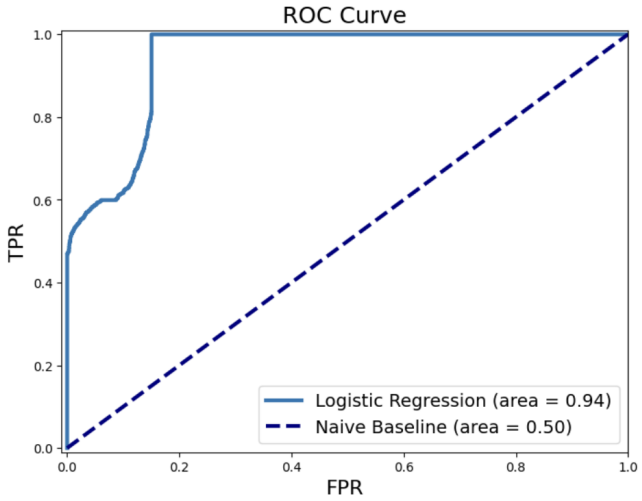

Figure 5: Logistic Regression 1 ROC Curve

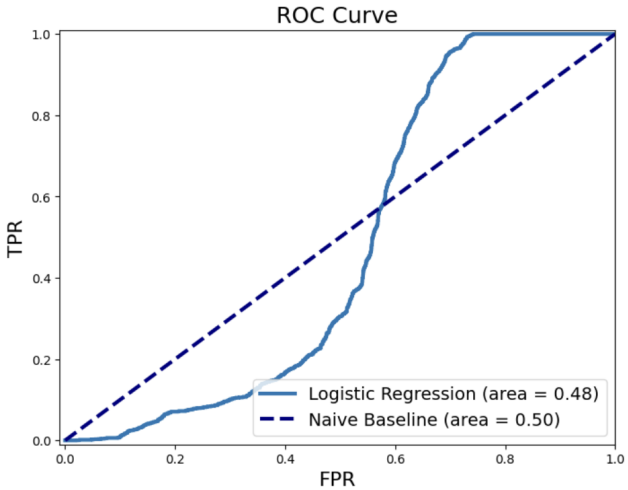

Figure 6: Logistic Regression 2 ROC Curve

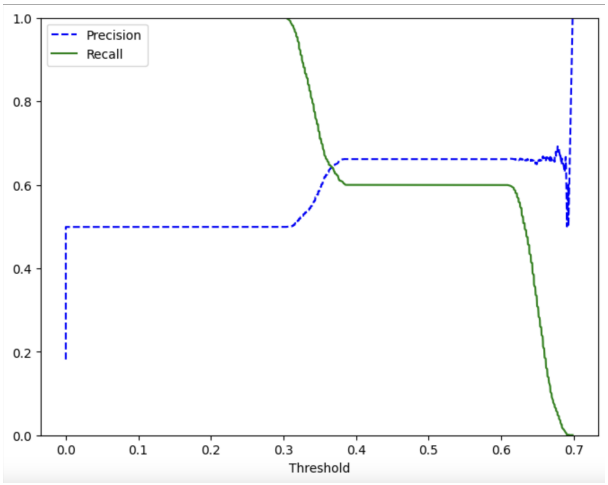

Figure 7: Precision Recall For P Threshold Value

Submission Details:

1. Data Sources:

https://bpd-transparency-initiative-berkeleypd.hub.arcgis.com/datasets/133261097d954b02af81961f9721f841_0/explore (Cases)
https://bpd-transparency-initiative-berkeleypd.hub.arcgis.com/datasets/3be134af40954e19a3d308779a65f175_0/explore (Calls for Service)

2. Code Link:
https://colab.research.google.com/drive/1uQasgh-sOcaV-R3IkWXjPaxlZQGNgiGN#scrollTo=iwFLY7hWmBNd

3. Optional Links:
Official Document for categorizing part one crime:
https://drive.google.com/drive/u/0/folders/1_-45RhRJcoIZYs4RVCga8DgnWQ7WWKw1