**How to Get Arrested in Chicago**

Using Logistic Regression to predict the chance of an arrest in Chicago Crime Data

Don Ulfig

Warren Kress

Pamela Germain

Rohit Chundru

Wahhaj Farooq

**Table of Contents**

1. Abstract.............................................................................................................................................2
2. Introduction...............................................................................................................................……2
3. Methodology.....................................................................................................................................4

4. Analysis and Results.......................................................................................................................10

5. Conclusion......................................................................................................................................17

6 Future Work……………………………………………………………..………………………18

7. Appendix ........................................................................................................................................20

8. Reference........................................................................................................................................52

1. **Abstract**

The goal of our research was to find out how various factors can contribute to an arrest. Using logistic regression and publicly available crime data obtained from the Chicago Data Portal, we analyzed various factors such as date, type of crime, and location to see how they had an effect on whether an arrest was made or not. In our findings we saw that only 26.6% of crimes have led to an arrest. One trend we noticed overall was that the violent crime rate in Chicago has been gradually increasing over the years, while drug related crimes have been decreasing. Another observation we have noticed is that there are more crimes occurring in central Chicago during the summer months than any other region or time, with the most common crime being theft.

1. **Introduction**

Crime in the city of Chicago has historically been higher than the national average, particularly in terms of violent crime. Reported crime in Chicago rose consistently in the 1990s before seeing a drop in the early 2000s. This continued until a large spike in violent crime and murder began in 2016, despite the overall fall of reported crime in the country as a whole. The increase in homicides in 2016 was of such a magnitude that it increased the murder rate of entire nation. While there are many theories, from gang activity to uneven deployment of police officers forsaking more economically vulnerable areas, the root cause of these diverging trends of the past few years remains a mystery. Additionally, as crime has risen arrest and conviction rates have failed to keep pace. The stark truth is that it is easier to commit a crime without subsequent punishment now than in the past.

Arrests for drug related crimes have been declining significantly since the year 2000. Much of this is due to the easing of regulations regarding marijuana. In 2000 the Chicago Police Department made of 58,000 narcotic related arrests. By 2017 that number was down to less than 12,000. Marijuana related arrests once accounted for roughly half of all drug related arrests. In 2017, that portion had fallen to 28%. From a geographical standpoint, the reduction in narcotic related arrests has not been evenly distributed. The North side of the city has seen the most dramatic reduction regardless of neighborhood, while the south side has seen mixed results based on neighborhood, and the Central side has seen the least reduction overall.

Property crime is not the first thing that comes to mind when lawlessness in Chicago is discussed. However, crime including theft and destruction of private property have accounted for the majority of the transgressions in the city. Property crimes can range from misdemeanor to felony level offences and carry varying punishments. These crimes include but are not limited to burglary, arson, shoplifting, defacing public or private property, and even fraud and identity theft. Along with violent crime rates, the occurrence of crimes against property have been declining nationally for the past several years.

As violent crime, including shootings and homicides, have been increasing in Chicago, arrests, let alone convictions, had been dismally low. It is reported that only approximately 17% of Chicago’s homicide reports result in a suspect being taken into custody and charged with the crime. The arrest rate for shootings is even less, at approximately 5% according to a report by the Chicago Tribune. Causes for the low arrest rates of violent crimes are varied and multiple according to the police department, ranging from the sheer number of incidents in recent years in relation to manpower to a culture of community distrust of the police.

Given the varied nature of the evolution of crime over the past two decades in Chicago, it is clear that some of the factors that contribute to a report of a crime culminating in an arrest warranted a closer examination in order to determine what has the greatest impact on the outcome.

1. **Methodology**

3.1 Data

The data was obtained from the Chicago Police Department Data Portal located at data.cityofchicago.com under the category of Public Safety. The dataset contains reported incidents of crime with the exception of murder where data exists for the victim. The Chicago Police Department’s system is called CLEAR (Citizen Law Enforcement Analysis and Reporting). The data contains a total of 22 columns ranging in years from 2001 until the present minus 7-10 days. The fields are at case number and date of occurence row level with descriptives in each of the columns and no quantitative values. Each case is described with the primary type of crime category along with a more detailed description, a brief location description, such as a home, the sidewalk, a bar, the police district and beat number, latitude/longitude coordinates, with the provided deepest level of detail for locations at a block level to protect the privacy of the victims, a few additional informational codes, and two true/false columns to denote a domestic incident and if an arrest was made.

3.2 Data Preprocessing and Cleaning

The data used included many fields of which were different classifications on location and other crime descriptors. The method used to scrape the data involved creating a python script to pull the wanted fields. Due to how the data is recorded the data needed further inspection due to the Chicago Police Department allowing fields to contain commas which interfere with the SAS program. The fields were replaced with empty characters to compensate. Some of the fields contained empty values and when translated to SAS were not seen as NULL values, so those attributes were replaced with NULL for proper reading of the fields. The data fields our team sought appropriate for the analysis of how to predict the likelihood of the arrest are as follows Domestic, Date, Primary Type, Description, Location Description, Arrest, District, and Year. The years used for our analysis are 2001, 2009, and 2018 containing over one million data points. The description of each variable as well as there recoded values are as follows:

**Domestic - true/false.**

Domestic is a categorical variable which indicates whether or not the crime was a domestic crime or not. It is recoded to 0/1 with 1 indicating that it was a domestic related crime.

**District - North/South/Central.**

District is a categorical variable which indicates the police district the crime took place. Districts 11, 14, 15, 17, 19, 20, 24, 16, 25 fall unter North. Districts 4, 5, 6, 7, 22 fall under South. Districts 1, 2, 3, 8, 9, 10, 12, 13, 18 fall under central. North is coded as 0/1 with 1 indicating it is North. South is coded as 0/1 with 1 indicating that it is south. Central is coded as 0/1 with 1 indicating that it is central.

**Primary\_type - Violent/Drugs/Stealing.**

Primary\_type is a categorical variable which indicates the type of crime that was committed. Violent indicates crimes that fall under battery, sexual assault, homicide, kidnapping, and domestic violence. Drugs indicate crimes that fall under narcotics, and other narcotic violations. Stealing indicates crimes that fall under theft, burglary, robbery, and motor vehicle theft. Violent is coded as 0/1 with 1 indicating that it was a violent crime. Drugs is coded as 0/1 with 1 indicating that it is a drug related crime. Stealing is coded as 0/1 with 1 indicating that it is a stealing related crime. All other crimes a represented by the base case.

**Season - Winter/Spring/Summer.**

Season is a categorical variable which indicates the season when the crime took place. This is taken from the date where the months of December, January and February are categorized as Winter. The months of March, April and May are categorized as Spring and the months of June, July and August are categorized as Summer. Winter is coded as 0/1 with 1 indicating that the crime took place in Winter. Spring is coded as 0/1 with 1 indicated that the crime took place in the Spring. Summer is coded as 0/1 with 1 indicating that the crime took place in the Summer. If all variables appear as 0 that means the crime took place in Fall, which falls under the months of September, October, and November.

**Year - 2001/2009/2018.**

Year is a categorical variable which indicates the year the time took place. 2009 is coded under d\_2009 with possible values of 0/1, with 1 indicating that the crime took place in 2009. 2018 is coded under d\_2018 with possible values of 0/1, with 1 indicating that the crime took place in 2018. If both d\_2009 and d\_2018 are 0, then that means the crime took place in 2001.

**Location - Home/not home.**

Location is a categorical variable which indicates where the crime took place. Home falls under residence and apartment, everything else falls under not home. Home is coded as 0/1 with 1 indicating that the crime took place in a home.

**Arrest - True/False.**

Arrest is our dependent binary variable which indicates whether or not an arrest has been made. This is the dependent variable. Arrest is coded as 0/1 with 1 indicating that an arrest was made.

**HomVio - True/False**

HomVio is an interaction categorical variable between Home and Violent indicating whether or not a violent crime took place in and apartment or residence. If HomVio is equal to 1 then both Home and Violent are true, otherwise it is equal to 0.

3.3 Data Exploration

3.3a Descriptives

Between the dummy variables and the qualitative columns within the data set, the descriptives for mean, media, standard deviation, minimum and maximum do not provide much in the way of meaningful statistics. They are all categorical variables so descriptive statistics are not meaningful in our exploration.

3.3b Frequency

When looking at the frequency table for when an observation resulted in an arrest, where 0=false and 1=true, the majority of the reported incidents at 74% did not end with an arrest. When looking at the frequency table for when observation was recorded as domestic or not, where 0=false and 1=true, the majority of the reported incidents at 86% were not domestic related. (*Table 2*)

The frequency tables for the region locations (*Table 3*), where 0=false and 1=true, show the majority of the reported incidents at 40% happening in districts 1, 2, 3, 8, 9, 10, 12, 13, 18 defined as the Central location. This is followed by districts 11, 14, 15, 17, 19, 20, 24, 16, 25 defined as the North location as 36% and finally districts 4, 5, 6, 7, 22 defined as the South location at 24%.

When it comes to the seasons Winter, Spring and Summer, where 0=false and 1=true, it is Summer that accounts for the largest portion of reported incidents at 30%, followed by Spring at 25% and Winter at 17%. The remaining percentage was not flagged in a season we were tracking at the time of this report. (*Table 4*)

The type of crime was organized into three categories, violent, stealing/theft and drug related, where 0=false and 1=true. The frequency results show that theft has the highest percentage of reported incidents at 31%, followed by violent at 19% and drug related at 9%. The remaining percentage was not flagged for a type of crime we were tracking at the time of this report. (*Table 5*)

We tracked 3 years from the overall data set, 2001, 2009 and 2018. The frequency results (*Table 6*), where 0=false and 1=true, showed the years 2001 and 2009 having similar number of reported incidents at 37% each. The remaining 26% belongs to the year 2018.

The last variable in the list is in regards to the reported incident occurring at home, where 0=false and 1=true. The frequency table (*Table 7*) shows that the majority of incidents at 69% did not occur at home.

All of this data was then compared to the arrest factor in Figures 1-14. This allowed for a visual representation of the frequency data to be compared to the arrest factor to give more meaning. This meaning can be interpreted as the ratio of arrest based on the true and false of each independent variable.

The frequencies alone are only able to compare to the entire sample as opposed to dependent variable that we are analyzing. By forming the graphs the frequencies can be seen against the sample and against the dependent variable allowing a better descriptive look at arrests against different factors.

1. **Analysis and Results**

4.1 Techniques

Our dependent variable, arrest, has to possible values: true or false. When considering the different techniques to use we had the option between linear and logistic. Linear regression can be used when the dependent variable can have an infinite amount of possible outcomes, while logistic regression can be used when the dependent variable has a finite number of outcomes. Since arrest only has two possible outcomes, we have determined that a logistic regression technique would be the most appropriate approach when analyzing the dataset.

4.2 Transformation

The nature of the dataset was mostly qualitative data containing the values of 0 and 1. Therefore there was no need for any form of transformation since there are only two possible values for each variable. The adjustments made to the data sets was the cleaning of the data as well as creating the dummy variables for each and all categorical variables.

4.3 Assumptions

Logistic regression does require the same type of linear regression assumptions, such as residuals not needing to be normally distributed. What is required is the dependent variable must be binary, in our case if an arrest for the reported incident is true or false. In addition, the observations need to be independent of each other. Each reported incident in the data set is assigned an individual case number and there is no multicollinearity among the independent variables which can see seen in multiple steps of the model selection process where we look at the correlation between variables. Lastly, our sample size is large, well over 2,000 records and there are a significant amount in each cases where arrest = 1 and arrest = 0 as shown by the frequency tables.

4.4 Diagnostics

4.4a Outliers and Influential Points

Outliers and influential points were not a factor in the analysis, as all the variables in the set are descriptive and can only have either a 0 or 1 value. Running influence and iplots would show that there are no outliers or influential points since we are dealing with binary categorical variables in our model.

4.4b Collinearity

When fitting the model to the full dataset, there was multicollinearity between the geographic variables, North, South and Central variables depeicting each of those regions in Chicago. This was corrected during the model selection processes. After running the analysis on final model mode, it was determined there was no collinearity among our independent variables using the pearson correlation matrix post model selection methods and for our final model.

4.5 Model Training and Testing and Final Mode Selection

The crimes data set was split into training and test using a 75/25% split. Since we had a very large number of cases, we did not need to worry about having to give more cases to train the model. In order to select a final model that included all our significant predictors, we ran both forward selection and backward elimination processes. The results for both selection methods can be seen in Figure 16a-e and Figure 17a-f. In the forward selection process, we can see in Figure 17.b that we have an r-square value of 0.2700, an AIC of 726850.29, and SC of 727013.58. This method removes north variable as can be seen in Figure 17.c due to it being and insignificant variable in the model. In the backward elimination process we can see in Figure 16.b that we have an r-square value of 0.2700 an AIC of 726850.10 and SC of 727013.39. In this selection method, we can see in Figure 17.c that the only variable deemed insignificant in analysis is the central variable so it was removed.

Both forward selection and backward elimination have the same r-squared value, but AIC and SC values differ very slightly.. We looked at the number of variables in each model and it was 13 for both. All other measurements being the same, we decided to go with backward elimination. The resulting final model and its corresponding parameters can be seen in Figure 18a-e. The final model contains the following 13 independent variables to determine arrest: d\_domestic north south violent drug stealing summer winter spring d\_2009 d\_2018 home homvio, which can be seen in Figure18.c. Using the likelihood ratio in Figure 18.d, we can see that since its p value < 0.05, the null hypothesis can be rejected meaning an empty model is not greater at predicting than our current model. ALso all the p value for each individual variables are significant since all p values < 0.05. To recheck for multicollinearity, we can see in the correlation matrix in Figure 18.e, that no value is > 0.9, thus no collinearity. The most significant variable in our model is drug. The final model equation is as follows:

log(arrest=1/arrest=0) = -0.7510 -0.0806\*d\_domestic-0.0966\*north-0.1227\*south-0.3719\*violent+6.6793\*drug-1.0042\*stealing-0.0208\*summer+0.0411\*winter+0.0213\*spring-0.0668\*d\_2009-0.3305\*d\_2018-1.0743\*home+1.2405\*HomVio

4.6 Interaction Term

In our analysis, the interaction term we chose was the multiplicative product of home and violent (coded as homevio). The reason we chose this as our interaction term was because it had the greatest positive effect on our r-squared value.

4.7 Prediction

The data with the highest significance according to the results from Figure 18.c indicate what would be considered the significant predictors pertaining to the probability of arrest based upon our independent variables. From the final model, the strongest predictors would be Domestic, South, North, Drug, Violent, Stealing, Summer, Winter, Spring, 2009, 2018, Home, and HomVio. Based upon the final model equation:

log(arrest=1/arrest=0) = -0.7510 -0.0806\*d\_domestic-0.0966\*north-0.1227\*south-0.3719\*violent+6.6793\*drug-1.0042\*stealing-0.0208\*summer+0.0411\*winter+0.0213\*spring-0.0668\*d\_2009-0.3305\*d\_2018-1.0743\*home+1.2405\*HomVio

variables that contain a negative value would impact the likelihood of an arrest not happening and those with a positive value would impact the likelihood of an arrest occurring.

**Negative Influencers:**

* Domestic
* North
* South
* Violent
* Stealing (Strong)
* Summer
* 2009
* 2018
* Home (Strong)

**Positive Influencers:**

* Drug (Strong)
* Winter
* Spring
* HomeVio (Strong)

The strongest predictors would consist of those when isolated that result in a definite event happening. Among these predictors, we can conclude that for the negative influencers Stealing and Home would be events, when isolated, that guarantee an event occurring. For the positive influencers Drug and HomeVio, when isolated, would imply an arrest to always occur.

Among the predictors there are some that contain an almost neutral stance which can come as unclear when describing the outcome of particular events. These events would consist of outside factors that may be influencing, or describe the outcomes as being a toss up. When looking at the resultants from the equation as a whole, the closer the value is to 0 would imply the combination to result in an arrest not happening. The closer the value is to 1 would imply a higher likelihood of an arrest.

4.8 Validation

Validation was done using a train and test split of 75% to 25% of the data respectively. The data was initially split in the model selection stage, before we started to build our model. For validation, we performed 2 thresholding methods to determine how well our model works at predicting arrest. First, we ran the SAS code to determine phat (the predicted arrest value), lcl, and ucl. Table 8a-c shows a sample of the predicted output along with the confidence intervals. Once we obtained these values, we were able to compute a confusion matrix based on a standard threshold of 0.5 to determine if the predicted probability is larger than this value, and if so, then it will be considered as an arrest (d\_arrest = 1). Once all of these predicted arrest values are detected, we generated a classification matrix to determine that for the test set, the true positive is 26585 , false positive is 92, true negative is 209881, and false negatives is 49859, as shown in Table 9. Using these values, we calculated our performance metrics, including accuracy, specificity, precision, recall, and the f-measure in Table 12. Based on a threshold of 0.5, we calculated an accuracy of 0.8256, precision of 0.9966, recall of 0.3478, specificity of 0.996, and F-metric of 0.5156. Using this method, we can see that the model has an overall good accuracy at 82.56%. However, the recall tells us that when an actual arrest does occur, it doesn't detect it as an arrest with a value of 0.3478. On the other hand, both precision and specificity are near perfect meaning when a nn arrest occurs, the model can predict it quite well and the proportion of arrests that were detected are indeed arrests.

The other method for measuring the performance of our model is to validate the predictions by determining the best cutoff value for our threshold. To determine what the threshold would be, we first create a classification table containing probability levels that range from 0.2 to 0.6, incrementing by 0.05, to see which of these possible threshold yields the best performance. The threshold that had the highest sum of its sensitivity and specificity was 0.30, which can be seen in Table 10. This was selected since it maximizes the sum of the specificity and sensitivity and is larger than all the other possible summations. Now using this measured predicted probability, we can compare our predicted values to see which of the predictions, thus generating the confusion matrix as Table 11. The counts for true positive is 37507, false positive is 21624, true negative is 188349, and false negative is 38927. We can see that for this option, we have an accuracy of 0.7886, a precision of 0.6343, recall of 0.4906, specificity of 0.8970, and an F-metric of 0.5533, as seen in Table 12. Using this method, we can see that the accuracy for the overall model is 0.7886, which is relatively good. Based on the precision, we can see that of all the arrests our model detected, only about 63% were actually arrests, which is close to average, but worse than the first method. It is still able to detect quite well when non arrest occurs since it has a specificity of 0.89, still relatively high. The change we see is the recall where this method has a better rate of actually detecting when an actual arrest occurs.

Comparing the performance metrics outcomes in Table 12 between the two methods, we can see that using a fixed threshold of 0.5 scores higher in all metrics except for the recall and f-measure where method 1 has a recall of 0.3478 and f-measure of 0.5156, whereas method 2 has a recall of .4906 and f-metric of 0.5533. To determine which threshold works best for validating our model, we take into consideration the outcomes of the prediction. We can't our model to be able to best detect arrests based on the given independent variables and of these arrests they should actually be arrests since the alternative would be to falsely classify a person for arrest. Thus we want the model that has the highest accuracy in this regard, which would the model that is validated using the fixed 0.5 threshold. It maximizes the most of the performance metrics with the least amount of tradeoff.

1. **Conclusion**

Using SAS software and data from the Chicago police we were able to study and research various factors that lead to an arrest. Based on the research we have done, we can conclude that there is a pattern among our data. Based on our findings after analyzing the data, Winter and violent related crimes had the lowest arrest rate among any other predictors. We can also see that you are most likely to get arrested for a drug related crime in central Chicago committed during the Summer or Spring in 2009, while you are least likely to get arrested for a theft crime in south Chicago during the Summer or Spring in 2018. Statistically speaking we can see that the arrest rate has dropped significantly from 2009 to 2018 while at the same time, frequency of crime has also went down, as seen in Figures 5 and 6.

1. **Future Work**

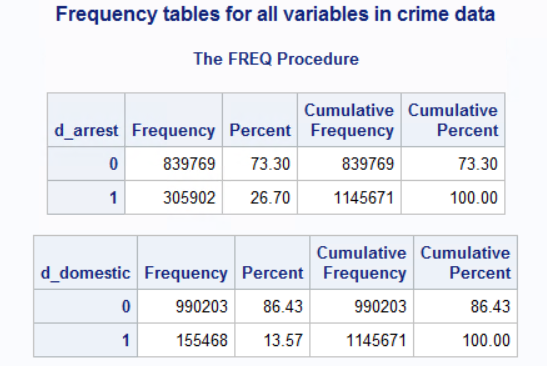
The nature of the dataset we have chosen contains mainly categorical data. Some of which, such as the Primary Type of crime, can be used as some form of scalar to give the data more meaning as opposed to dummy variables. By molding the data in this format it would allow for more representations based upon type of crime and would allow that comparison to be visual as well as numerical.

Since the data has many variations of combinations, all of which are considerable in regards to the outcome of an arrest, there are many variables which can be broken up into separate datasets which can yield a more accurate result. The years can be broken apart and analyzed against one another to acquire a different perspective on what crimes were occurring and if said crimes resulted in more or less arrests. Different years have access to different technology and methods which can influence whether a criminal gets away or is arrested for whatever crime is committed.

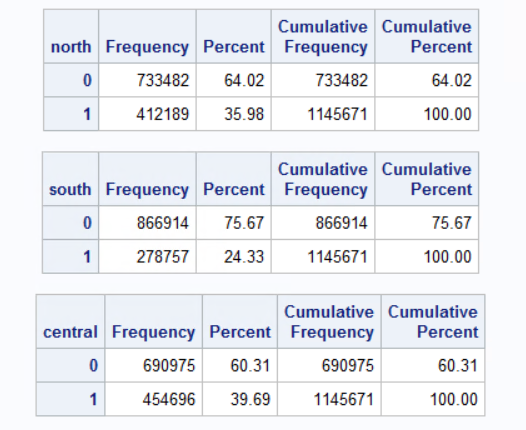
Adding data that is related to the population of the area in which a crime occurs such as census, financial and educational data. This data would give more reasoning as to why these crimes may occur more frequently in particular areas. The development of more data would allow for more patterns to be analyzed such as low income areas, low education areas, and population demographic to be considered as well as the type of crime that is committed.

Aggregation at date levels would lend itself to seasonal patterns throughout the year, or at a time level, patterns throughout the hours in the day. Different times of day may yield different types of crimes and different frequencies at which these crimes occur. The addition would allow for further comparisons that may prove to be influential on crime outcomes.

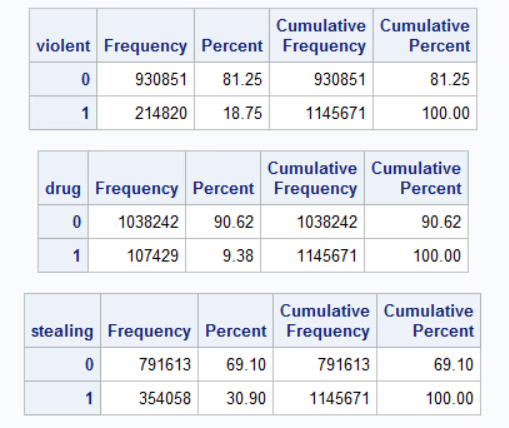
1. **Appendix**

****

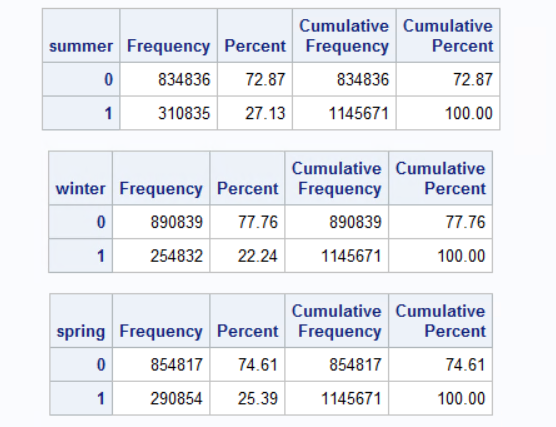
*Table 1: Number of cases for d\_arrest and d\_domestic in each class*

****

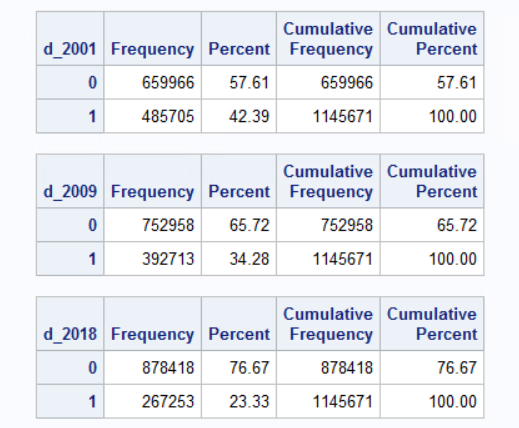
*Table 2: Number of cases for north, south, and central in each class*

****

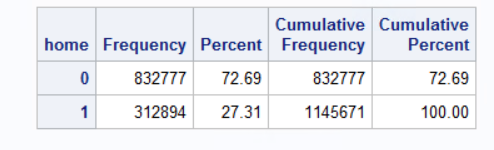
*Table 3: Number of cases for violent, drug, and stealing in each class*

****

*Table 4: Number of cases for summer, winter, and spring in each class*

****

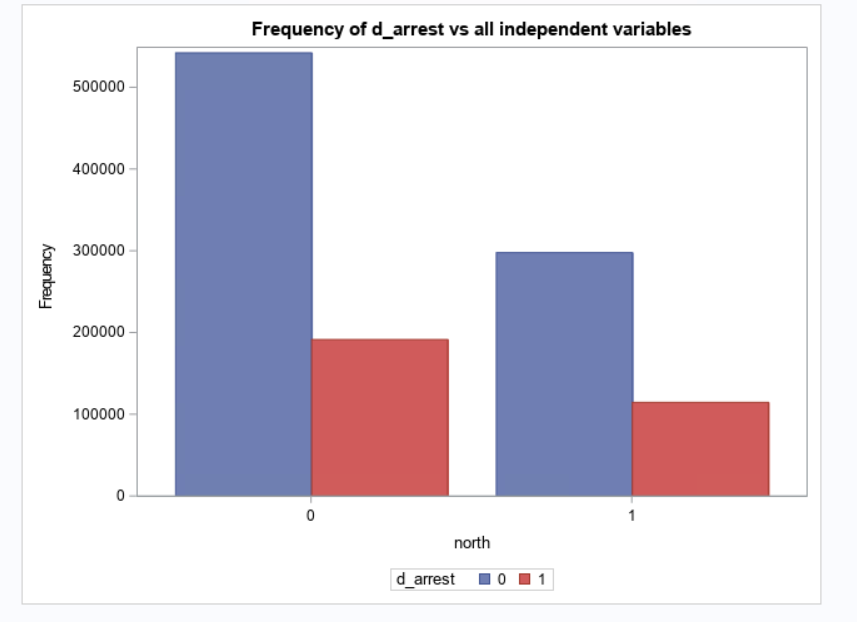
*Table 5: Number of cases for d\_2001, d\_2008, and d\_2018 in each class*

****

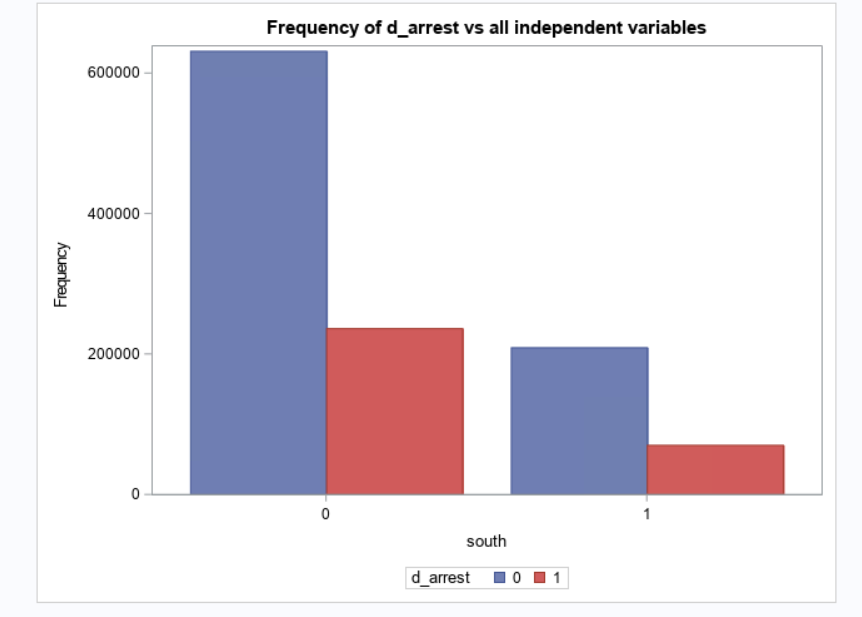
*Table 6: Number of cases for home in each class*

****

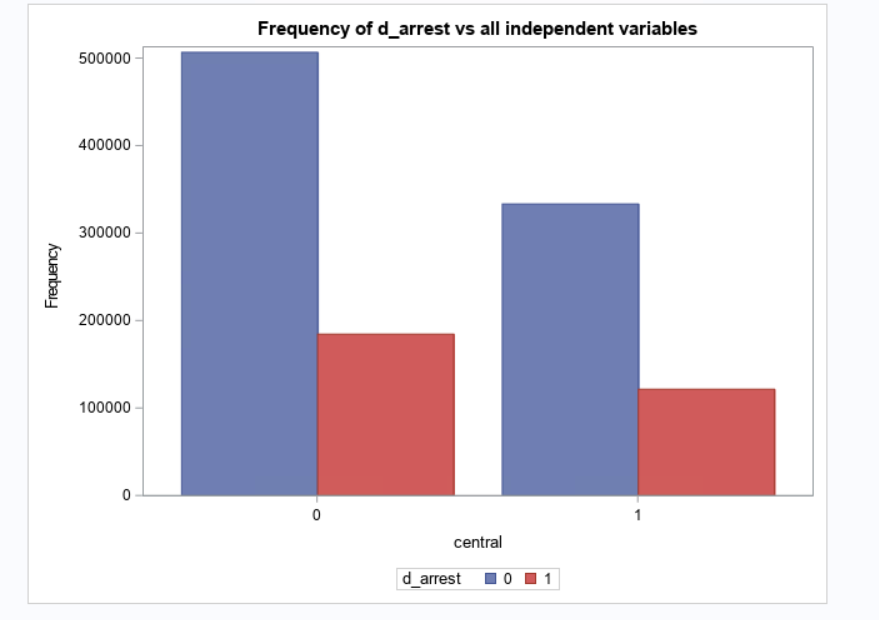
*Figure 1: Frequency of d\_domestic by d\_arrest*

****

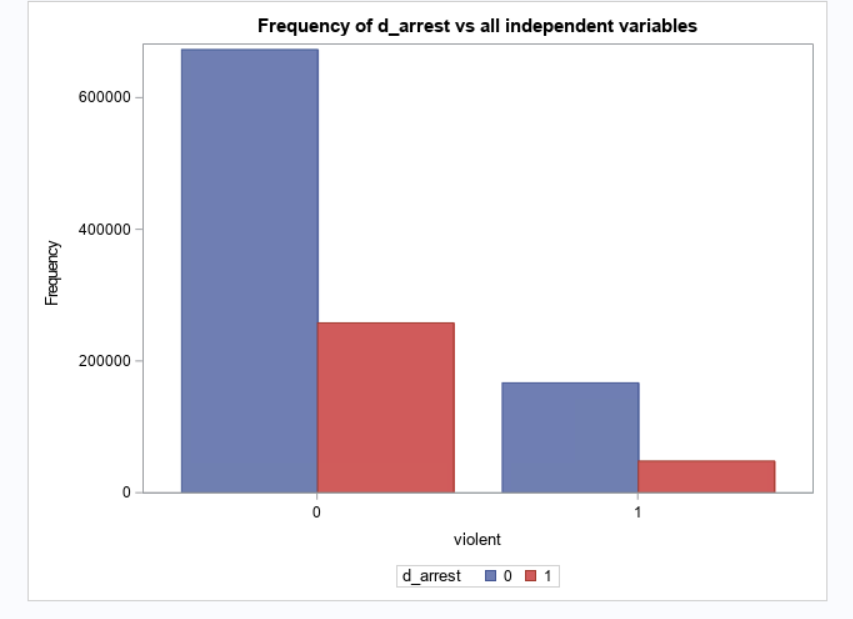
*Figure 2: Frequency of north by d\_arrest*

****

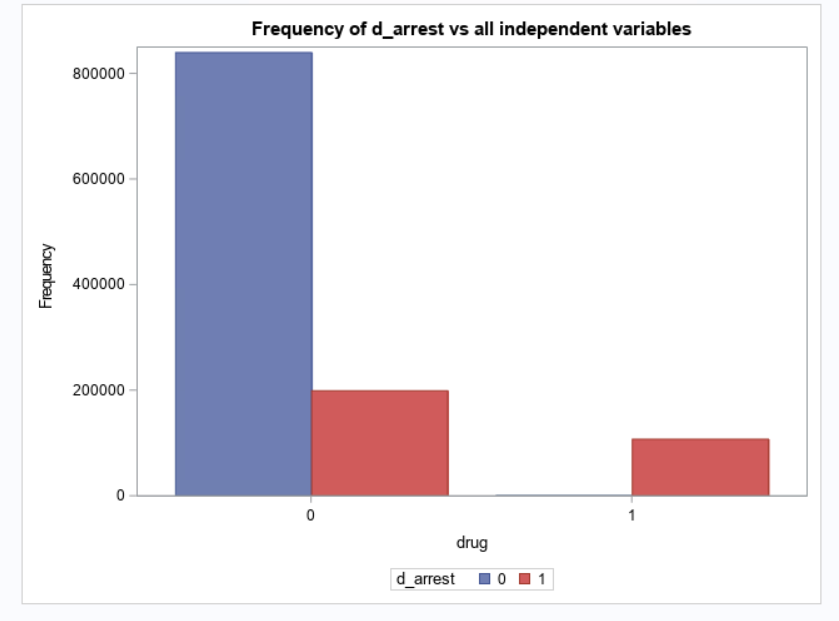
*Figure 3: Frequency of south by d\_arrest*

****

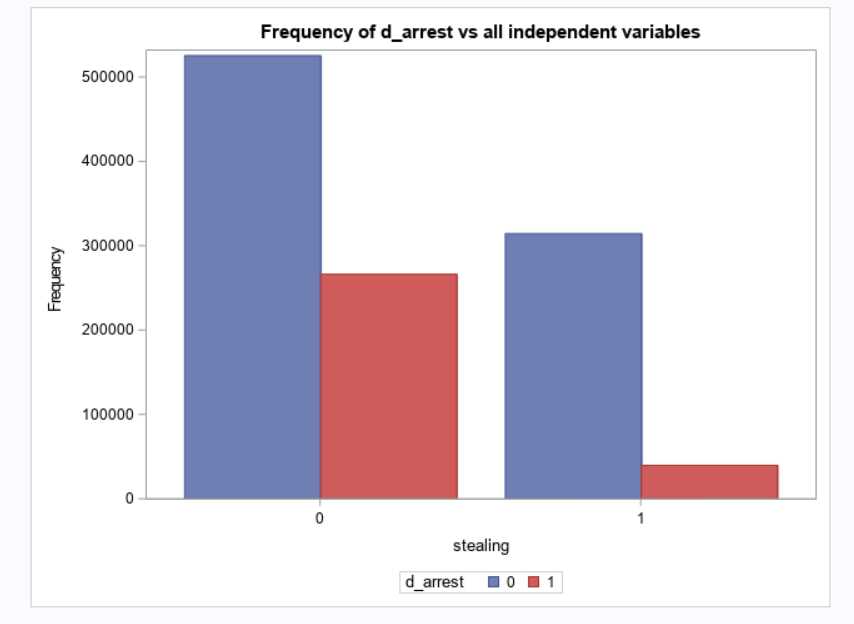
*Figure 4: Frequency of central by d\_arrest*

****

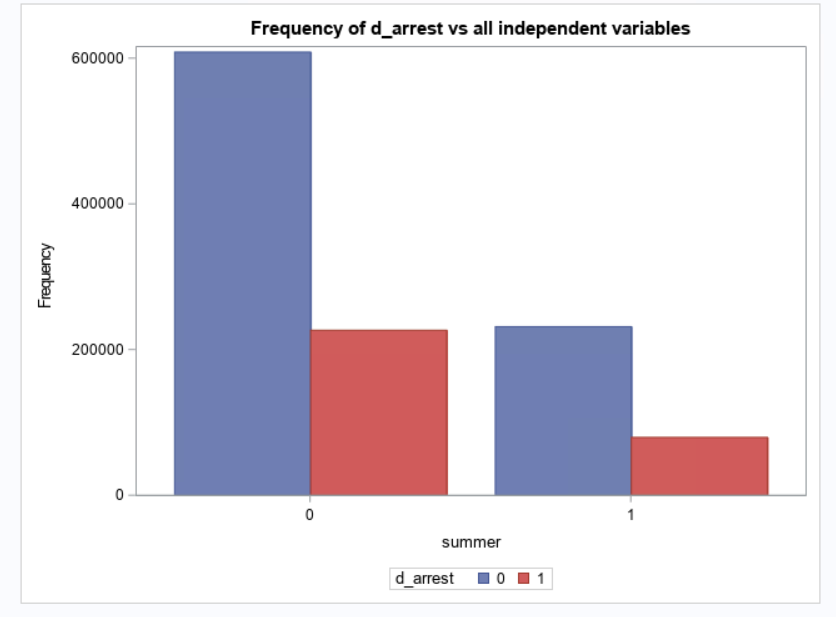
*Figure 5: Frequency of violent by d\_arrest*

****

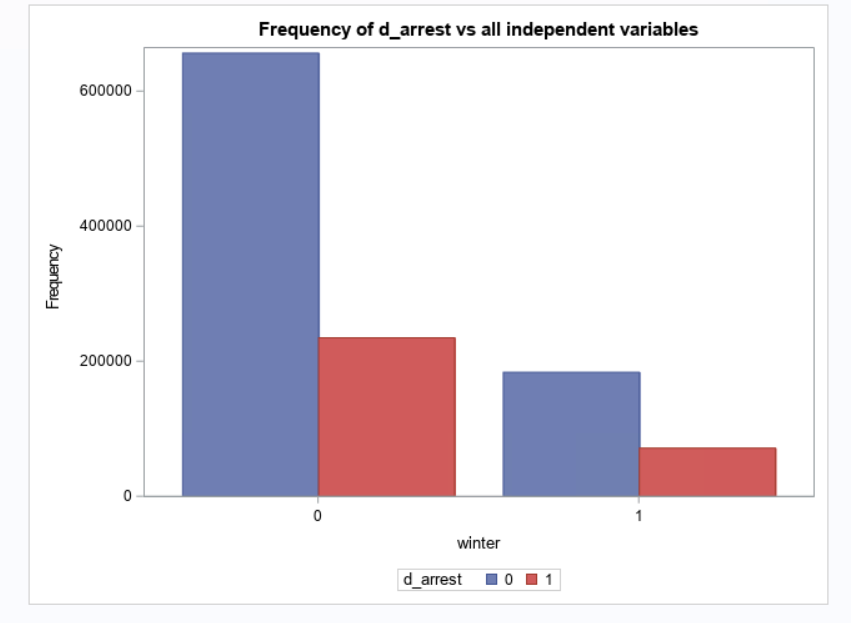
*Figure 6: Frequency of drug by d\_arrest*

****

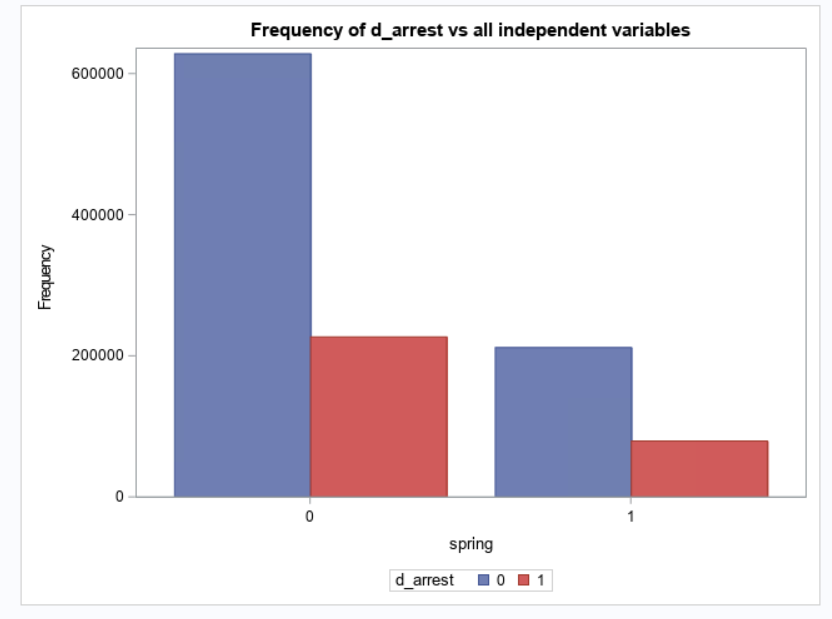
*Figure 7: Frequency of stealing by d\_arrest*

****

*Figure 8: Frequency of summer by d\_arrest*

****

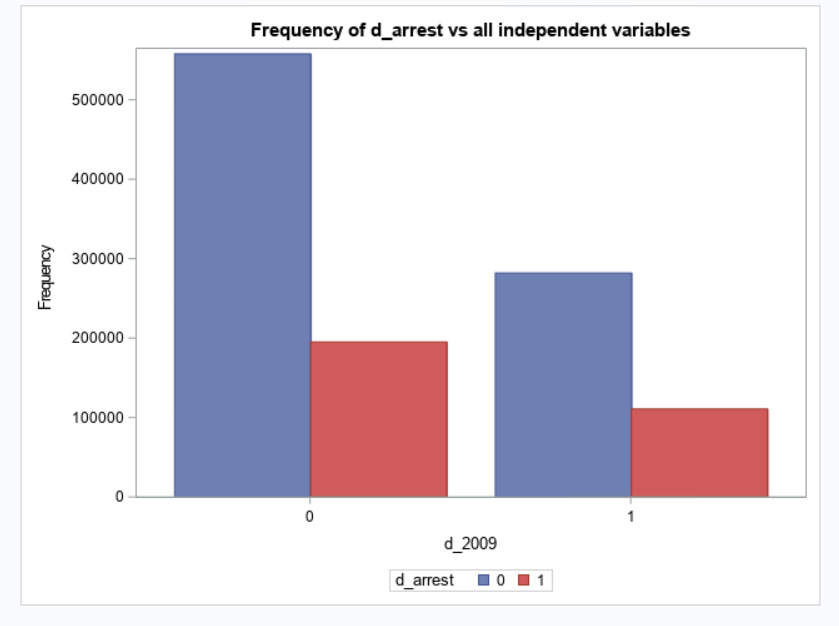
*Figure 9: Frequency of winter by d\_arrest*

****

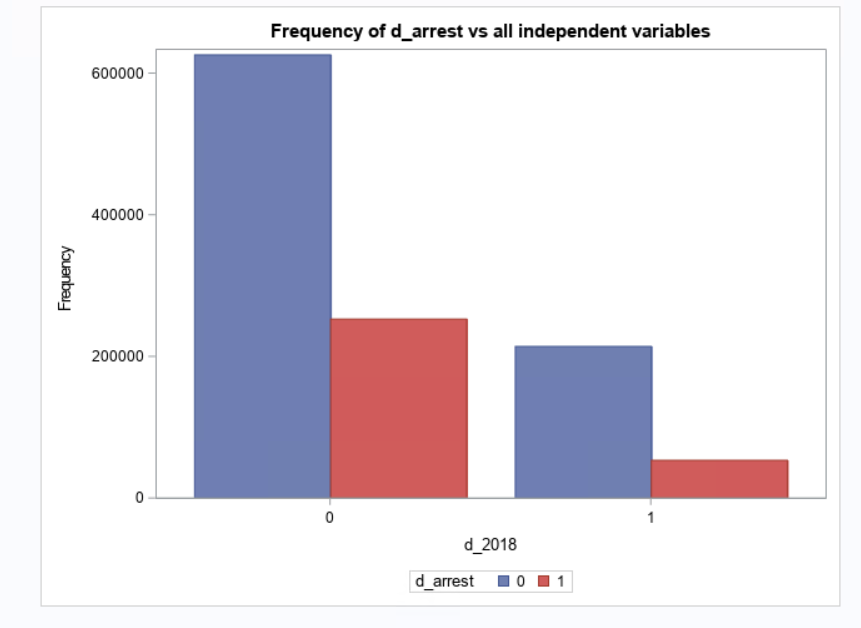
*Figure 10: Frequency of spring by d\_arrest*

****

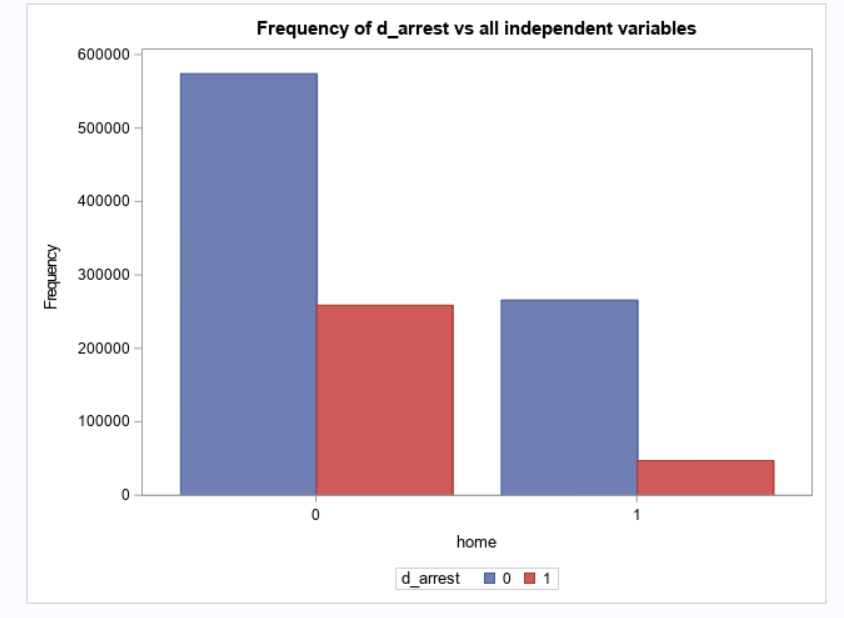
*Figure 11: Frequency of d\_2001 by d\_arrest*

****

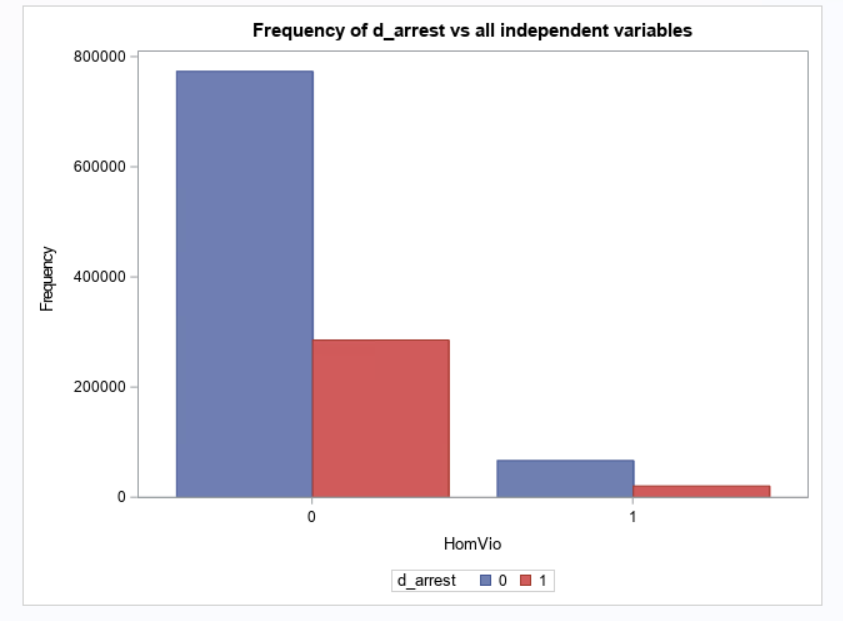
*Figure 12: Frequency of d\_2009 by d\_arrest*

****

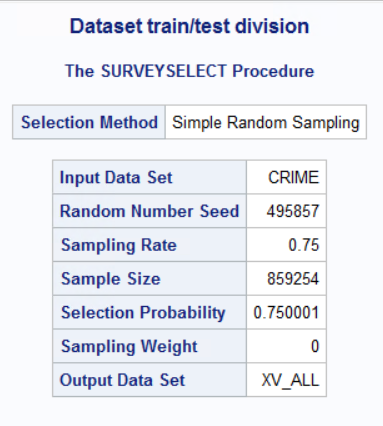
*Figure 13: Frequency of d\_2018 by d\_arrest*

****

*Figure 14: Frequency of home by d\_arrest*

****

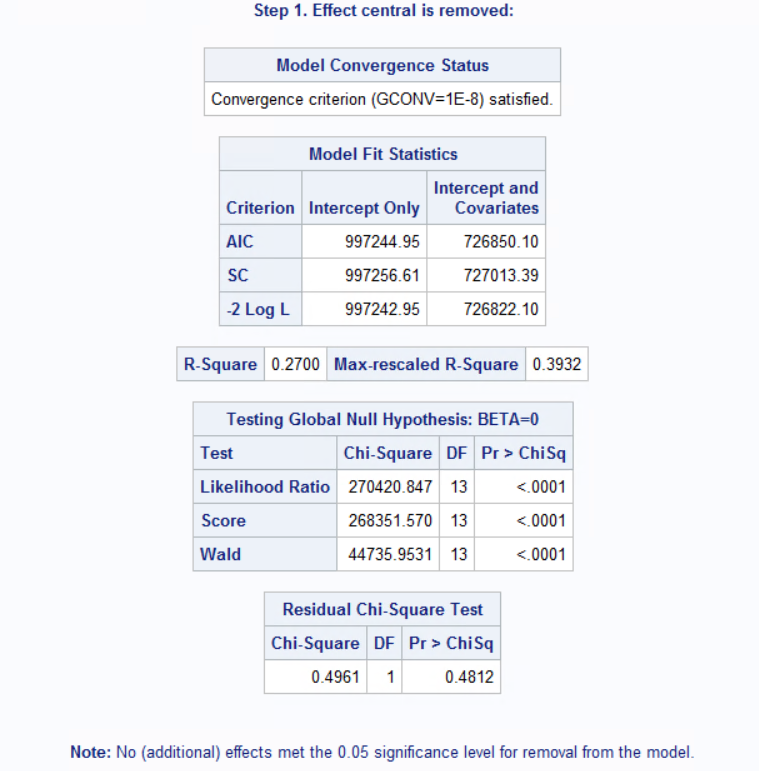
*Figure 15: Frequency of HomeVio by d\_arrest*

****

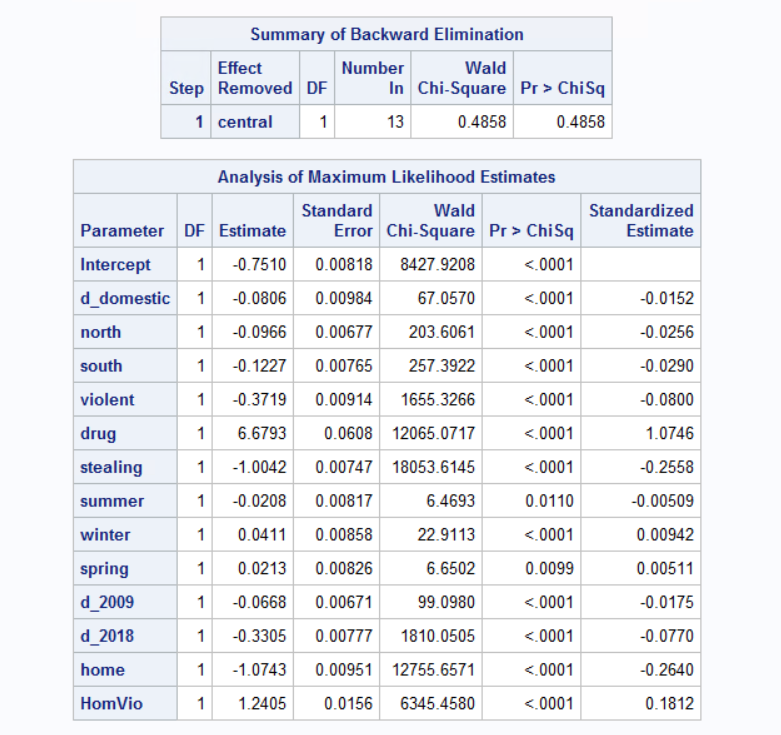
*Table 7: Crime data split on 75/25 Train/test splits*

****

*Figure 16.a: Backward selection model information*

****

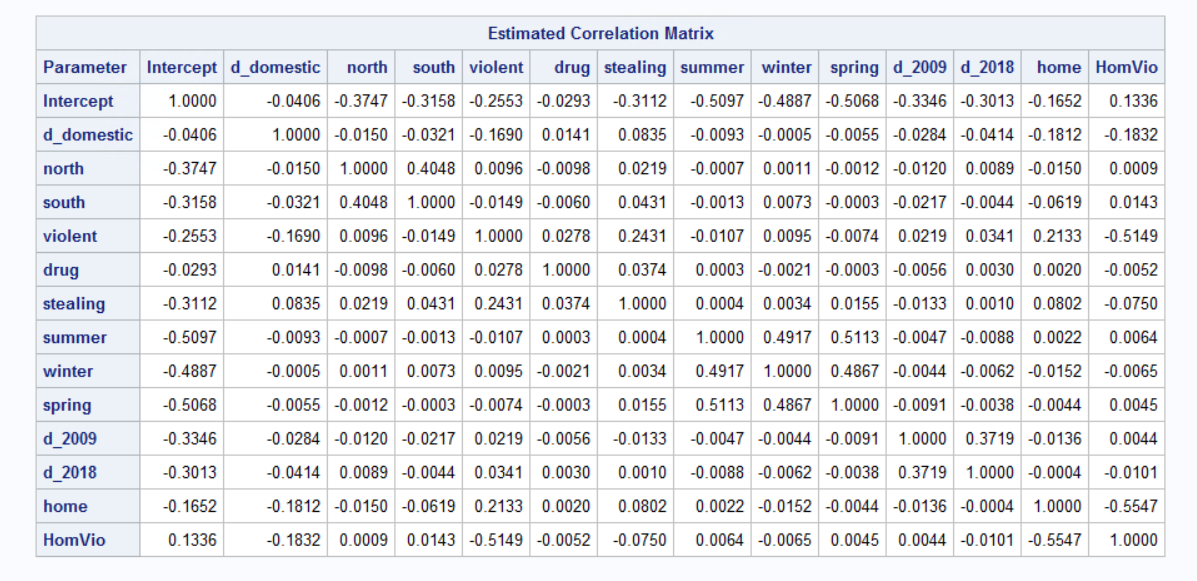
*Figure 16.b: Backward selection model statistics, r-square and chi-square charts*

****

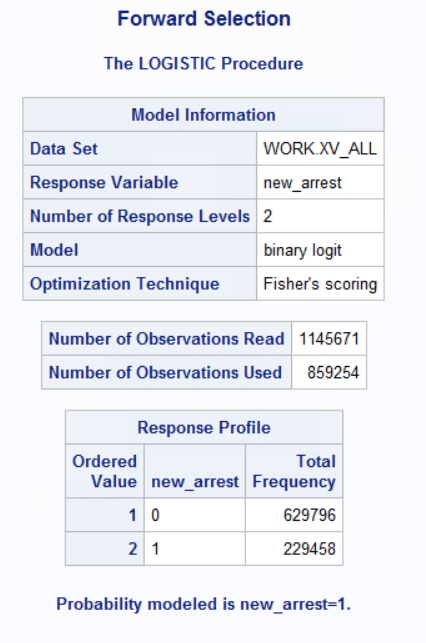
*Figure 16.c: Backward selection model summary and MLE*

****

*Figure 16.d: Backward selection model odds ratios estimates and predicted probability charts*

****

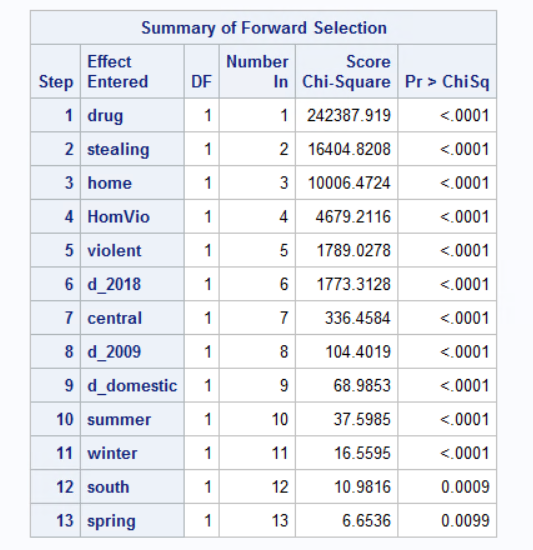
*Figure 16.e: Backward selection model correlation matrix*

****

*Figure 17.a: Forward selection model information*

****

*Figure 17.b: Forward selection model statistics, r-square, and chi-square tables*

****

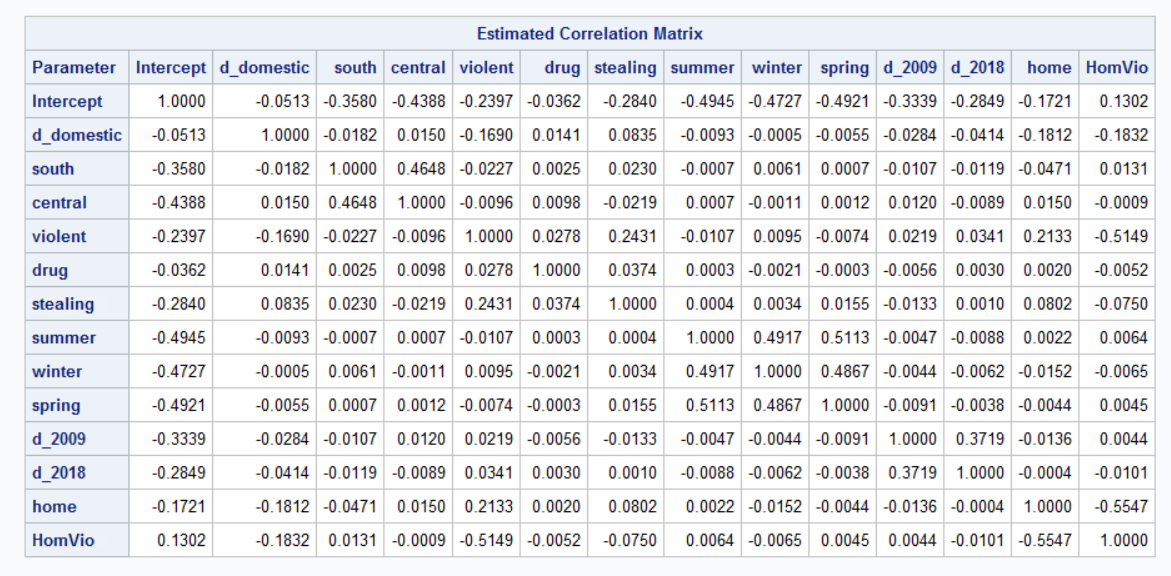
*Figure 17.c: Forward selection model summary table*

****

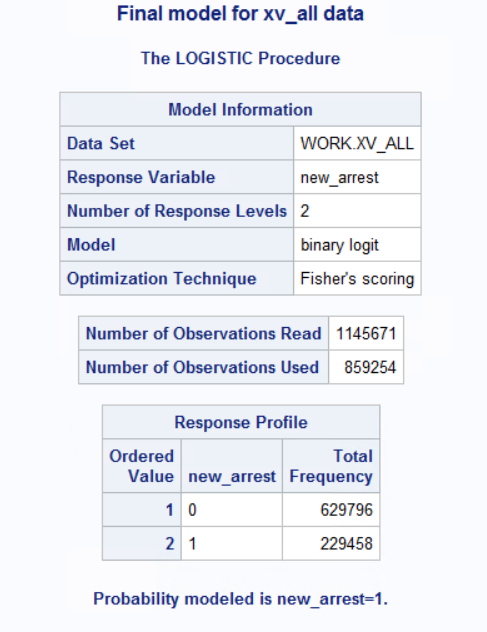
*Figure 17.d: Forward selection model MLE chart*

****

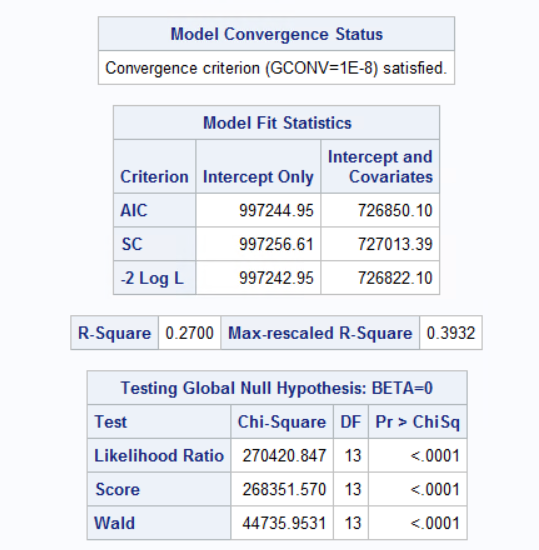
*Figure 17.e: Forward selection model odds ratio estimates and predicted prob. Charts*

****

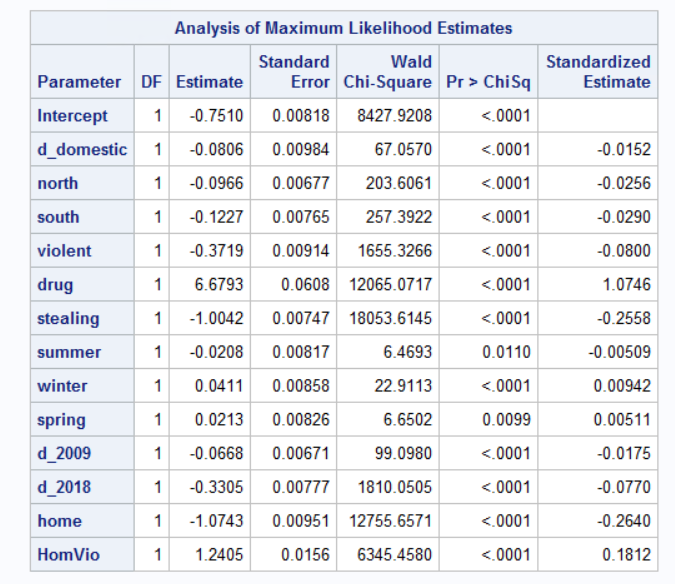
*Figure 17.f: Forward selection model correlation matrix*

****

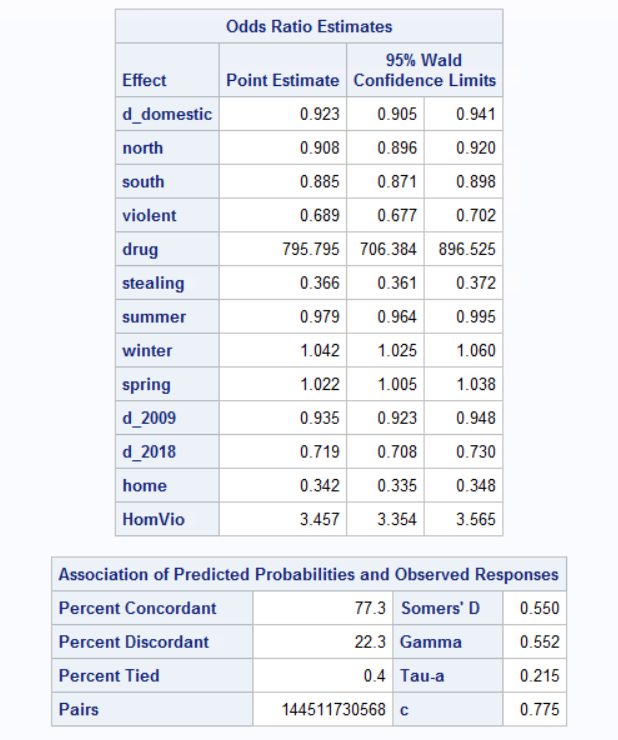
*Figure 18.a: Final model information*

****

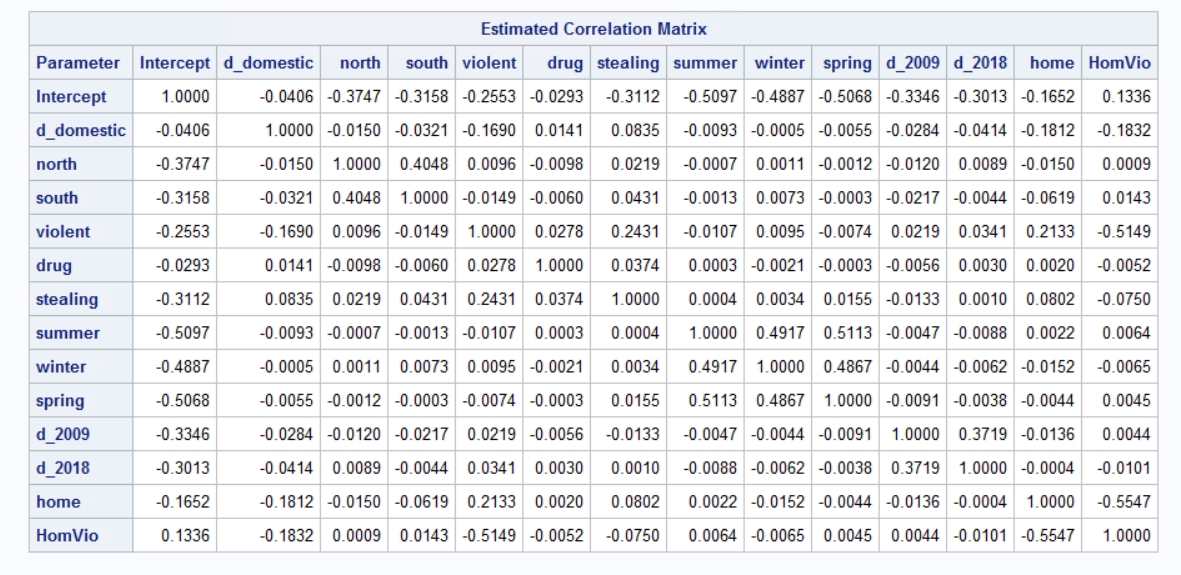
*Figure 18.b: Final model statistics, r-square, and chi-square charts*

****

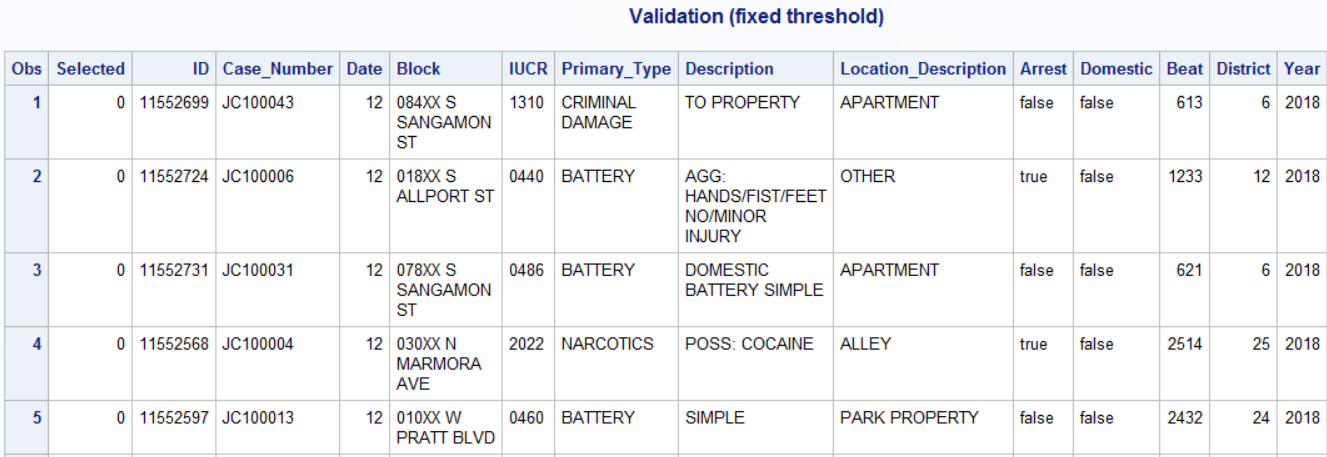
*Figure 18.c: Final Model MLE table*

****

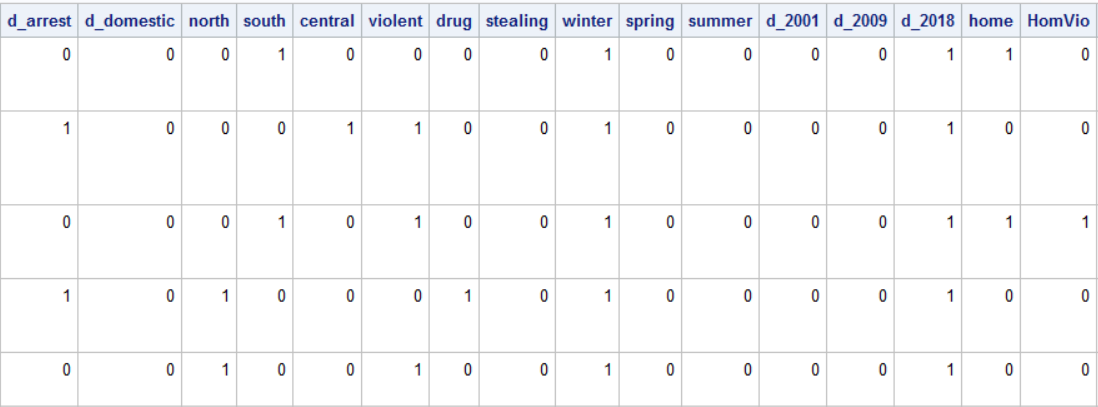
*Figure 18.d: Final Model odds ratio estimates and predicted probability charts*

****

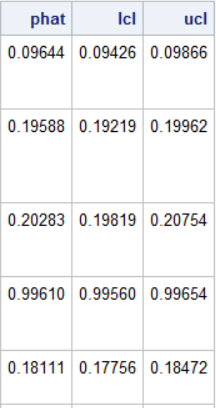
*Figure 18.e: Final model correlation matrix*

****

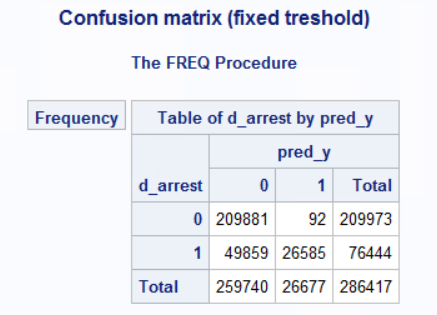
*Table 8.a: Test data set with predicted d\_arrest and confidence interval (5 observations)*

****

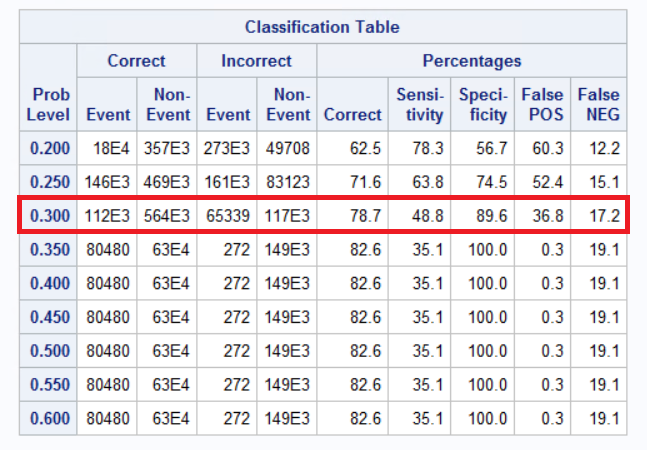
*Table 8.b: Test data set with predicted d\_arrest and confidence interval (5 observations)*

****

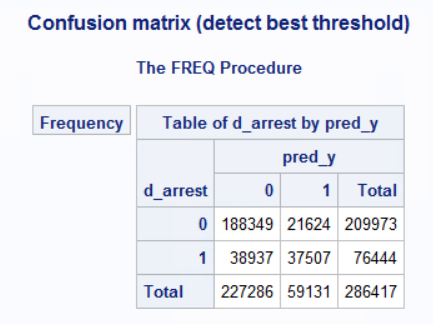
*Table 8.c: Test data set with predicted d\_arrest and confidence interval (5 observations)*

****

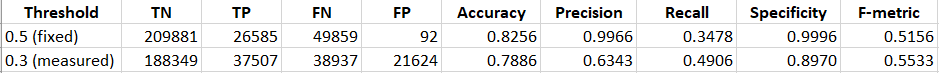
*Table 9: Confusion matrix for test data using fixed threshold of 0.5*

****

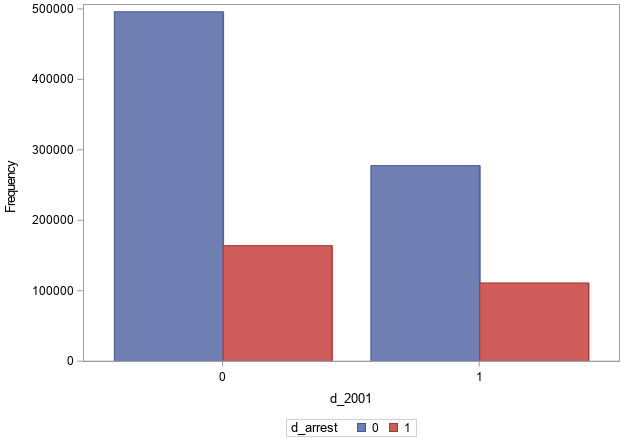
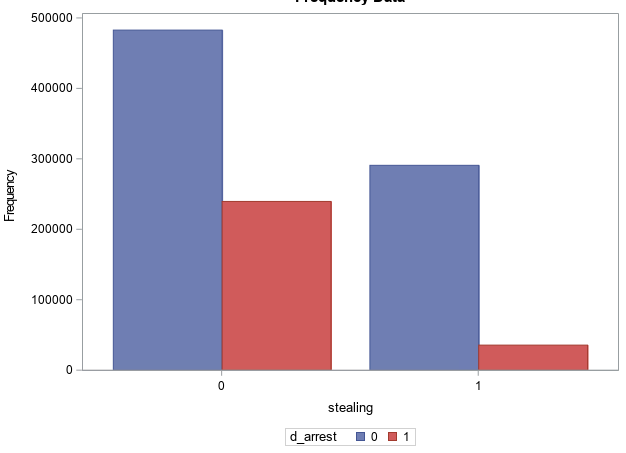
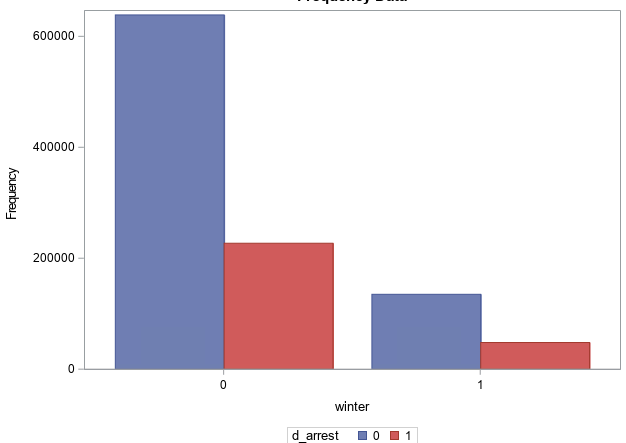
*Table 10: Classification table for predicted probabilities from 0.2 to 0.6 with best threshold value in red*

****

*Table 11: Confusion matrix for test data using best threshold of 0.3*

****

*Table 12: Performance metric for both fixed and best measured threshold value*

**Code**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Data Importing

Contains steps to import data set as crime, comma

seperated.

Source: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

proc import datafile='S:\Final Project/Final Project\crimes.csv' out = crime replace;

delimiter = ',';

getnames=yes;

run;

\*print current view of data set, first 50 obs shown;

proc print data=crime (obs=50);

run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Data Preprocessing/Cleaning

This section contains the code to create dummy variables.

We convert certain variables into relevant vars pertinent

to our analysis and remove ones we don't use.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* create dummy variables for arrest, where if arrest is now 1 if true;

\* and domestic is 1 if true;

data crime;

set crime;

\*arrest dummy;

d\_arrest = 0;

if arrest = 'true' then d\_arrest = 1;

\*domestic dummy;

d\_domestic = 0;

if domestic ='true' then d\_domestic = 1;

\*districts based on https://news.wttw.com/sites/default/files/Map%20of%20Chicago%20Police%20Districts%20and%20Beats.pdf;

\*combine chicago districts into three distinct regions: north, south, and central;

north = 0;

south = 0;

central = 0;

if district in (11 14 15 17 19 20 24 16 25) then north = 1;

if district in (4 5 6 7 22) then south = 1;

if district in (1 2 3 8 9 10 12 13 18) then central = 1;

\*primary\_type of arret we are focusing on by grouping types in 3 cats, violent, drug, or stealing crimes;

violent=0;

drug=0;

stealing = 0;

\* primary type is the kind of arrest made and this section combines similar crimes into 3 categories: violent, drug, and stealing;

if Primary\_Type in ('BATTERY', 'CRIM SEXUAL ASSAULT', 'HOMICIDE', 'KIDNAPPING', 'DOMESTIC VIOLENCE') then violent = 1;

if Primary\_Type in ('NARCOTICS', 'OTHER NARCOTIC VIOLATION') then drug = 1;

if Primary\_Type in ('THEFT', 'BURGLARY' 'ROBBERY' 'MOTOR VEHICLE THEFT') then stealing = 1;

\*season dummy;

\*here we convert our time information into yearly seasons to be more general. We use winter, spring, and summer;

winter=0;

spring=0;

summer=0;

if date in (1 2 12) then winter=1;

if date in (3 4 5) then spring = 1;

if date in (6 7 8) then summer =1;

\*year dummy;

\*we focus on the years of 2001, 2009 and 2018;

d\_2001 = 0;

d\_2009 = 0;

d\_2018 = 0;

if year = 2001 then d\_2001 = 1;

if year = 2009 then d\_2009 = 1;

if year = 2018 then d\_2018 = 1;

\*location dummy - set home to equal crimes where it occurs at homes as in

\* residence or an apartment;

home = 0;

if location\_description = 'RESIDENCE' or location\_description = 'APARTMENT' then home =1;

\*interaction variable - combination of type of location (home) and violent crimes;

HomVio= home \* violent;

run;

\*prints cleaned and processed data set (100 obs);

title "Crime data set post processing";

proc print data = crime (obs=100);

run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Data exploration step

The code in this section is used to see the frequency table for each variable in

the data set

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* create freq table for each var;

Title "Frequency tables for all variables in crime data";

proc freq data=crime;

tables d\_arrest d\_domestic north south central violent drug stealing summer winter spring d\_2001 d\_2009 d\_2018 home;

run;

\*The following code below is used to plot the frequency graphs;

\*for each independent variable agaisnt the binary d\_arrest values;

title "Frequency of d\_arrest vs all independent variables";

\* domestic vs arrest;

proc sgplot data=crime;

vbar d\_domestic/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\* north vs arrest;

proc sgplot data=crime;

vbar north/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\* south vs arrest;

proc sgplot data=crime;

vbar south/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\* central vs arrest;

proc sgplot data=crime;

vbar central/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\* violent vs arrest;

proc sgplot data=crime;

vbar violent/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\* drug vs arrest;

proc sgplot data=crime;

vbar drug/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\* stealing vs arrest;

proc sgplot data=crime;

vbar stealing/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\* summer vs arrest;

proc sgplot data=crime;

vbar summer/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\* winter vs arrest;

proc sgplot data=crime;

vbar winter/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\* spring vs arrest;

proc sgplot data=crime;

vbar spring/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\* d\_2001 vs arrest;

proc sgplot data=crime;

vbar d\_2001/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\* d\_2009 vs arrest;

proc sgplot data=crime;

vbar d\_2009/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\* d\_2018 vs arrest;

proc sgplot data=crime;

vbar d\_2018/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\* home vs arrest;

proc sgplot data=crime;

vbar home/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\* HomVio vs arrest;

proc sgplot data=crime;

vbar HomVio/group=d\_arrest GROUPDISPLAY = CLUSTER;

run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Data Train/Test Split

\*Divide Data in to train/test set using 75/25 split

\* and save into xv\_all dataset

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

title 'Dataset train/test division';

proc surveyselect data = crime

out= xv\_all seed=495857

samprate = 0.75 outall;

run;

\* display 100 obs after splitting;

proc print data =xv\_all (obs=100);

run;

\* set new\_arrest as train\_y variable for the train data;

data xv\_all;

set xv\_all;

if selected then new\_arrest=d\_arrest;

run;

\* display 20 obs after making train y var;

title "Train/test split data with ney\_y (new\_arrest)";

proc print data=xv\_all (obs=20);

run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Model Selection

The code in this section is used to determine the final model. We start with creating

a full model and using forward and backward selection techniques, to determine

the best possible model to fit our data

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*full model;

proc logistic data = xv\_all;

title 'Full Model for Crime data';

model new\_arrest (event='1')= d\_domestic north south central violent drug stealing summer winter spring d\_2009 d\_2018 home homvio/stb corrb rsquare;

run;

\*Model Selection and comparison;

\*Backward selection;

proc logistic data = xv\_all;

title 'Backward Selection';

model new\_arrest (event='1')= d\_domestic north south central violent drug stealing summer winter spring d\_2009 d\_2018 home homvio/

stb corrb rsquare selection = backward;

run;

\*Forwards selection;

proc logistic data = xv\_all;

title 'Forward Selection';

model new\_arrest (event='1')= d\_domestic north south central violent drug stealing summer winter spring d\_2009 d\_2018 home homvio/

stb corrb rsquare selection=forward;

run;

\*both produce the same model;

\*final model with most signiicant variables;

proc logistic data = xv\_all;

title 'Final model for xv\_all data';

model new\_arrest (event='1')= d\_domestic north south violent drug stealing summer winter spring d\_2009 d\_2018 home homvio/

stb corrb rsquare;

run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Model Validation

\*The code in this sectio is used for validation via 2 options: 0.5 threshold vs

finding best threshold sing predicted probailites.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*Calculate phat along with its intervals and create pred data for test data to determine;

\* if pred\_y for test is > 0.5;

proc logistic data = xv\_all;

title 'Validation (fixed threshold)';

model new\_arrest (event = '1') = d\_domestic north central violent drug stealing summer winter spring d\_2009 d\_2018 home homvio;

output out=pred(where=(new\_arrest=.)) p=phat lower=lcl upper=ucl

predprob=(individual);

run;

\*print 10 obs of pred data;

proc print data=pred (obs=10);

run;

\* calculats classification results for each obs using 0.5 threshold;

title "Confusion matrix (fixed treshold)";

data probs;

set pred;

pred\_y=0;

threshold=0.5;

if phat>threshold then pred\_y=1;

run;

\* generate confusion matrix for option 2;

proc freq data = probs;

tables d\_arrest\*pred\_y/norow nocol nopercent;

run;

\*Calculate phat along with its intervals and create pred data;

\* and generates prob levels from 0.2 to 0.6 so we can find best;

\* threshold value;

proc logistic data = xv\_all;

title 'Validation (detect best threshold)';

model new\_arrest (event = '1') = d\_domestic north central violent drug stealing summer winter spring d\_2009 d\_2018 home homvio/

ctable pprob= (0.2 to 0.6 by 0.05);

output out=pred(where=(new\_arrest=.)) p= phat lower=lcl upper=ucl predprob=(individual);

run;

\* computes predicted y for test set based on best threshold (0.25);

title "Confusion matrix (detect best threshold)";

data probs;

set pred;

pred\_y=0;

threshold=0.30;

if phat>threshold then pred\_y=1;

run;

\* generate confusion matrix for best threshold (option 3);

proc freq data = probs;

tables d\_arrest\*pred\_y/norow nocol nopercent;

run;

1. **Reference**

Chicago Data Portal. Crimes - 2001 to present (2019). Retrieved from <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>

Map of Chicago Police Districts and Beats - March 2012. Retrieved from

<https://news.wttw.com/sites/default/files/Map%20of%20Chicago%20Police%20Districts%20and%20Beats.pdf>