Course Project Part 2

Yu-Hsuan (Monica) Ko

Business Problem

In the high-crime city of Chicago, residents must prioritize their safety. This project aims to analyze crime trends in the area over the past five years. Our goal is to provide insights that can help residents live safely in Chicago. We can allocate resources effectively and enhance public safety by identifying these trends.

Logistic Regression Model

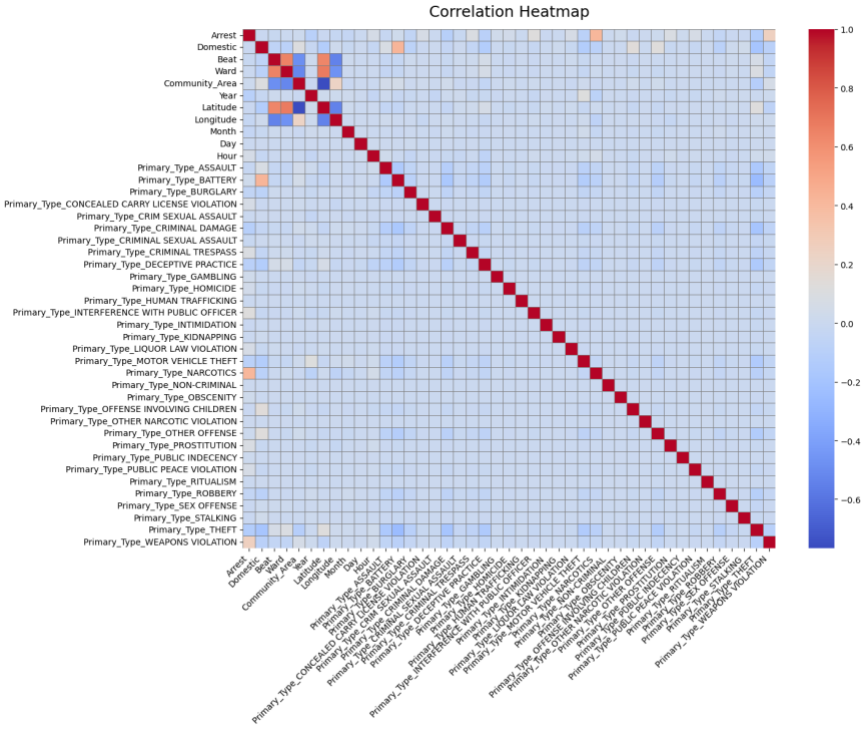
This project aims to predict whether a crime results in an arrest. We built a logistic regression model using 100,000 historical Chicago crime records from the past five years (2019-2023). We use arrest as the response variable and other helpful information as explanatory variables. The dataset includes several variables: crime type, location, date, and time. The model formula is

Arrest ~ Year + Longitude + Month + Hour + Primary_Type_ASSAULT + Primary_Type_BATTERY + Primary_Type_BURGLARY + Primary_Type_CONCEALED CARRY LICENSE VIOLATION + Primary_Type_CRIMINAL DAMAGE + Primary_Type_CRIMINAL TRESPASS + Primary_Type_DECEPTIVE PRACTICE + Primary_Type_GAMBLING + Primary_Type_HOMICIDE + Primary_Type_INTERFERENCE WITH PUBLIC OFFICER + Primary_Type_LIQUOR LAW VIOLATION + Primary_Type_MOTOR VEHICLE THEFT + Primary_Type_NARCOTICS + Primary_Type_OBSCENITY + Primary_Type_PROSTITUTION + Primary_Type_PUBLIC PEACE VIOLATION + Primary_Type_ROBBERY + Primary_Type_THEFT + Primary_Type_WEAPONS VIOLATION

This model has 23 parameters with $R2 = 0.29$. See Appendix 1 for the model summary and estimated coefficients.

Key Considerations in Modeling

This dataset primarily consists of categorical variables, with only a few numerical variables included. While some of the variables utilize dummy coding methods, most numerical variables exhibit low correlations with one another. Consequently, the current variables may not be sufficient for developing a model with strong predictive power and a higher R-squared value, given their low correlation. This situation reflects a near-realistic scenario: crime incidents occur randomly, and the allocation of police resources cannot be determined with precision. Additionally, we have filtered out variables with an absolute correlation higher than 0.2.



Adjust the variables for model use

In Project 1, we transformed some data and retained most of the variables for this project. However, some things could have been improved when we initially included these variables in the logistic model.

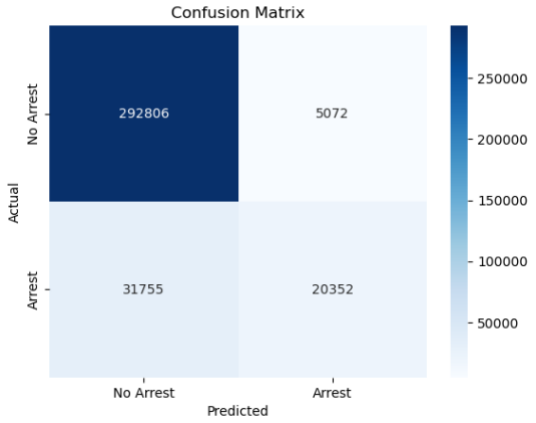
1. The dataset has some highly similar variables like district, longitude, latitude, x coordinates, and y coordinates. We decided to keep only longitude and latitude.
2. In the cleaned data from Project 1, the local description and description have too many different values, and we decided to remove them.
3. The original dataset has many variables, which may cause noise for model fitting. We decided to check correlations to retain some variables that might be useful.

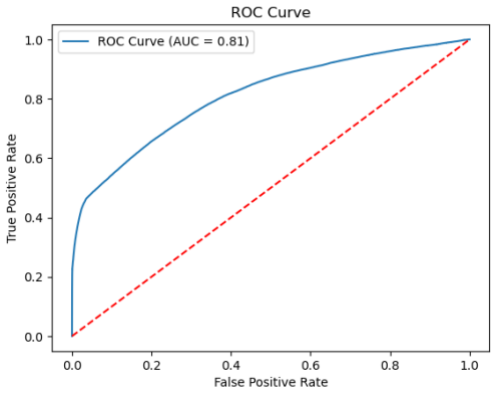
Model Results

After cleaning the data again, we put the remaining variables into the logistic model; the model result is below.

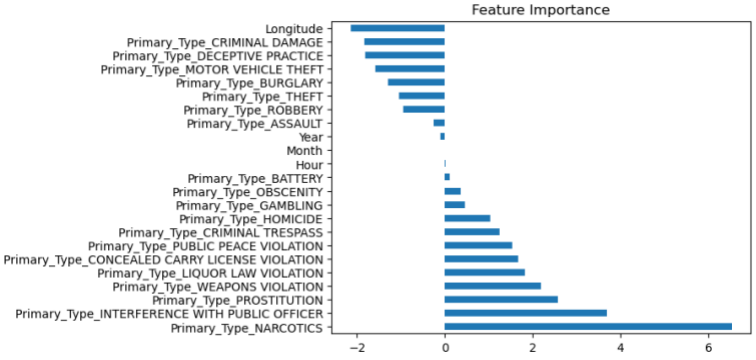
```
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.98      0.94    297878
           1       0.80      0.39      0.53     52107

    accuracy                           0.89    349985
   macro avg       0.85      0.69      0.73    349985
weighted avg       0.89      0.89      0.88    349985
```

The confusion matrix provides insight into the classification performance of the logistic regression model. The model demonstrates a reasonable ability to distinguish between instances of arrest and non-arrest but has a relatively higher false-negative rate, indicating that some arrests are missed by the model. The precision rate is 0.80, indicating the model's caution in predicting arrests and ensuring that most "Arrest" predictions are accurate. However, it may overlook some actual arrests, leading to false negatives. The recall rate is 0.39. While the model accurately predicts arrests, it fails to identify a significant portion of actual arrests, which may be critical in contexts where missing arrests is a concern. An AUC score of 0.81 suggests that the model has good discriminative power, meaning it is effective at distinguishing between arrests and non-arrests.





The feature importance reveals that narcotics-related offenses are the strongest predictor of arrests, followed by interference with public officers, prostitution, and weapons violations. Conversely, longitude, criminal damage, and deceptive practices are negatively associated with arrests, indicating a lower likelihood of these crimes leading to arrests. Policy efforts should focus on high-arrest predictors such as narcotics and weapons violations while examining geographical trends to address potential underreporting or underenforcement.



Other Insights

The analysis shows that while the model achieves strong precision and an AUC of 0.81, its low recall (0.39) highlights the need for additional contextual data and advanced modeling techniques to improve performance. Key predictors, such as narcotics and weapons violations, should guide policy focus, while geographical trends warrant further investigation to address potential biases or underreporting.

Appendix 1

Logit Regression Results

Dep. Variable:	Arrest	No. Observations:	1166616
Model:	Logit	Df Residuals:	1166592
Method:	MLE	Df Model:	23
Date:	Thu, 12 Dec 2024	Pseudo R-squ.:	0.2928
Time:	02:34:15	Log-Likelihood:	-3.4717e+05
converged:	False	LL-Null:	-4.9088e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0411	nan	nan	nan	nan	nan
Year	-0.1504	0.002	-68.362	0.000	-0.155	-0.146
Longitude	-3.4718	nan	nan	nan	nan	nan
Month	-0.0206	0.001	-21.720	0.000	-0.022	-0.019
Hour	0.0134	0.000	28.689	0.000	0.012	0.014
Primary_Type_ASSAULT	-0.1700	0.013	-12.764	0.000	-0.196	-0.144
Primary_Type_BATTERY	0.2081	0.011	19.207	0.000	0.187	0.229
Primary_Type_BURGLARY	-1.0520	0.024	-43.668	0.000	-1.099	-1.005
Primary_Type_CONCEALED CARRY LICENSE VIOLATION	5.2373	0.183	28.596	0.000	4.878	5.596
Primary_Type_CRIMINAL DAMAGE	-1.2992	0.016	-78.745	0.000	-1.332	-1.267
Primary_Type_CRIMINAL TRESPASS	1.3616	0.016	83.241	0.000	1.330	1.394
Primary_Type_DECEPTIVE PRACTICE	-1.6583	0.022	-73.875	0.000	-1.702	-1.614
Primary_Type_GAMBLING	6.2522	0.711	8.795	0.000	4.859	7.645
Primary_Type_HOMICIDE	1.3752	0.037	37.093	0.000	1.303	1.448
Primary_Type_INTERFERENCE WITH PUBLIC OFFICER	4.1851	0.063	66.375	0.000	4.062	4.309
Primary_Type_LIQUOR LAW VIOLATION	6.7411	0.380	17.760	0.000	5.997	7.485
Primary_Type_MOTOR VEHICLE THEFT	-1.4221	0.022	-65.917	0.000	-1.464	-1.380
Primary_Type_NARCOTICS	6.0732	0.048	127.609	0.000	5.980	6.166
Primary_Type_OBSCENITY	2.7154	0.144	18.802	0.000	2.432	2.998
Primary_Type_PROSTITUTION	6.1070	0.231	26.422	0.000	5.654	6.560
Primary_Type_PUBLIC PEACE VIOLATION	1.7569	0.030	58.344	0.000	1.698	1.816
Primary_Type_ROBBERY	-0.7874	0.021	-36.990	0.000	-0.829	-0.746
Primary_Type_THEFT	-0.9204	0.012	-74.212	0.000	-0.945	-0.896
Primary_Type_WEAPONS VIOLATION	2.3655	0.014	171.519	0.000	2.339	2.393