Crime Analysis in Austin:

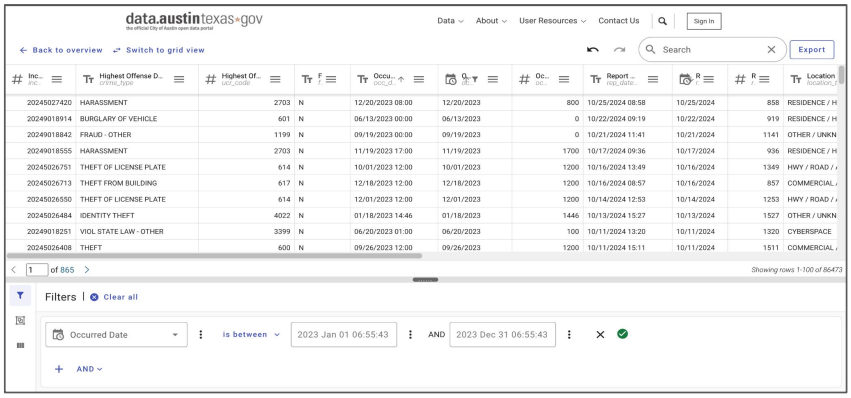
Predicting Crime Based on Patterns in Time and Location

Connor Shen

University of Texas at Austin
SDS 322E
Fall 2024

Introduction

This project analyzes crime patterns in Austin, Texas, using data from the Austin Police Department's crime reports. By examining when and where crimes most frequently occur, as well as the types of crimes common to different locations, we aim to uncover trends that could inform public safety strategies. My analysis explores patterns in time, such as peak hours and seasonal shifts, as well as spatial trends across Austin's council districts. Understanding these patterns can help guide targeted policing efforts, improve resource allocation, and ultimately enhance community safety in Austin.



Crime Reports from Jan 1st, 2023 to Dec 31st, 2023

2,505,051 rows | 19 columns

Unique Row: Each row represents an individual crime report filed in Austin.

Main Variables:

- Highest Offense Description: Details the specific crime (theft, assault).
- Occurred Date Time: Timestamp of the crime, used to extract hour, date, and month for analysis.
- Location Type: Broad category indicating the location (residential, commercial).
- Council District: Specifies the city district where the crime occurred

Predictor Variables

Hour, Council District, Location

Outcome Variable

Crime Type

What are the **time** and **location patterns** of crime in Austin, Texas and how can we use these to try to predict what type of crime will occur?

Methods

crime <chr>	hour <int>	month <ord>	date <date>	location <chr>	council_district <dbl>
Disturbance	8	Dec	2023-12-20	Residential	1
Theft	0	Jun	2023-06-13	Residential	4
Fraud	0	Sep	2023-09-19	Other	9
Disturbance	17	Nov	2023-11-19	Residential	7
Theft	12	Oct	2023-10-01	Transportation/Transit	4
Theft	12	Dec	2023-12-18	Commercial	5
Theft	12	Dec	2023-12-01	Transportation/Transit	7
Fraud	14	Jan	2023-01-18	Other	10
Other	1	Jun	2023-06-20	Other	3
Theft	12	Sep	2023-09-26	Commercial	9
Other	15	Mar	2023-03-07	Residential	7
Fraud	0	Apr	2023-04-01	Other	3
Other	22	Aug	2023-08-26	Residential	1
Theft	12	Oct	2023-10-02	Transportation/Transit	2
Fraud	12	Dec	2023-12-21	Other	9
Theft	7	Oct	2023-10-04	Residential	5
Theft	0	Mar	2023-03-14	Commercial	9
Fraud	0	Oct	2023-10-15	Residential	1
Disturbance	12	Oct	2023-10-27	Residential	2
Disturbance	17	Jun	2023-06-01	Other	4
Assault	12	Oct	2023-10-28	Residential	5
Fraud	12	Dec	2023-12-27	Residential	7
Disturbance	13	Dec	2023-12-05	Residential	3
Assault	22	Mar	2023-03-26	Residential	4
Theft	9	Nov	2023-11-09	Residential	3
Fraud	0	Apr	2023-04-07	Residential	1
Fraud	0	Jul	2023-07-21	Other	7
Other	10	Nov	2023-11-25	Residential	1

1–28 of 85,791 rows Previous 1 2 3 4 5 6 … 36 Next

Clean Dataset

85,791rows | 19 columns

Changed “Council District” to lowercase

- council_district

Split “Occurred Date Time” into Three

- hour
- month
- date

Categorized Crime Types from “Highest Offense Description”

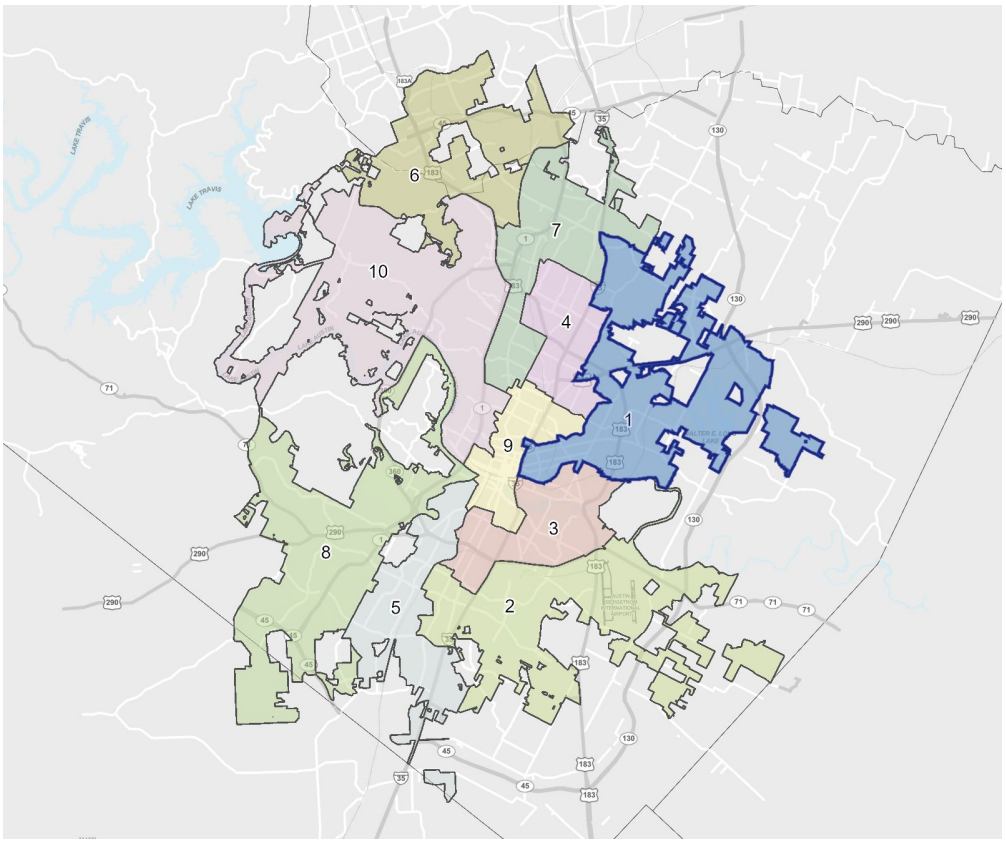
- crime

Categorized Location Types from “Location Types”

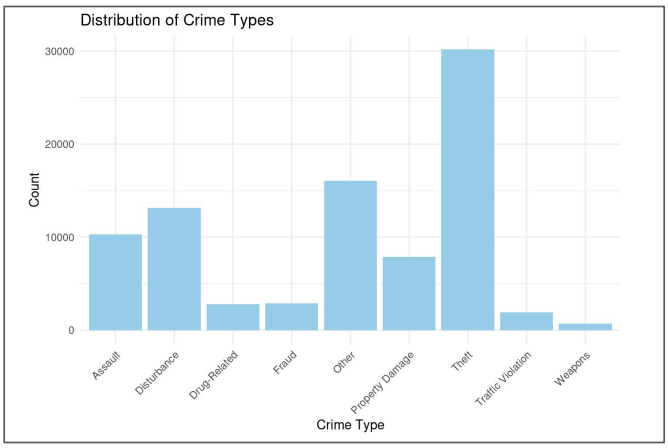
- location

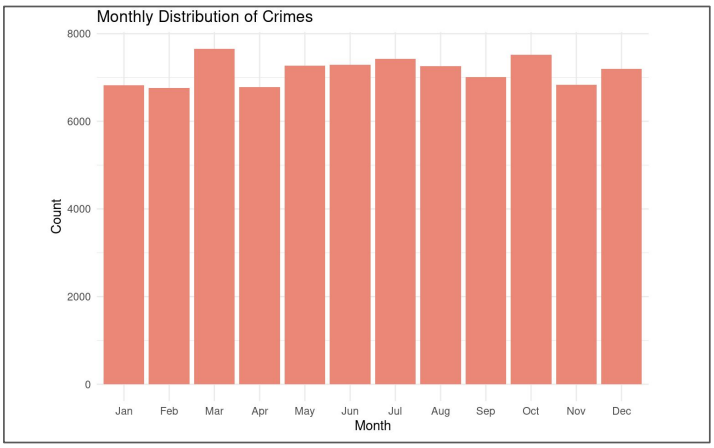
Removed all other columns

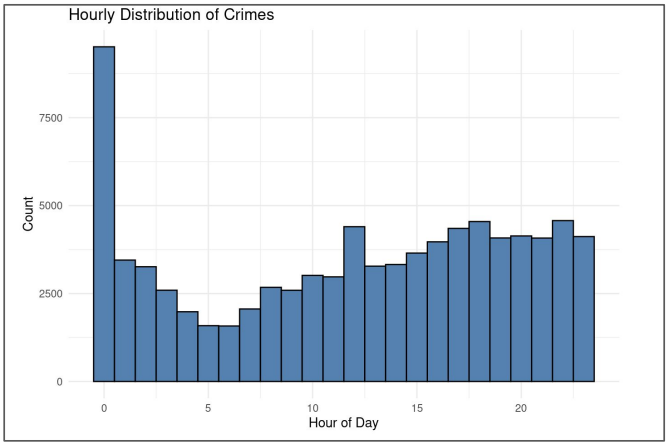
Council Districts



E D A







Distribution of Crime Types:

- Theft is the most common crime type, significantly outpacing others, followed by disturbances and property damage. Crimes like drug-related offenses and weapons violations occur much less frequently, suggesting these are more isolated incidents or less frequently reported.

Monthly Distribution of Crimes:

- Crime rates remain relatively stable throughout the year, with no major seasonal variations. However, minor fluctuations could indicate slightly higher reporting during spring and early summer (March to May).

Hourly Distribution of Crimes:

- Crime rates peak around midnight, likely influenced by nightlife and reduced public oversight during late-night hours. There is a significant dip during early morning hours (5 AM–6 AM) and a gradual increase throughout the day, with another smaller peak in the evening hours (8 PM–10 PM).

Crime Types by Hour of Day:

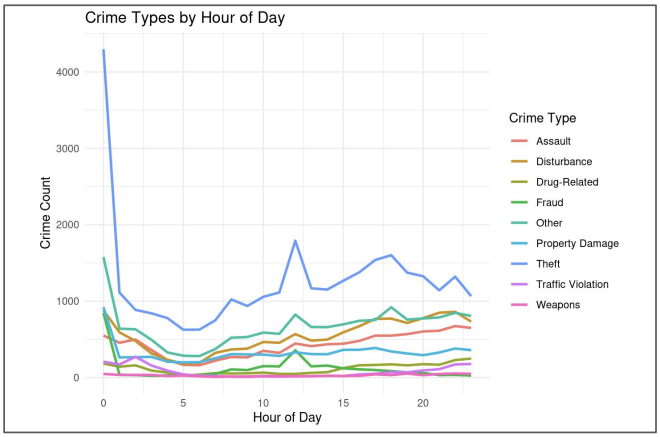
- Theft and disturbances peak sharply at midnight, revealing that late-night hours are the most vulnerable.

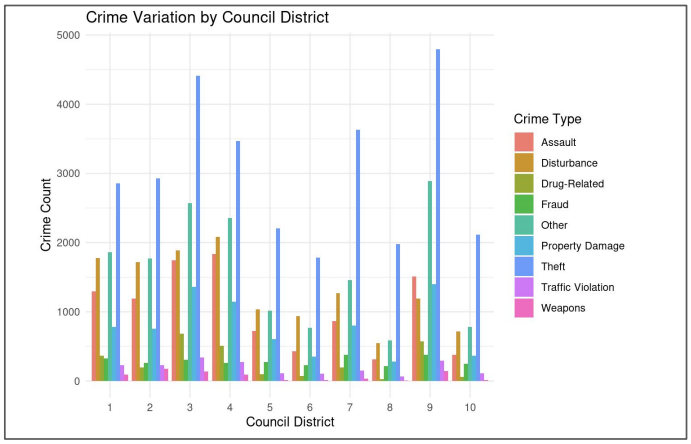
Crime Variation by Council District:

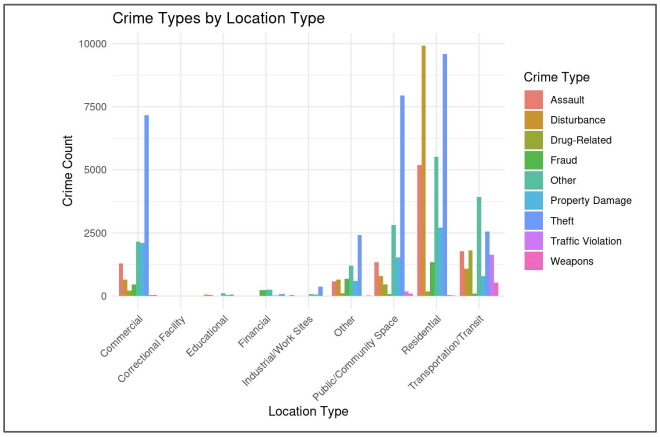
- Districts 3 and 9 consistently report the highest crime counts, making them hotspots for targeted policing and community safety initiatives.

Crime Types by Location Type:

- Commercial and residential areas experience the majority of crimes, especially theft, emphasizing the need for increased security measures in these locations.







Model: K-Nearest-Neighbors

```r
{r, warning=FALSE}
library(pROC)
library(caret)

# Clean up the columns
crime_data <- crime_data |>
  mutate(
    location = as.character(location),
    crime = as.character(crime)
  ) |>
  mutate(
    location = as.factor(location),
    crime = as.factor(crime)
  )

# Normalize numeric columns because kNN is a distance metric
crime_data <- crime_data |>
  mutate(
    hour_scaled = scale(hour),
    district_scaled = scale(council_district)
  )

# Creating the model
fit_knn <- knn3(
  crime ~ hour_scaled + district_scaled + location,
  data = crime_data,
  k = 5  # Number of neighbors
)

predictions <- predict(fit_knn, crime_data) |> as.data.frame()
predictions
```

Ensured Location Type and Crime Type are Characters

- used as.character() in mutate()

Converted Location Type and Crime Type to Categorical Variables

- used as.factor() in another mutate()

Normalized Hour and Council Districts

- used scale() in mutate()
- The reason these variables need to be normalized is because k-nearest-neighbors is a distance metric and therefore, different scales used in numerical variables greatly affect the model
- In this case, hours only contains values from 0 to 23 while council districts contains values from 1 to 11.

Fit Model Over the Entire Dataset and Make Predictions

Prediction Results

Assault <dbl>	Disturbance <dbl>	Drug-Related <dbl>	Fraud <dbl>	Other <dbl>	Property Damage <dbl>	Theft <dbl>	Traffic Violation <dbl>	Weapons <dbl>
0.152941176	0.23529412	0.000000000	0.094117647	0.18235294	0.064705882	0.27058824	0.000000000	0.000000000
0.134419552	0.24439919	0.006109980	0.050916497	0.14460285	0.103869654	0.31568228	0.000000000	0.000000000
0.017621145	0.03524229	0.004405286	0.185022026	0.24229075	0.052863436	0.45374449	0.004405286	0.004405286
0.119496855	0.33333333	0.000000000	0.069182390	0.13207547	0.075471698	0.27044025	0.000000000	0.000000000
0.163265306	0.06122449	0.020408163	0.000000000	0.28571429	0.081632653	0.32653061	0.061224490	0.000000000
0.132352941	0.07352941	0.000000000	0.044117647	0.14705882	0.117647059	0.48529412	0.000000000	0.000000000
0.300000000	0.05000000	0.050000000	0.000000000	0.35000000	0.100000000	0.15000000	0.000000000	0.000000000
0.125000000	0.06250000	0.000000000	0.125000000	0.18750000	0.000000000	0.50000000	0.000000000	0.000000000
0.125000000	0.20833333	0.041666667	0.000000000	0.33333333	0.041666667	0.25000000	0.000000000	0.000000000
0.111111111	0.03968254	0.000000000	0.031746032	0.18253968	0.119047619	0.51587302	0.000000000	0.000000000
0.092307692	0.32307692	0.015384615	0.069230769	0.16153846	0.069230769	0.26153846	0.007692308	0.000000000
0.015267176	0.06870229	0.000000000	0.290076336	0.10687023	0.038167939	0.48091603	0.000000000	0.000000000
0.221818182	0.33454545	0.007272727	0.003636364	0.15636364	0.080000000	0.19636364	0.000000000	0.000000000
0.058823529	0.04411765	0.044117647	0.029411765	0.30882353	0.132352941	0.20588235	0.014705882	0.161764706
0.100000000	0.02000000	0.000000000	0.140000000	0.20000000	0.160000000	0.38000000	0.000000000	0.000000000
0.078125000	0.34375000	0.015625000	0.015625000	0.10937500	0.140625000	0.29687500	0.000000000	0.000000000
0.049281314	0.02053388	0.004106776	0.039014374	0.11293634	0.082135524	0.68377823	0.000000000	0.008213552
0.122065728	0.16901408	0.009389671	0.063380282	0.17136150	0.089201878	0.36854460	0.004694836	0.002347418
0.100840336	0.24369748	0.004201681	0.105042017	0.13865546	0.054621849	0.35294118	0.000000000	0.000000000
0.058823529	0.15686275	0.019607843	0.039215686	0.31372549	0.058823529	0.33333333	0.000000000	0.019607843
0.107784431	0.22155689	0.000000000	0.089820359	0.16167665	0.065868263	0.35329341	0.000000000	0.000000000
0.062500000	0.21634615	0.004807692	0.144230769	0.13461538	0.062500000	0.37500000	0.000000000	0.000000000
0.169082126	0.28502415	0.000000000	0.028985507	0.19323671	0.057971014	0.26086957	0.000000000	0.004830918
0.267100977	0.37133550	0.009771987	0.009771987	0.15635179	0.061889251	0.12052117	0.003257329	0.000000000
0.105555556	0.25555556	0.005555556	0.033333333	0.21666667	0.150000000	0.23333333	0.000000000	0.000000000

```r
{r}
# Convert predictions to probabilities for AUC calculation
# (One-vs-All approach: Calculate AUC for each crime type)

# Get all unique crime types
crime_types <- levels(crime_data$crime)

auc_results <- lapply(crime_types, function(type) {
  roc_response <- as.numeric(crime_data$crime == type)
  roc_predict <- predictions[, type]
  auc(roc(roc_response, roc_predict))
})

# Combine AUC results into a data frame
auc_df <- data.frame(
  Crime_Type = crime_types,
  AUC = unlist(auc_results)
)

auc_df
print(paste("Mean AUC: ",mean(auc_df$AUC)))
```

"Mean AUC: 0.78440670592984"

Retrieve Unique Crime Types

- levels()

Iterate Through Crime Types and Create a Binary Response for each Crime

- roc_response contains binary response
- 1 indicates the current crime type being evaluated.
- 0 indicates all other crime types.

Get the Crime Type Predicted Probabilities and Calculate AUC

Crime_Type <chr>	AUC <dbl>
Assault	0.6702061
Disturbance	0.7734737
Drug-Related	0.8763343
Fraud	0.8480792
Other	0.6309812
Property Damage	0.6794751
Theft	0.7093358
Traffic Violation	0.9410249
Weapons	0.9307501

Cross-Validation

```r
{r, warning=FALSE}
# Set number of folds
k <- 5

# Set a seed for reproducibility
set.seed(322)

# Randomly shuffle the rows of the dataset
crime_data <- crime_data[sample(nrow(crime_data)), ]

# Create k folds
folds <- cut(seq(1, nrow(crime_data)), breaks = k, labels = FALSE)
```

```r
{r}
# Initialize a list to track performance for each crime type across folds
perf_k_all_types <- list()

# Get unique crime types
crime_types <- unique(crime_data$crime)

for (i in 1:k) {

  # Split data into train and test data
  train_not_i <- crime_data[folds != i, ]
  test_i <- crime_data[folds == i, ]

  # Normalize numeric predictors
  preproc <- preProcess(train_not_i[, c("hour", "council_district")], method = c("center", "scale"))
  train_not_i[, c("hour_scaled", "district_scaled")] <- predict(preproc, train_not_i[, c("hour", "council_district")])
  test_i[, c("hour_scaled", "district_scaled")] <- predict(preproc, test_i[, c("hour", "council_district")])

  # Train the model
  train_model <- knn3(
    crime ~ hour_scaled + district_scaled + location,
    data = train_not_i,
    k = 5  # Number of neighbors
  )

  # Calculate performance for all crime types
  auc_per_type <- sapply(crime_types, function(type) {
    roc_response <- as.numeric(test_i$crime == type)
    roc_predict <- predict(train_model, test_i)[, type]

    # Calculate AUC
    auc(roc(roc_response, roc_predict))
  })
  perf_k_all_types[[i]] <- auc_per_type
}

# Combine AUC results into a data frame
perf_k_df <- do.call(rbind, perf_k_all_types)

# Calculate average and standard deviation of AUC for each crime type
perf_summary <- data.frame(
  Crime_Type = crime_types,
  Mean_AUC = colMeans(perf_k_df, na.rm = TRUE),
  SD_AUC = apply(perf_k_df, 2, sd, na.rm = TRUE)
)
perf_summary
print(paste("Mean AUC: ", mean(perf_summary$Mean_AUC)))
print(paste("Mean SD: ", sd(perf_summary$SD_AUC)))
```

Create Folds
- 5 Folds

Set up Performance Tracking
- used a list to store all crime types

K-Fold Cross-Validation
- Split data into train and test sets
  - train_not_i: Contains all rows except those in the i-th fold
  - test_i: Contains only the rows in the i-th fold
- Normalize the numeric predictors
  - Scales hour and council_district to have mean 0 and standard deviation 1 for both train and test data, ensuring fair distance calculations in kNN
- Train the model
- Calculate all AUC for each Crime
  - Converts the response variable into a binary outcome (1 = current type, 0 = all others)
  - Extracts the predicted probabilities for the current crime type
  - Computes the AUC

Aggregate Results
- Combines the list of AUC values for all folds into a single data frame. Rows represent folds, and columns represent crime types

Crime_Type <fctr>	Mean_AUC <dbl>	SD_AUC <dbl>
Assault	0.6126830	0.005123114
Theft	0.6837116	0.001138050
Property Damage	0.5993899	0.008615689
Other	0.5720460	0.003761259
Disturbance	0.7420505	0.005133014
Drug-Related	0.7919132	0.004296576
Fraud	0.7690373	0.009261496
Weapons	0.7982226	0.020215634
Traffic Violation	0.8825788	0.013192249

```
"Mean AUC:  0.716848090942626"
"Mean SD:  0.00584421850406O6"
```

Results

Comparison Between Average Cross-Validation Performance and Overall Model Performance

Overall Model Performance: fitting the model to the entire dataset shows reasonably high AUC values for several crime types, such as:

- Traffic Violation: 0.94
- Weapons: 0.931
- Drug-Related: 0.876
- Fraud: 0.848

However, for some types like "Other" (0.630), "Property Damage" (0.679), and "Assault" (0.670), the performance is lower, indicating difficulties in separating these classes effectively. That said, the model as a whole is fairly good at determining crime types as its average AUC is 0.78

Cross-Validation Performance: The cross-validation results reveal slightly lower AUC values on average, which I expected because cross-validation evaluates the model's ability to generalize to unseen data:

The highest-performing categories still show strong AUCs:

- Traffic Violation: 0.882
- Weapons: 0.798
- Drug-Related: 0.791

However, the same categories as before – "Property Damage" (0.599), "Assault" (0.612), "Other" (0.572) – underperform. The average AUC for the entire model also follows the same trend (0.72).

Takeaway
The average AUC values from cross-validation are slightly lower than the overall model performance, which is expected. The overall performance reflects how well the model fits the dataset it was trained on, whereas cross-validation assesses generalization. Categories like "Traffic Violation" and "Weapons" are being accurately predicted, while others like "Assault" and "Other" remain challenging to model effectively.

How Well Does the Model Predict New Data?

Strong Prediction for Some Crime Types: For categories like "Traffic Violation," "Weapons," and "Drug-Related," the high cross-validated AUCs indicate that the model can effectively predict these crimes on new data, as it consistently distinguishes these classes across folds.

Weaker Prediction for Others: The low cross-validated AUCs for "Other," "Property Damage," and "Assault" suggest difficulty in predicting these categories. This may result from:

- Overlap in the predictors (e.g., similar patterns of occurrence for these crime types).
- Insufficient distinguishing features in the dataset for these categories.

Takeaway
The model is well-suited to predicting specific crime types with clear patterns in the data, such as "Traffic Violation" and "Weapons." However, it struggles with more ambiguous or overlapping classes like "Other" and "Assault," indicating the need for additional predictors or feature engineering to improve performance for these categories. The overall ability for the model to predict what crime happened is good, with an AUC of 0.72.

Discussion

Key Patterns in Time and Location:

- **Theft**: The most frequent crime type and often concentrated in commercial areas and peaking during late-night hours (8 PM–12 AM). This makes sense since this period has reduced business activity and public oversight.
- **Traffic Violations**: Very consistent in both location and time, often occurring during peak commuting hours in districts with high traffic density.
- **Districts 3 and 9:** Consistently reported the highest crime rates, suggesting these are potential hotspots for targeted interventions.

Modeling Performance:

- **High Predictability:** Crimes like traffic violations (AUC: 0.88) and weapons offenses (AUC: 0.80) show strong predictive performance due to clear temporal and spatial patterns.
- **Low Predictability:** Crimes such as assault (AUC: 0.61) and property damage (AUC: 0.60) are harder to model, likely due to overlapping causes or the lack of specific predictors.
- **Cross-Validation Results:** Slightly lower average AUCs compared to the overall model fit, reflecting generalization ability. Predictability is consistent for high-performing crime types but remains a challenge for less-defined categories like "other."
- **Overall:** The model performs pretty well with an AUC of 0.78 for the entire dataset and 0.72 when generalizing new data. This means the model is not overfitting or underfitting and is good at generalizing on new data.

Implications for Austin:

- **District-Specific Interventions:**
  - Focus resources in Districts 3 and 9, especially during theft-prone late-night hours.
- **Traffic Monitoring:**
  - Deploy traffic enforcement during peak hours to reduce violations and related offenses.
- **Enhanced Security in Commercial Areas:**
  - Strengthen preventive measures, such as surveillance and patrols, to deter theft.

Predictive Modeling as a Tool:

While the model effectively predicts crimes with strong patterns, its limitations for more ambiguous categories really exposes the need for further refinement and feature engineering, requiring more data which is very costly to obtain.

Ethical Considerations

Bias in Data Collection:

- The dataset may overrepresent certain areas, such as Districts 3 and 9, due to systemic biases in policing practices, which could inflate crime statistics.
- Underreporting in less-policed districts might hide underlying issues. These biases could lead to disproportionate interventions of specific communities.

Community Impact:

- Predictive policing models must balance the need for proactive interventions with fairness. Transparent use and communication of these tools are important to maintaining trust and preventing potential harm.
- Poorly implemented models can also very easily lead to false positives, which might alienate the population

Data Collection Practices:

- Standardize crime type categories and improve consistency in location data to ensure reliable insights.
- Address gaps in socioeconomic and demographic data, as these are critical for understanding broader crime dynamics.

Unexpected Insights and Challenges

Theft Being the Most Common Crime Committed:

- Theft being the most common crime committed was very surprising because I didn't expect it to be that common in a city. I had thought that that theft would be more common in suburban areas or areas with a smaller concentration of people.

Weaker Patterns in Certain Crimes:

- Low AUC values for crimes like assault and property damage highlight the need for additional predictors, such as socioeconomic factors, weather, or real-time incident reporting.
- There could also be a lot of overlap in their predictors, making it very hard to distinguish between crimes accurately.

Data Limitations:

- Inconsistencies in location and crime type coding required significant preprocessing, potentially affecting model accuracy as information was lost when putting crimes in buckets.

Future Directions

Enhance Data Integration:

- Use socioeconomic data, weather patterns, and event schedules to provide deeper context for predictions.
- Transition from council district-level data to more granular neighborhood or hotspot-specific insights.

Expand Analytical Scope:

- Conduct time-series analyses to evaluate long-term trends in crime rates and assess the impact of past interventions.
- Investigate patterns within more specific crime categories to find deeper, underlying patterns..

Reflection, Acknowledgments, and References

Reflection:

Challenges Faced:

- Data Preparation: Consolidating the dataset into a tidy format was one of the most demanding aspects because of inconsistencies in the Location Type and Highest Offense Description fields. Custom categorizations for crime types and locations required me to carefully create groupings.
- Modeling Complexities: Certain crime categories, such as "Other" or "Property Damage," were harder to model due to overlapping patterns or insufficient distinguishing features in the dataset. These challenges really revealed the limitations of working with real-world, imperfect data.
- Feature Engineering: The lack of contextual variables, such as socioeconomic factors, weather, or real-time events, limited the model's ability to improve predictions for less structured crime types.

Lessons Learned:

- Data Cleaning and Categorization: Handling real-world data showed the importance of structured and consistent data entry, as small inconsistencies can significantly affect analyses. Developing efficient cleaning pipelines using R improved data accuracy and interpretability.
- Model Evaluation: Cross-validation is a robust framework to evaluate a model's ability to generalize, revealing strengths for predictable crime types and limitations for ambiguous categories.

Acknowledgments:

I would like to express my gratitude to Professor Guyot and the teaching assistants for their valuable guidance throughout the project. Their feedback greatly enhanced the rigor of my data analysis and interpretation. Special thanks to the City of Austin's Open Data Portal for making crime data publicly accessible, enabling this study to contribute actionable insights.

References:

Dataset:

Austin Crime Reports, available from the City of Austin's Open Data Portal:

https://data.austintexas.gov/Public-Safety/Crime-Reports/fdj4-gpfu/about_data

Background Research:

Kuo, Pei-Fen, et al. "A Promising Example of Smart Policing: A Cross-National Study of the Effectiveness of a Data-Driven Approach to Crime and Traffic Safety." Case Studies on Transport Policy, Elsevier, 5 Sept. 2019. https://www.sciencedirect.com/science/article/pii/S2213624X19301336.

Whitworth, Adam. "Local Inequality and Crime: Exploring How Variation in the Scale of Inequality Measures Affects Relationships Between Inequality and Crime." Urban Studies, vol. 50, no. 4, 29 Aug. 2012, pp. 725–741. https://doi.org/10.1177/0042098012455716.

External Resources:

- dplyr: Documentation for data manipulation functions.
  - https://dplyr.tidyverse.org/
- ggplot2: Documentation for data visualization functions.
  - https://ggplot2.tidyverse.org/
- caret: Comprehensive guide for training and evaluating machine learning models.
  - https://topepo.github.io/caret/
- pROC: Documentation for Receiver Operating Characteristic (ROC) curve analysis and AUC calculation.
  - https://cran.r-project.org/web/packages/pROC/index.html