ECO481:

Arrest Prediction Tool

Joseph Junior Nchendia Nkeng (*nkengjos)*
joseph.nkeng@mail.utoronot.ca

Vivaan Bhaskar (*bhaskarv)*
vivaan.bhaskar@mail.utoronto.ca

Yin-Jie Tseng (*tsengyi2)*
yinjie.tseng@mail.utoronto.ca

Crime not only has a negative impact on the victim but poses significant social and economic effects. Areas of high crime correlate with lower property value and neighborhood satisfaction, furthermore, increased crime and arrest rates have significant negative effects on determining election results.

The negative socio-economic and political effects of crime lead us to our main research question : “Will a reported crime result in an arrest?”. Exploring further into this question and analyzing statistics on crimes lead to our next question: “What factors contribute to an arrest being made?”. Our research paper seeks to answer these questions by constructing a model that predicts whether a reported crime will result in an arrest. We then proceed by looking at the most important factors that contribute to an arrest being made in the built model.

Using our model for the prediction on crime and arrest can assist various stakeholders in their decision making, by understanding features and patterns to optimize resource allocation while being a preventative and exploratory tool towards crime.

A prime example can be seen in Raymond Surette’s article entitled “Crimes, arrests, and elections: Predicting winners and losers”. Surette’s shows using data about counties in the US that crime and arrest rates are a determining factor in a sheriff’s election or re-election in their county.

Our motivation and methodology drew from three key research papers within this field. The first being the aforementioned paper by Raymond Surette. Surette’s research centers around the “sheriff-election” model, which found a negative relation between arrests rates and the probability of a sheriff’s re-election. To be precise, sheriffs in counties with high arrest rates were more likely to lose elections. Hence, Surette argues; during elections, Sheriffs should include crime and arrest rates in their campaign strategy. Surette’s research shares similarities with our research in that it studies the rate of arrest as well as the greater implications of crime and arrests on society. His research however, covers a greater scope than ours. As his research

focuses on factors that determine sheriff-elections, his study explores additional factors such as political affiliation which is not within our scope.

While exploring existing research on crimes and arrest, nearly all research topics were based on predicting future crime rates. One such paper was Khatun et al.'s "Data mining technique to analyze and predict crime using crime categories and arrest records" uses machine-learning models to predict crime type and location of crime on predicting criminal behavior and patterns in crime. This research was beneficial in helping us construct our model as their study was also based on machine learning models. However, their research, like many before it, aimed to predict criminal behavior based on type of crime and location, while our model aims to find probability of arrest and the determining features that would lead to an arrest.

One of the few papers that we found relating to our research question was: "Predicting arrest probability across time: An exploration of competing risk perspectives" by Michelle Bolger. In this paper, she analyzes some of the predictors that could increase the probability of arrests in a group of individuals. She found consistent delinquency by individuals over a period of 10 years increased the probability of arrest. Although we found it challenging to collect tracking information about individuals in our dataset, we made sure to include all the crimes reported throughout all the years in our dataset, hoping it could capture Bolger's findings. It is worth noting that in Bolger and Surrete's research, they predefined the arrest determining variables whereas in our study, we take a more global look at these contributing variables by fitting all possible variables in our dataset to the model, then selecting the best.

Our area of study already consists of many existing research on crime and arrests. However, our study is distinct in that existing studies aim to predict future crimes and arrest rates given specific features, while our study takes a broader scale to look at factors that determine whether an arrest is made. In this already saturated research space, we did not want to limit ourselves through our own biases in determining what features to study. Instead, we wanted to use machine learning techniques to find patterns that could lead to further in-depth studies into

specific features. One such finding was the discrepancies our studies discovered in arrest rates for domestic and non-domestic reports on crime, in which we could find no existing research analyzing the one-third difference in rates of arrest between the two.

In this paper, we fit four models: decision trees, random forest, naive bayes and logistic regression to our data and compare them based on accuracy and precision on test data. The best model is then chosen to be Random Forest.

DATA

This section introduces the datasets used to conduct our analysis. We choose to focus our analysis on the **City of Chicago, Illinois, USA**. We source data from the City of Chicago database (1). The dataset that we choose includes **reports of incidents of crime** from 2001 to present (updated to two weeks before the time data is accessed). We use data from **2001-2022**.

The dataset includes the following information of a reported incident of crime:

1. **Date and Time** of the report: This variable is cleaned to have separate columns for each component of date (day, month, year) and time (hour, minute, second). We require the month, year and hour observations.
2. **Location**: This includes several variables - the Block, District, Community Area and Ward of where the crime was reported from. They are different ways to categorize areas in the city. For our analysis, we focus on Community Area codes. A description of the location where the reported crime was observed. Additionally, the Beat is also reported, which is the station which was contacted to report said crime.
3. **Type of Crime**: It categorizes the type of crime into primary type and offers a description (short). There are two codes that represent how reports are recorded as per code reporting systems - IUCR (Illinois State) and the FBI code. We focus on IUCR codes as they follow

the same format of primary type, which is represented by the first two digits (from left), followed by the description.

4. **Arrest** Variable: Dummy variable indicating whether the report ended up having found 'probable cause' to make an arrest.
5. **Domestic** Variable: Dummy variable indicating whether the report was domestic, i.e. the suspect and victim are family members.
6. The exact geographical location of the reported crime (we do not use these statistics).

Here, we will add new columns for:

1. **Time of Day**: We will divide the day into four parts, in six hour intervals, starting from midnight (12AM). This is done to account for changes in the number of people out during different times of the day, as well as different officers for day vs night.
2. **Quarter**: Group months by quarters, this also helps in accounting for seasonal effects.
3. **IUCR primary**: Creating a new variable to split the IUCR in the middle, to consider only primary type and have a numerical value for the same.
4. Community Area statistics: We look to include data on demographics - population, median income and visible minority percentage in the area. The data is taken from (2).

METHODOLOGY

This section introduces our approach taken to explore the methods in which various stakeholders concerned with crime statistics can use Machine Learning techniques to optimize their decision making strategies. This could, as mentioned earlier, involve the optimization of revenue allocation, personnel allocation and most of all, help in creation of prevention measures.

The first part of our analysis is exploratory. We look at the different variables in the data sets individually (mentioned in the Data section) to understand their pattern and distribution over 20 years of data. This will help us in our understanding of putting machine learning algorithms to use for these features. It also helps to know the descriptive statistics of the dataset to make any further analysis. We will also use regressions to understand, first, the relationship between community level statistics - population, income, and minority percent, with the arrest percentage in that area. Next, a control is added for each statistic for whether a report was domestic or not. Lastly, the relationship between IUCR codes with arrest percentage are seen (with the domestic control). The domestic control is introduced to explore whether a report that is domestic has an impact on a reported incident of crime resulting in an arrest.

The second part of our analysis focuses on building a predictive model which predicts whether a reported crime will result in an arrest, depending on the features of the report. We compare the predictive ability of four machine learning techniques:

1. **Logistic Regression**: The model predicts the log-odds of a success, in this context, an arrest being made, given the features that are input into the model. The interpretation requires the linearity assumption, hence, if the inputs are linear, the model will make more sense. In our situation, the inputs (excluding categorical variables) are linear.
2. **Gaussian Naive Bayes**: This predictive model predicts the probability of an event (arrest being made) conditional on an event (a given feature). This model is interesting in this context, as our research question can be answered using conditionality, however, it requires

an independence assumptions (variables are independent from one another) which may not be satisfied.

3. **Decision Trees**: This model learns patterns from the data which are used to make classifications depending upon feature values. It starts from a root and ends on a leaf with multiple such paths. These paths are classification rules, which can be multiple (two or more conditional statements) or simple (single).
4. **Random Forest**: This method is an extension of the Decision Trees model which uses many uncorrelated decision trees to make the classification rules. It uses bagging, where individual data points are used more than once to train data on multiple sample sets; and feature randomness, generating random subset of features.

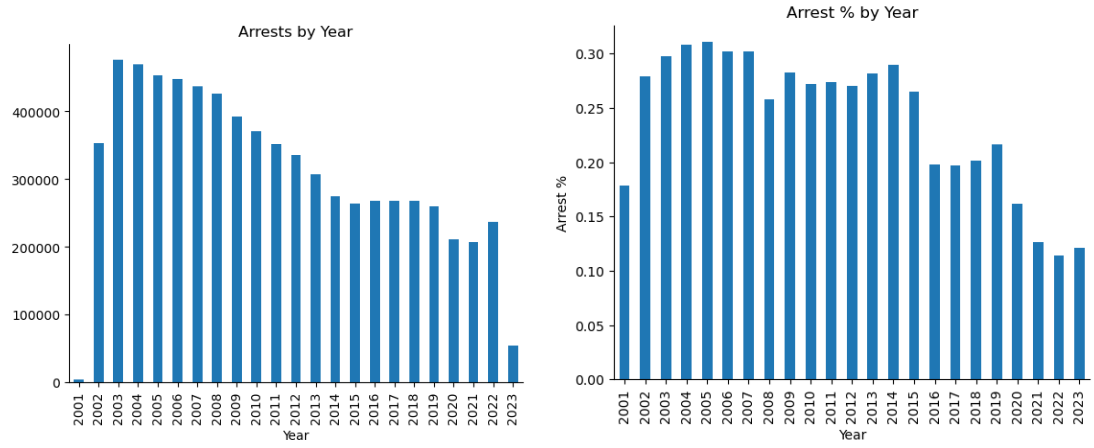
We will calculate the Accuracy, Precision and Area Under Curve (AUC) of each of the models and select the one which performs the best. For the purposes of choosing the best model, we will use cross-validation - expand here.

Lastly, we display the feature importance from the top model and discuss how the findings can help the various stakeholders concerned with crime and arrest rates.

Descriptive Statistics

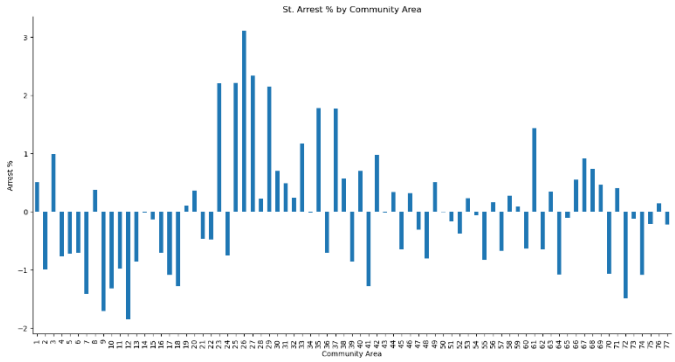
This section presents descriptive statistics of key variables in the dataset that we consider.

1. **Arrest Variable**

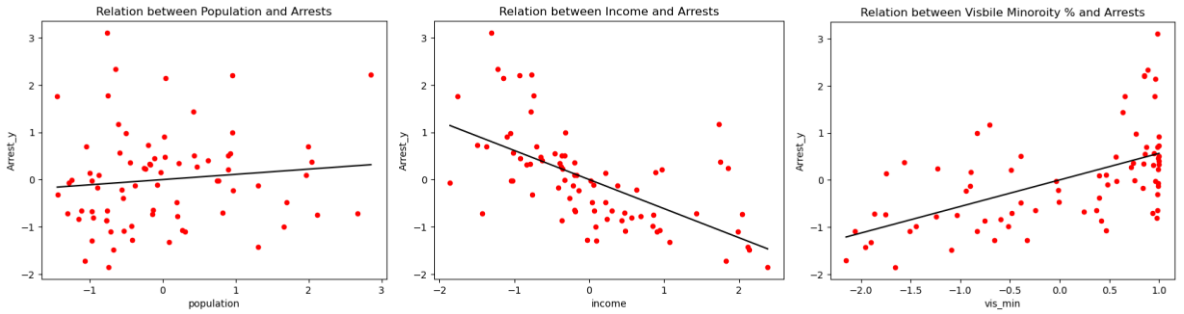


The dataset has 7.14 million reports of incidents of crime, with an average arrest rate of approximately 26%. The graph on the left shows the number of arrests have fallen over the years,from 2002-2013, which has steadied since. We see a considerable drop during 2020-21, which can be attributed to the pandemic*. The graph on the right shows arrest rates over the timeframe, which too has fallen in recent years.

2. **Community Variable**



The graph above shows the standardized arrest rates by community area which goes from 1 to 77 from left to right. We can see that some areas have considerably lower arrest rates, with the same having much higher. We now explore a few features of community areas to understand these differences. Scatterplots of arrest rates with population, income and visible minority percentages are shown below.



We can see that there is no clear relation with population, while income has a clear negative relation, implying that higher income community areas have lower arrest rates on average. Inversely, a negative relationship is seen between visible minority and arrest percentage. The regression outputs for individual models as well as the combined model is shown below. All variables in the analysis are standardized.

OLS Regressions

	Model 1	Model 2	Model 3	Model 4
const	0.00	0.00	0.00	0.00
	(0.11)	(0.09)	(0.10)	(0.09)
population	0.11			0.23**
	(0.11)			(0.09)
income		-0.62***		-0.48***
		(0.09)		(0.16)
vis_min			0.56***	0.21
			(0.10)	(0.16)
R-squared	0.01	0.38	0.32	0.44
R-squared Adj.	-0.00	0.37	0.31	0.41
R-squared	0.01	0.38	0.32	0.44
No. observations	77	77	77	77

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

The first model is not able to explain any variation in the data and does not have a statistically significant relation. The second model, with income, shows a large negative value, which means that when median income is increased (decreased) by one standard deviation, the arrest rate of that community area goes down (up) by roughly .6 standard deviation, on average. This result is statistically significant. Similarly, the positive coefficient implying a direct relationship, an increase in visible minority percentage by one standard deviation will increase the arrest rate of that community area by half a standard deviation on average. The last model, which includes all variables, explains 44 percent of variation in the data, and has statistically significant results for all the metrics. The coefficients are given and have similar implications, however, an increase is to be interpreted with other variables as constant.

3. **Domestic Variable**

Domestic	Arrest_x	Arrest_y
0	6142039	0.270415
1	997508	0.187425

OLS Regressions

	Model 1	Model 2	Model 3
const	0.26***	0.26***	0.13***
	(0.00)	(0.00)	(0.00)
vis_min	0.04***		
	(0.00)		
income		-0.05***	
		(0.00)	
IUCR_primary			0.01***
			(0.00)
Domestic	-0.10***	-0.10***	-0.06***
	(0.00)	(0.00)	(0.00)
R-squared	0.01	0.01	0.06
R-squared Adj.	0.01	0.01	0.06
R-squared	0.01	0.01	0.06
No. observations	7139547	7139547	7139547

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

We can see from the first table that (given by the Arrest_y column) arrest rates are a third lower for domestic than non-domestic reports. In the second output, we run regressions with average arrest rates as the independent variable. For this regression, no values are standardized. The domestic variable for each model has a negative coefficient (the last model having lower coefficient), which implies that if the report is domestic, it has a 10 percentage point lower chance of resulting in an arrest, on average. The third model has an interesting interpretation as IUCR codes are coded as per the severity of crime as decided by the State of Illinois. Hence, can be interpreted as, the arrest rate predicted by the model for an IUCR code 1 crime, as 14%, which increases by one percentage point as the severity of crime goes down (1 being highest to 51 being lowest).

4. **Time of Day**:

Time_of_Day	Arrest_x	Arrest_y
1	1199414	0.211935
2	1452736	0.223041
3	2214279	0.259038
4	2273118	0.306214

1 represents the start of a new day, from midnight and categorizes the day into four parts, with six hour intervals. They can be thought of as night, morning, afternoon, and evening. We see that as the day progresses, more reports are made and the arrest rate increases drastically, with it increasing by roughly 50% from morning (2) to evening (4).

RESULTS

LogReg

	precision	recall	f1-score	support
0	0.79	0.98	0.87	297084
1	0.44	0.06	0.10	82571
accuracy			0.78	379655
macro avg	0.61	0.52	0.49	379655
weighted avg	0.71	0.78	0.71	379655

RF

	precision	recall	f1-score	support
0	0.87	0.94	0.90	297084
1	0.68	0.47	0.56	82571
accuracy			0.84	379655
macro avg	0.77	0.71	0.73	379655
weighted avg	0.83	0.84	0.83	379655

Tree

	precision	recall	f1-score	support
0	0.86	0.93	0.90	297084
1	0.65	0.47	0.54	82571
accuracy			0.83	379655
macro avg	0.76	0.70	0.72	379655
weighted avg	0.82	0.83	0.82	379655

Naive_Bayes

	precision	recall	f1-score	support
0	0.79	0.94	0.86	297084
1	0.35	0.11	0.16	82571
accuracy			0.76	379655
macro avg	0.57	0.53	0.51	379655
weighted avg	0.69	0.76	0.71	379655

Model Comparison

IUCR_primary	0.475721
Month	0.205274
Beat	0.111038
Time_of_Day	0.083463
Ward	0.040645
Domestic	0.030745
Arrest_y	0.014920
District	0.008465
income	0.008268
vis_min	0.007266
Community Area	0.005555
Arrest_x	0.004977
population	0.003662

Random Forest Most Important Variables

CONCLUSION

The findings of our 4 machine learning models can be seen in “Model Comparison”. “Accuracy” put simply, is the percentage of correct predictions. We see the Random Forest has the highest accuracy at 84%.

Because our datasets are imbalanced, with 297084 non-arrests and 82571 arrests made. We need to also analyze the “Precision”, which is the ratio between True Positive and all Positives. We find this is also highest in Random Forest, where the macro avg is at 77% Furthermore, when looking at the “Recall”, which the ratio of True Positive over True Positive and False Negatives. We find that the recall is the highest in Random forest, where the macro avg is at 71%

From our model comparison, we can conclude that our optimal prediction model is chosen to be Random Forest, as it scores highest in accuracy, precision and recall.

Using our Random Forest model, we find the ten most important variables that contribute to an arrest are: IUCR_primary,Month, Beat, Time_of_Day,Ward, Domestic, Arrest_y, District, income, vis_min.

Notably, IUCR_Primary has a coefficient of 0.475, hence it is by far the most important variable in determining whether an arrest is made. This tells us that the type of crime committed is the main determinant in whether an arrest is made.

The only other variables with coefficient above 0.1 are Month (0.2) and Beat (0.111). These are variables indicating time and location where crime is reported.

Hence, from our random forest model, we conclude that the type of crime is the most important determining factor, followed by Month and where the crime was reported.

Though our model can give a prediction to 84% accuracy, our model may be limited by the limitation of the variables available within our data set. Notably, there is not a variable for the race of the individual that was or was not arrested. With growing discussion about police race relations, this seems to be an important variable that was not included. Furthermore, as we did not want to impose our own biases, our model may face the curse of dimensionality, which is when a dataset has too many features, making it difficult for the model to identify patterns.

REFERENCES

(1) https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data

(2) https://github.com/dssg/411-on-311/blob/master/data/chicago-community-areas.csv

(3) Bolger, M. A. (2018). Predicting arrest probability across time: An exploration of competing risk perspectives. Journal of Criminal Justice, 59, 92–109. https://doi.org/10.1016/j.jcrimjus.2018.05.008

(4) Surette, Raymond. "Crimes, Arrests, and Elections: Predicting Winners and Losers." Journal of Criminal Justice, vol. 13, no. 4, 1985, pp. 321–327., https://doi.org/10.1016/0047-2352(85)90002-9.

(5) Khatun, M. R., Ayon, S. I., Hossain, M. R., & Alam, M. J. (2021). Data mining technique to analyse and predict crime using crime categories and arrest records. *Indonesian Journal of Electrical Engineering and Computer Science*, *22*(2), 1052-1060 https://doi.org/10.11591/ijeecs.v22.i2.pp1052-1060