Perform Analysis and Document Finding

# Data

## Exclusions (Row Reduction)

Before modeling, we excluded records that would introduce potential noise into our analysis. Our initial dataset included 261,254 records of incidents reported by CMPD from 2012 to 2014. First, given our focus is on predicting clear rates for crimes, we removed non-crimes like Missing Persons and Suicides that are labelled as “Non-Crimes” according to the NIBRS official classification[[1]](#footnote-1). As shown in Figure 1, we removed 25,992 non-crime records. After removing these incidents, the total number of incidents each year drops about 8-12% to 79,730 in 2012, 76,696 in 2013 and 78,836 in 2014 (see Appendix 4.1). We then removed 30,593 records that were cleared due to being reported as unfounded or miscellaneous clear status. This included crimes where the victim did not prosecute or District Attorney opted not to press charges. Appendix 4.2 details the list of clearance statuses that were under this category. Last, since our focus is on CMPD only, we removed 1,367 crimes (< 1%) that were committed relayed to Police units outside of CMPD. Most of these crimes were sent to other local units like Matthews, Pineville or Mint Hill. After removing non-Charlotte crimes, our final model dataset includes 203,302.

|  |  |  |
| --- | --- | --- |
| **Step in Preparing Model Dataset** | **Change** | **Records** |
| Starting Population: Original Dataset |  | 261,254 |
| Remove Non-Crimes | -25,992 | 235,262 |
| Remove Unfound and Misc. Clear Status | -30,593 | 204,669 |
| Remove Non-CLT Crimes (e.g. Matthews) | -1,367 | 203,302 |
| **Final Model Dataset** |  | **203,302** |

*Figure 1: Data Set Exclusions*

## Feature Generation

Our next step is feature generation to add additional fields to our data set. First, we created a list of variables based on several hypotheses (see 2nd report paper for details). Most of the additional variables were created through transformation of our original dataset. However, some variables used appended data like neighborhood demographic variables from the 2012 Charlotte Quality of Life Survey.[[2]](#footnote-2) In total, 35 variables were created from either the original data while 15 variables were created with external data (10 demographic variables, 2 extreme weather day flags and 3 proximity flags for crimes committed within 500 feet of homeless shelters, churches or schools).

Figure 2: Predictor Variables by Category

Figure 2 above categorizes the 50 variables considered into broad categories of variables. Appendix 4.4 provides more details of the 50 variables considered and a description for each variable.

## Dependent variable

We set up our problem as a binary classification problem to predict whether crimes are “cleared” (i.e. result in an arrest) or not, i.e. remain open and have not resulted in an arrest. A column (Clear\_Flag) was created where a 1 was assigned if the incident was cleared, 0 if the incident was still open at the time of the original dataset generation.

## Variable Importance (Filter Method)

Before running our models, we evaluated on a “filter” basis the variable of importance of each predictor using a statistical approach (Chi-Square). We chose Chi-Square given than nearly all of the variables were categorical and given that all variables were originally screened to ensure that they aligned to one of our hypotheses. However, as we explain later, most of our methods (like GBM and GLM with regularization) have their own wrapper based feature selection algorithms that will further refine the list of variables.

Figure 3: Variable Importance

Figure 3 shows the top 20 variables ranked from the highest Chi-Square value. The top variables are primarily related to the type of crime (NIBRS\_Hi\_Class, Category, Against). In addition, several additional variables came in related to victim (PUBLIC\_FLAG, NCSTATE\_FLAG), property value (RankPropertyValue) and the amount of recent crime activity (RollSevenDayNorm).

## Dataset partition (training, validation, test)

We split our final model data set three ways, 70% was allocated for training the models, 20% for validation and 10% for testing. Each model is built and tuned using the training and validation datasets. The models were not run against the test until they were completed. The test dataset is to be used to evaluate performance across each model.

## Model Performance Metric

Given that our problem is binary classification, we selected Accuracy (true results divided by total incidents) and Area under the Curve (AUC) to be our two main model performance metrics for each model. First, we selected Accuracy because we lack the domain expertise to weigh the cost-benefit of false negatives versus false positives in this context of cleared crimes. Therefore, for simplicity purposes, we assume that both false negatives and false positives have the same cost/benefit structure and accordingly, accuracy would be a sufficient measurement rather than sensitivity or specificity. We also want to consider AUC as it is a more holistic measurement of ROC performance of each model and will help us in evaluating each model’s performance.

# Model Development (Training, Validation and Testing)

## Models used

For classification, we surveyed a range of models going from simple and intuitive (CART) to more complex, black box models like Gradient Boosting Models and Deep Learning. For more advanced models, we used the H2O R Wrapper to run H2O. H2O is an open-source machine and deep learning suite of applications used to increase the scalability for a broad range of algorithms.[[3]](#footnote-3) It uses in-memory compression to run millions of rows of data with a small cluster.

We ran seven models:

1. “Simple” CART
2. CART
3. Naïve Bayes
4. Generalized Linear Regression (with Regularization) (GLM)
5. Gradient Boosting Model (GBM)
6. Deep Learning
7. Random Forests

First, we ran a “simple” CART that restricted our decision tree to only the top variables (from filter selection) in order to gain intuition on our dataset. In particular, we restricted the “type of crime” variable to the variable “Against” rather than the more detailed “NIBRS\_Hi\_Class” or “Category” because this variable had far fewer classes (only four versus 30+) which made the interpretation much easier. Appendix 4.4 provides the “Simple” CART decision tree.

Next, we ran a more advanced CART which considered all potential variables. This model became much more complicated but its predictive power was much stronger. Third, we ran a Naïve Bayes on a limited number of variables with a Laplace smoother (lambda = 3). Fourth, we ran Regularized (Lasso) Generalized Linear Regression. We ran regularization on the model in order to reduce unnecessary and redundant features that are included in the dataset. We selected regularized instead of stepwise given that only regularization was available in the H2O package. We selected Lasso (Alpha = 1) instead of Ridge (Alpha = 0) because we found the Lasso performed better on the validation dataset.

In addition to the traditional methods (GLM, CART, Naïve Bayes), we ran three more advanced, black box methods: GBM, Deep Learning and Random Forests. For all three of these models, there were several tuning parameters (e.g. the number of trees and the maximum tree depth for GBM or the number of hidden neurons for Deep Learning).

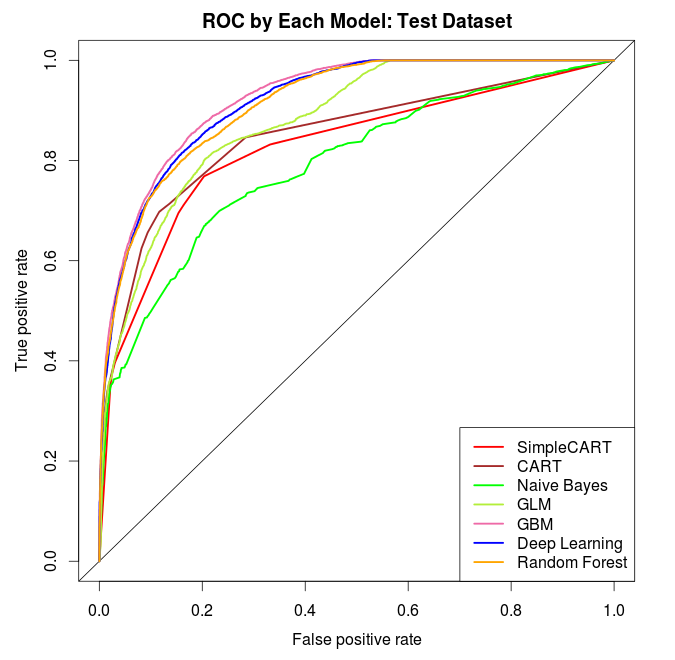
## Model Results

The tables below provide the Accuracy and Area-Under-the-Curve for all models across the three datasets used. We will discuss our results in the next section.



Figure 4: Model Performance Accuracy and Area under the Curve (AUC)

# Output and Results Analysis

Overall, our models performed very well. All models had an accuracy of nearly 80% on the test dataset, with the exception of s Bayes. All models had an AUC of nearly 0.8 or higher on the test dataset.

In particular, GBM had the best performance on the test dataset in both accuracy (84.8%) and AUC (0.9241). Interestingly, GBM showed the highest amount of overfitting on the training dataset, in which it had nearly an 88% accuracy. Nevertheless, even after removing overfitting, GBM still had the best performance on the test dataset. This is shown in the test ROC curve in which GBM (pink) has the highest ROC curve.

Next, Deep Learning and Random Forests performed nearly identically from both an accuracy and AUC perspective. Both showed less overfitting relative to the training dataset and have slightly lower results than GBM’s performance.

The CART and “Simple” CART models performed well too. The CART model performed better than the simpler model, showing the trade-off that simplicity and interpretability can be exchanged for increased predictive power. Even better, both models showed little signs of overfitting as its performance was nearly identical on the training, validation and test dataset.

GLM showed signs of overfitting. Its training accuracy was 82.6% while its test accuracy was 78.3%, which was lower than the “Simple” CART model. Likely, more rigorous feature transformation for non-linearities and perhaps other feature selection techniques (e.g. forward or backward stepwise) may provide less overfitting results.

In conclusion, from a predictive accuracy point of view, GBM was the best model and predicted clear rates with nearly 85% (out-of-sample) accuracy. Nevertheless, this model remains largely a “black box” model in which its components are difficult to interpret. Therefore, for practical use, we recommend that CART models can perform quite well along with interpretable results that practitioners may find usable than black box algorithms like GBM and Deep Learning.

# Appendix

## Exclusion Step 1: Remove Non-Crimes



The dataset includes some non-crimes (e.g. Missing person, suicides, etc.). These crimes are listed as “Not a Crime” according to the NIBRS official classification. In order to reduce noise in the dataset, we propose to remove all “Not a Crime” incidents. After removing these incidents, the total number of incidents each year drops about 8-12% to 79,730 in 2012, 76,696 in 2013 and 78,836 in 2014.

## Exclusion Step 2: Remove Rare Clearance Types

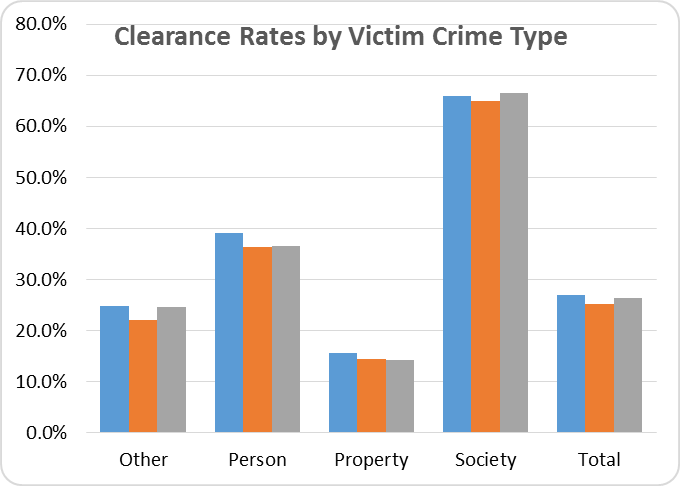


There are several rare clearance statuses that we intend to exclude from our analysis. We’re excluding them because they’re exceptional circumstances that would introduce unnecessary noise into our analysis. These clearance statuses include situations where the offender died, prosecution was declined by the District Attorney and when clearance occurred by another agency. The total number of incidents are reduced by about 300 – 1,000 incidents (< 2%).

## Clear Rates by Crime Victim Type



After applying the two exclusions in the previous two sections, we examine the clearance rates by Crime Victim Type. “All” is a category of small crime types (Category = “All Other Offenses”) that are a combination of person, property, society. “Person” are crimes against individuals (e.g. most are assault crimes). “Property” are crimes against property (e.g. larceny, theft). “Society” are crimes against society in general (e.g. drugs, trespassing, weapon).

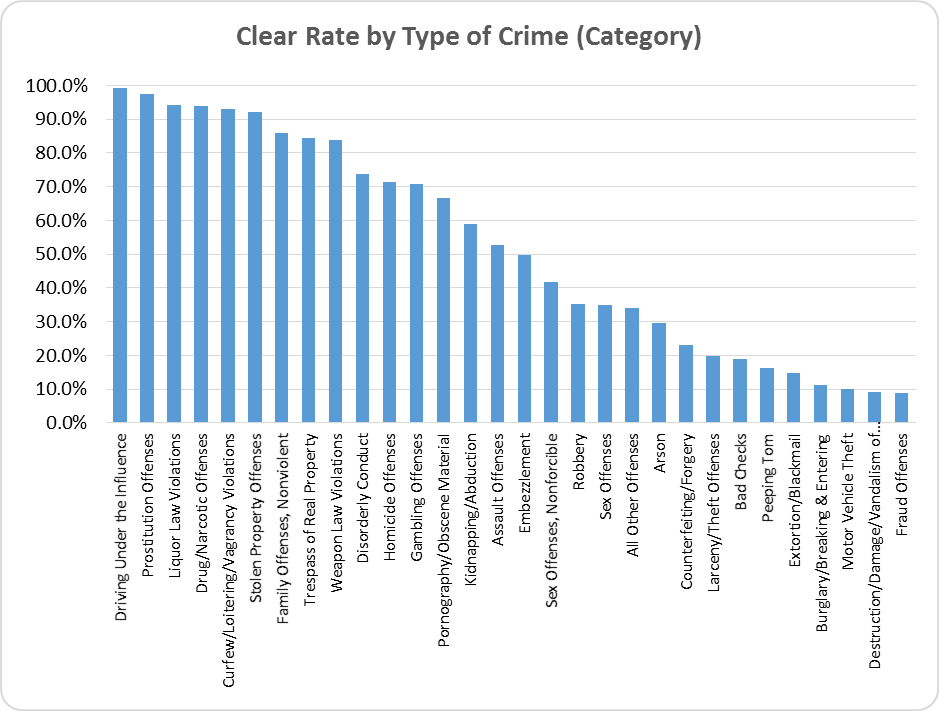


*Classifying crimes by type of victim provides the best clusters for determining clearance rate if all types of crime are controlled.*

We make two observations. First, as expected, the clearance rates differ substantially by the type of crime (as defined as the victim). Crimes against Society have the highest clearance rate while property crimes have the lowest clearance rates. This is what would be expected as crimes against society and people would likely be deemed to be the highest priority to public officials; therefore, they probably receive more resources to solve these types of crimes.

Second, holding the type of crime (by victim) constant, the clear rates are stable across time. This implies that holding for the type of crime, these crimes typically have a normal clearance rate.

In later work, we will consider segment classification by these four types of crimes. In other words, build four models given that we observe and expect that each crime type would have different factors determining its clear rate.

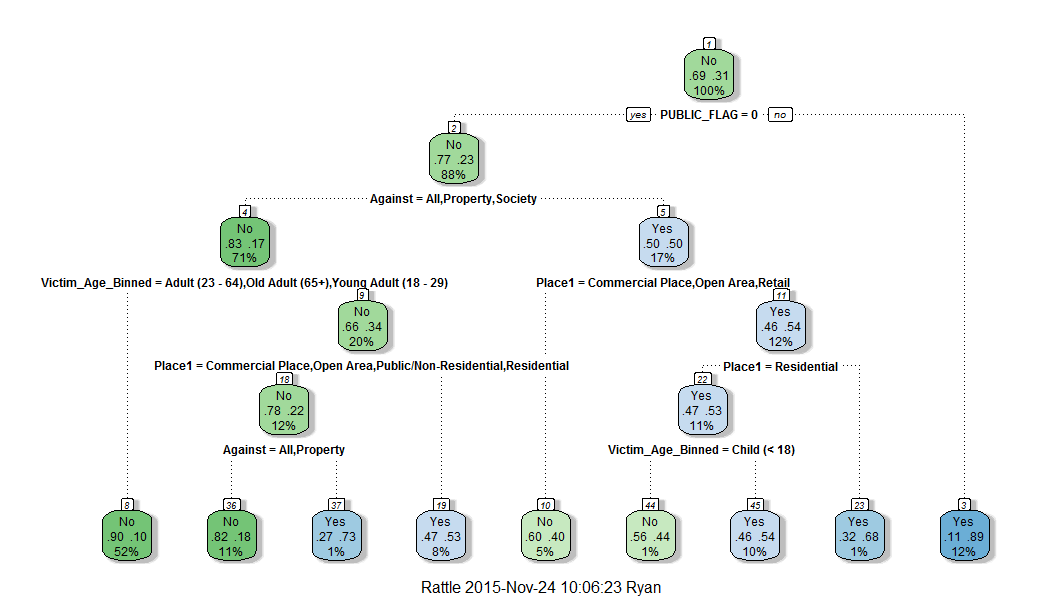


*Some crimes have a higher clear rate due to the nature of the crime and how they are discovered by law enforcement. All the crimes before the first elbow at Stolen Property are mostly cleared presumably because the officer observers the crime being committed, before an arrest is made.*

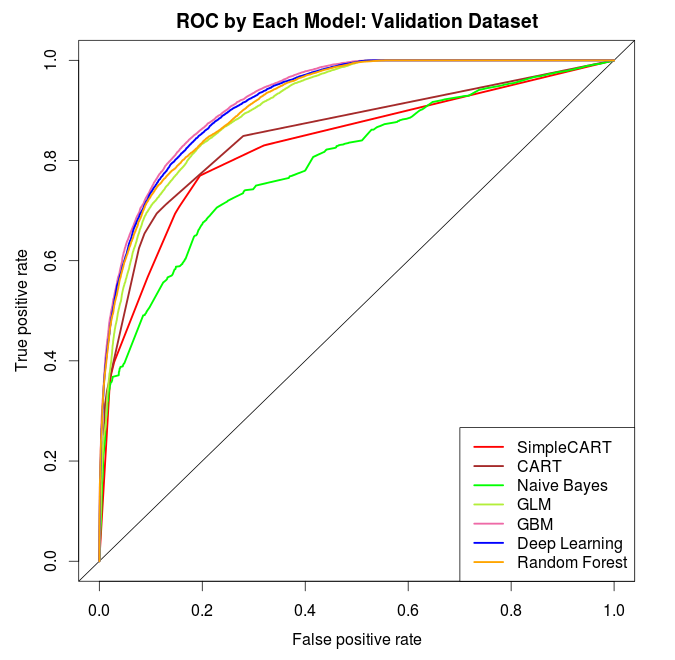
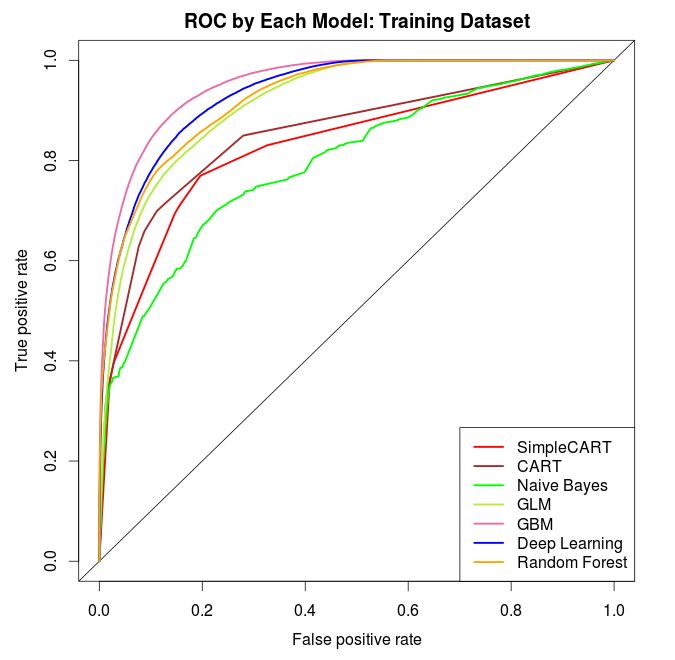
## Predictor Variables



## Simple CART Decision Tree



## ROC Curves for Training and Validation Datasets



1. https://www.fbi.gov/about-us/cjis/ucr/nibrs/2012/resources/nibrs-offense-definitions [↑](#footnote-ref-1)
2. http://clt.charlotte.opendata.arcgis.com/datasets/36d83cc2db854958a092254999992e8f\_0 [↑](#footnote-ref-2)
3. See this site for more details: http://h2o.ai/ [↑](#footnote-ref-3)