Teaching Case

Can You Beat the Case? Predicting an Acquittal Verdict Using CRISP-DM and R

Frank Lee
flee@gsu.edu

Clinton Baxter
cjbaxter1989@gmail.com

J. Mack Robinson College of Bueinss
Georgia State University
Atlanta, Georgia 30303, USA

Abstract

The Cross Industry Standard Process for Data Mining (CRISP-DM) framework was developed in the 1990s and has been widely used as the most relevant and comprehensive leading principle for conducting analytics projects. Despite the wide acceptance and adoption of the CRISP-DM framework, the current business analytics discipline often focuses on the modeling phase and overlooks the interplays between the phases. Consequently, students often lack a comprehensive understanding of the entire analytics process. This teaching case is created to demonstrate the importance of the data analytics life cycle and how six phases collectively contribute to the success of analytics projects using R. This case collects real-life data and follows the six CRISP-DM phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. At the end of the project, students will learn the importance of the data analytics life cycle, especially the data understanding and preparation phases, which often receive minimal attention in business analytics projects. This project will also demonstrate the importance of storytelling, ensuring that critical insights are conveyed to the audience.

Keywords: Data Mining, Legal Analytics, CRISP-DM, Analytics Project, R

1. INTRODUCTION

Under the United States Justice system and the Constitution, defendants are afforded indelible rights such as the right to remain silent, the right to counsel, and of course, the right to a speedy and public trial by an impartial jury. If these rights are breached or violated, this could trigger a cascading of events —generally in the form of appeals—that could result in the original verdict being overturned and the case being dismissed or retried. In a perfect system, these indelible rights are always upheld under the U.S. Constitution, and for the remainder of this paper, you will assume they are and have been. Although these rights are indelible, they are not binding, meaning defendants can choose to waive their rights at any time. For example, a defendant may waive their right to remain silent and testify in a court of law or waive their right to a jury trial and plead the case. This project is designed to delve deeper into those cases that go to trial and their associated outcomes; guilty or not guilty (acquittal).

Your first mission is to understand whether surface-level variables like the age, sex, ethnicity of the defendant (and plaintiff), and other variables play a significant role in a jury's verdict of guilty or not guilty. Electing for a jury trial can be very time-consuming, stressful (for all parties), and unpredictable. You need to provide a defendant on trial for murder in Cobb County, Georgia, with a probability of receiving an acquittal verdict.

The Cross Industry Standard Process for Data Mining

The Cross Industry Standard Process for Data Mining (CRISP-DM) framework was developed in the 1990s and has been widely used as the most relevant and comprehensive leading principle for carrying out analytics projects (Wirth & Hipp, 2000). CRISP-DM consists of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment, with arrows indicating the most important and frequent dependencies between phases. The sequence of the phases is flexible and can be customized easily. Project work can occur in several phases simultaneously, and the movement can be either forward or backward between phases, as necessary.

Learning Objectives

By completing this assignment, you will be able to:

- Describe the data analytics project lifecycle and critical elements of each phase
- Obtain sufficient relevant data and conduct data analytics using scientific methods
- Apply appropriate and powerful connections between quantitative analysis and real-world problems
- Present descriptive statistics and models in business context and employing appropriate data visualizations
- Apply advanced techniques to conduct thorough and insightful analysis
- Interpret the results correctly with detailed and useful information

2. CASE BACKGROUND

Case Text

Sarah Brown, a defendant in an upcoming murder case in Cobb County, Georgia, has enlisted the services of the local law firm Confident Cases LLC. Clint Baxter will be representing her as legal counsel throughout her case. As an astute lawyer, Clint understands that a defendant enjoys the presumption of innocence. At the same time, the onus relies on the State's prosecution to provide evidence beyond a reasonable doubt that the defendant is guilty of the charges. The choice is always up to the defendant, who ultimately must make the decision regardless of Clint's recommendations. Clint wants his defendant to make the best suitable decision for her (and her family) as these decisions will have enormous life-changing impacts. Clint knows prior cases are public records, allowing him to data mine and collate specific data points into a working data set. Using his prior statistical knowledge, Clint wants to give Sarah the probability of beating the case (being acquitted of murder), so Sarah can make the most informed decision.

The Data Source

You will use data from the Cobb County Clerk of the Superior Court query to look up individual murder cases from the last ten years in the Cobb County area. The Cobb County Clerk of the Superior Court allows you to filter the court cases by charge type using a specific date range. Once you load the query for murder cases from the last ten years, you will look through each case's sentencing documents to determine if the defendant had pleaded out their case or chose a jury trial. If they went to a jury trial, you will include their case in the model as a data point and gather all the case variables. If they plead out, their information can be disregarded from the model.

Assumptions

Jury selection is part of the judicial process but is extremely time-consuming and outside this analysis's scope. The critical assumption in analytics projects is that the future will continue to be like the past. MacCoun (1989) used Bayesian methodology to conduct analysis on mock juries to uncover any innate biases each juror may have prior to their selection. The study concluded that it is difficult to predict human behavior on such a vast scale with many variables. Therefore, we will not be diving into each juror's prior disposition but rather assume each juror is a "rational" person. While we understand those variables could certainly play a role in each outcome, we only want to look at known variables. In addition, you should also assume all constitutional rights have been upheld, so you will not be looking at any future data.

3. PROJECT ACTIVITY

Business Understanding

This stage focuses on understanding the

intentions and requirements of the project from the business perspective and converting this knowledge into an analytics problem. This stage also determines the aim of the project and designing the analytics plan. This activity aims to understand whether surface-level variables like the age, sex, ethnicity of the defendant (and plaintiff), and other variables play a significant role in a jury's verdict of guilty or not guilty. To accomplish the objective, you need to provide a defendant on trial for murder with a probability of receiving an acquittal verdict.

Data Understanding

This phase involves an initial data collection and proceeds to activities that help the participants become familiar with the data. This project intends to use individual murder cases from the last ten years from the Cobb County Clerk of the Superior Court in Georgia. To find the relevant data, students first look through each case's sentencing documents to determine if the defendant pleaded out their case or went to a jury trial.

The following steps will help guide you along during your data collection. Please make sure to follow each of the steps in order.

1. Visit https://ctsearch.cobbsuperiorcourtclerk.com /CaseType and filter by case type "criminal" for murder cases from 2004 to 2019. (Appendix A, Figure 1).

2. Click on the paper icon located in the "view" column next to the defendant's name. (Appendix A, Figure 2).

3. Go to pleadings and search for "jury list" to identify if it is a jury trial case. Note: If you do not find a jury list, the case was more than likely pleaded out. Keep in mind that the jury list will never be available to protect each juror's anonymity. (Appendix A, Figure 3).

4. You can also obtain the prosecutor's name, case I.D., and the defendant's name at the top of the "Case Details." (Appendix A, Figure 4).

5. Next, click on the "Attorneys" tab to identify the attorneys the defendant retained. If the status shows active, you assume the attorney represented the defendant until trial completion. If it says released, they are not counted as an attorney for the defendant in your model. In the example provided, there are four. (Appendix A, Figure 5).

6. Next, look under the "defendants" tab to see how many codefendants there are. In the example provided, there are none. (Appendix A, Figure 6).

7. Next, look under the pleading tab for "list of witnesses." This pleading pdf will be locked (to protect the identity of the witnesses involved), but if you see it listed, you know the case involved eyewitness testimonies. (Appendix A, Figure 7).

8. Next, look under the pleadings tab once again for the indictment pdf. Once opened, scroll through the indictment to determine how the murder was committed, when the murder took place, and the total number of victims. For example, you may find that a murder took place with one victim on 11/05/2003 by firearm. You'll need to categorize the murder methods by "Firearm," "Stabbing," or "Other." (Appendix A, Figure 8).

9. Next, go to the "offenses" tab to identify how many charges were brought against the defendant(s). (Appendix A, Figure 9).

10. You'll need to obtain the verdict handed down by the jury. This can be found in the "verdict" document under the "pleadings" tab or in the "sentence" document if the verdict document is sealed. Remember, if they were acquitted of other counts but guilty of even one count of murder, the case is considered a loss for the defendant. (Appendix A, Figure 10).

11. The last two variables needed are prior criminal convictions and age at the time of the murder. They are readily available public information.

    Visit http://www.dcor.state.ga.us/GDC/OffenderQuery/jsp/OffQryForm.jsp

    Use the search to locate the convicted inmate. You'll also find the convict's DOB as well as any other prior convictions (calculate the age of the defendant during the trial by subtracting the sentencing date from their DOB). More digging might be necessary to identify the remaining variables if the defendant was proven not guilty. (Appendix A, Figure 11).

Data Preparation

This phase selects a subset of the data, performs data cleansing, and prepares the data for analysis. You are looking for completeness,

ISSN: 2473-4901
v8 n5731

consistency, and accuracy in the data. You must ensure all columns were filled appropriately with their corresponding values and spot-checked any inconsistencies before loading into R. (Appendix A, Figure 12).

You must provide a few aggregated/ summarized statistics before data preparation and modeling. The summary statistics allow you to identify patterns while improving your understanding of the data. During your data collection, you gathered various attributes, including the sex of the defendant, the defendant's age, and whether the defendant had any prior criminal convictions.

First, find and visualize the distribution of the defendants' age in your dataset. You may create age buckets such as <18, 18-25, 26-35, 35-50, and 50+. Discuss the findings.

Second, find and visualize the distribution of the method of the murder. Discuss the findings.

Third, visualize and discuss the distribution of "Guilty" vs. "Not Guilty" between sex.

Finally, create a visualization that clearly shows the relationship between the "Age" of the defendant and the "Number of Charges." Do you notice a pattern or any relationship? Discuss your findings.

Modeling

This phase involves the selection and development of analytics techniques and models. In addition, portions of a data set are often set aside for training and validating the model(s). This teaching uses the programming language R for illustration, but all the analytics tasks can be similarly completed with any other software such as Python or RapidMiner.

Decision Trees models are quite popular supervised models for various reasons: they are easy to implement and interpret, and the complexity of a full tree can be optimized by incorporating pruning. You may start your analysis with a Decision Tree that uses the Classification and Regression Trees (CART) algorithm and move to an Ensemble methodology (Bagging and Boosting) which can help with the overfitting problem seen with single decision trees.

Classification Tree

To run the decision tree model in R, you need to do data preprocessing by converting your categorical data into factors using the **as factor ()** function. (See R Codes in the Appendix B).

Next, partition the data into a train and validation set using a 60/40 ratio to create the default tree.

Create the default classification tree using the ***repart*** function.

Next, create a full tree that you can prune appropriately based on the cp (complexity parameter) results. Find the best pruned three with the least complexity. To identify the cp value associated with the smallest cross-validated classification error, use the ***printcp*** function to display the complexity parameter table.

```
        CP nsplit rel error  xerror    xstd
1 0.545455      0  1.000000 1.00000 0.26629
2 0.090909      1  0.454545 0.45455 0.19285
3 0.045455      3  0.272727 0.81818 0.24697
4 0.022727      7  0.090909 1.09091 0.27454
5 0.000000     11  0.000000 1.18182 0.28196
```

Figure 1 Complexity Parameter Table

Here you can see the best-pruned tree with the least complexity is the second one with the lowest ***xerror*** score of 0.45455, which is still the lowest when factoring in the ***xtd*** score (0.45455+0.19285 = 0.6474).

Next, run the prediction and create the confusion matrix as well as the Lift, Decile-wise, and ROC charts. Evaluate the model using accuracy, sensitivity, and specificity.

```
          Reference
Prediction  0  1
         0 22  3
         1  4  4

               Accuracy : 0.7879
                 95% CI : (0.6109, 0.9102)
    No Information Rate : 0.7879
    P-Value [Acc > NIR] : 0.5994

                  Kappa : 0.3969

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.5714
            Specificity : 0.8462
         Pos Pred Value : 0.5000
         Neg Pred Value : 0.8800
             Prevalence : 0.2121
         Detection Rate : 0.1212
   Detection Prevalence : 0.2424
      Balanced Accuracy : 0.7088

       'Positive' Class : 1
```

Figure 2 R CART Confusion Matrix

ISSN: 2473-4901
v8 n5731

The model has a decent accuracy at 0.7879 and a quite reasonable specificity at 0.8462. However, the model lacks the ability to correctly classify the target class, which in our case is a verdict of Not Guilty with a subpar score of 0.5714.

You still want to understand more about the model's overall performance and finish this model by completing the Lift, Decile-wise, and ROC charts.

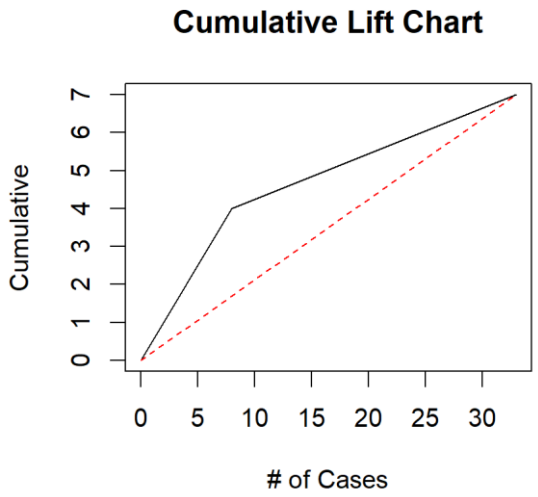


Figure 3 Cumulative Lift Chart

As you can see, though the model's sensitivity is not within our acceptable range, you know the model is better at predicting a Not Guilty verdict when compared to a random guess.

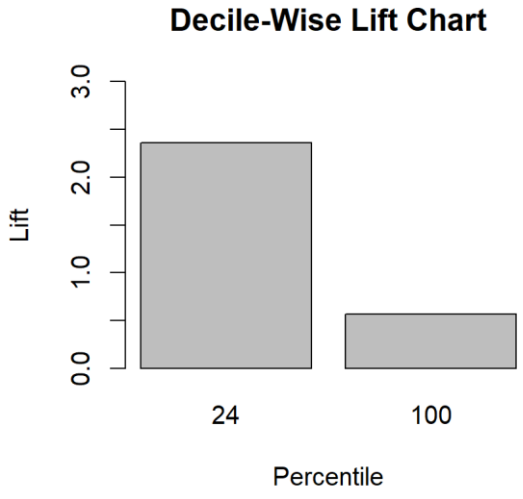


Figure 4 Decile-Wise Lift Chart

Additionally, after reviewing the decile-wise lift chart, you can conclude that the model's top 24% of the observations contain 2.25 times as many Class 1 cases as the 24% of the randomly selected observations.

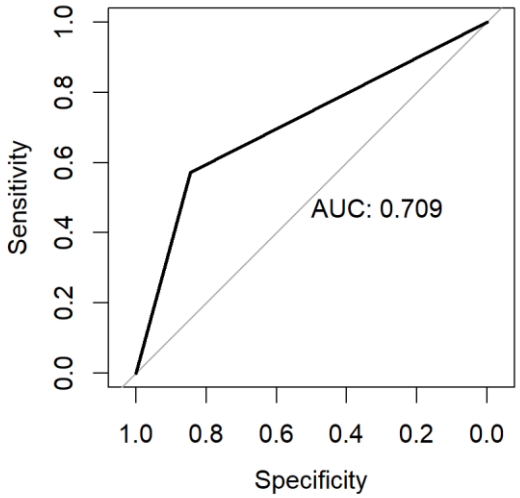


Figure 5 ROC Chart

You can see by viewing our ROC that while the sensitivity doesn't quite match up to what you want or expect, the specificity is still quite good. You can validate this by displaying the AUC score, which is right around the 0.71 mark in this model's case.

Ensemble (Bagging)

We now want to know how well our ensemble models will perform, so for this, you will need to complete the same minor preprocessing step as from our CART tree model.

Again, you will begin by splitting the data set into a train and validation set to maintain consistency across all models. You will use a 60/40 split. Once complete, run the model using the **randomForest()** function and specify the number of variables by setting the **mtry** option equal to 10—this tells the model to use a bagging strategy by using all ten predictor variables in the model.

While running the model, you also want to know how important each feature is to the model. Using the ***varImpPlot()*** function, you can visually identify which variables are important in terms of an average decrease in accuracy if they were omitted.

ISSN: 2473-4901
v8 n5731

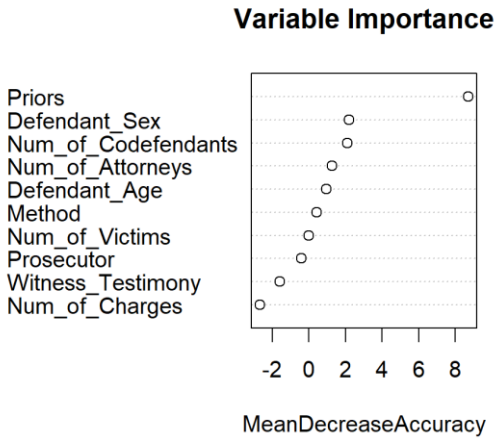


Figure 6 Variable Importance Chart

Not surprisingly, whether a defendant has any prior criminal convictions is extremely important to the model, meaning the model would suffer a tremendous decrease in accuracy if we dropped this variable. Conversely, we could drop the Num_of_charges variable and possibly notice a slight increase in accuracy—which makes sense as prior criminal history is oftentimes suppressed during a trial. After running the model, you want to view the confusion matrix, as shown below.

```
          Reference
Prediction  0  1
         0 20  2
         1  6  5

               Accuracy : 0.7576
                 95% CI : (0.5774, 0.8891)
    No Information Rate : 0.7879
    P-Value [Acc > NIR] : 0.7462

                  Kappa : 0.4

 Mcnemar's Test P-Value : 0.2888

            Sensitivity : 0.7143
            Specificity : 0.7692
         Pos Pred Value : 0.4545
         Neg Pred Value : 0.9091
             Prevalence : 0.2121
         Detection Rate : 0.1515
   Detection Prevalence : 0.3333
      Balanced Accuracy : 0.7418

       'Positive' Class : 1
```

Figure 7 R Bagging Confusion Matrix

This model delivers a much better sensitivity rating than the previous decision tree model. We do notice a slight tick down in the precision, meaning this bagging model might present clients with a false hope to beat the murder charges as the model is classifying more Not Guilty verdicts that are in reality Guilty verdicts. You also should notice a degradation in the specificity meaning this model is not quite as good as classifying our non-target class (Guilty verdicts). Just as you did in the previous model, you will need to create cumulative lift, decile-wise, and ROC charts.

Ensemble (Boosting)

You will use another ensemble method with a boosting strategy in your final model. You will again prepare the data using the same techniques as the previous model.

Setting ***mfinal*** equal to 100 tells the model to repeatedly sample across multiple weak learner single trees.

```
          Reference
Prediction  0  1
         0 23  2
         1  3  5

               Accuracy : 0.8485
                 95% CI : (0.681, 0.9489)
    No Information Rate : 0.7879
    P-Value [Acc > NIR] : 0.2701

                  Kappa : 0.5692

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.7143
            Specificity : 0.8846
         Pos Pred Value : 0.6250
         Neg Pred Value : 0.9200
             Prevalence : 0.2121
         Detection Rate : 0.1515
   Detection Prevalence : 0.2424
      Balanced Accuracy : 0.7995

       'Positive' Class : 1
```

Figure 8 R Boosting Confusion Matrix

As you can see, the confusion matrix for the boosting model looks quite promising, excelling in each performance statistic well above the others. This model provides a high accuracy, great sensitivity, precision, and specificity rates.

Evaluation and Deployment

This evaluation phase involves reviewing and interpreting the analysis results in the context of the business objectives and success criteria

described in the first phase. Lastly, the deployment stage translates the knowledge gained from data analysis into a set of actionable recommendations.

4. PROJECT REPORT

You need to write a comprehensive project report. The project report should provide an executive summary, introduction, data collection, data preparation, methodology, conclusion, reference, and appendix. Specifically, 1) after evaluating and running each model, you should be able to compare their results. 2) Your discussion should focus, in particular, on the results that are most interesting, surprising, or important. 3) Interpret the results with detailed and valuable information. It would be best if you also discussed the consequences or implications. 4) Finally, if the answers or findings are unexpected, see whether you can find an explanation for them, such as other factors that your analysis did not include.

5. REFERENCES

MacCoun, R. J. (1989). Experimental Research on Jury Decision-Making. *Science*, 244(4908), 1046-1050. http://www.jstor.org/stable/1703992

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*.

ISSN: 2473-4901
v8 n5731

APPENDIX A
Case Figures

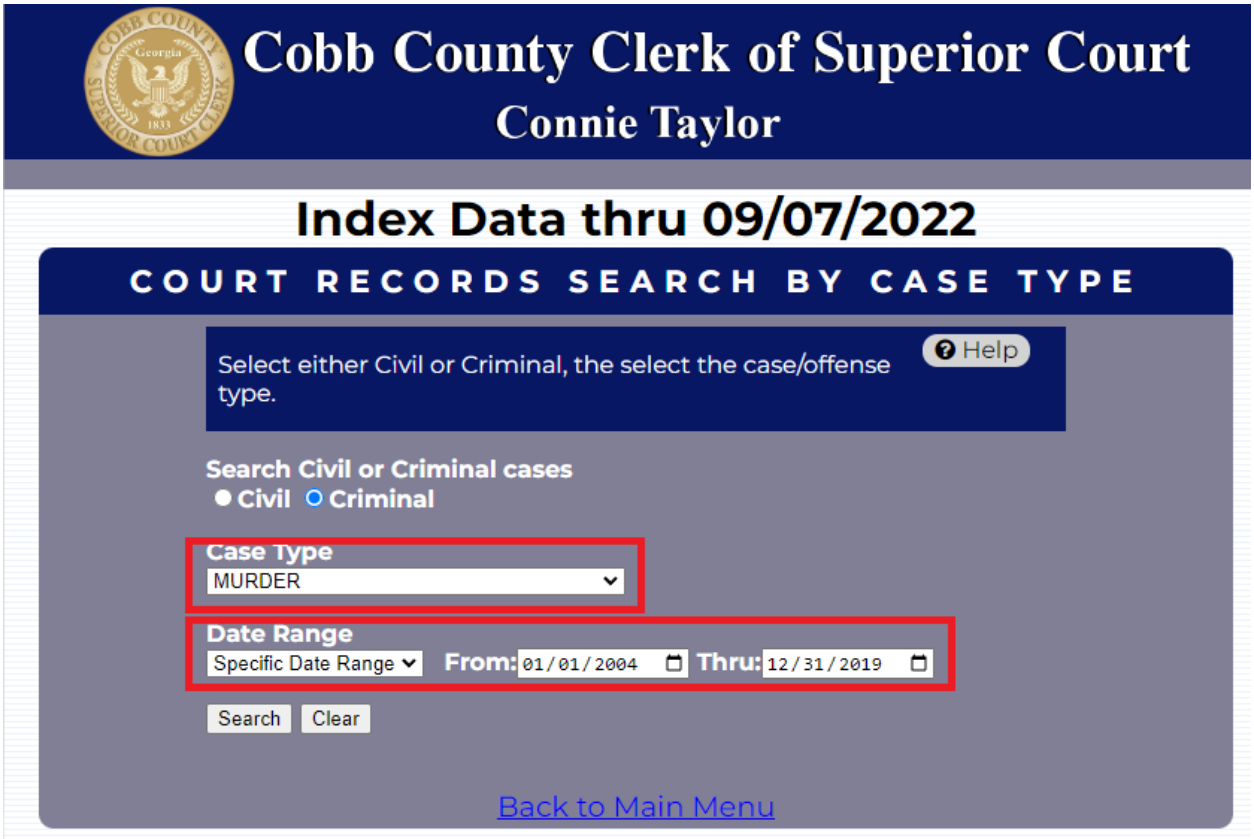


Figure 1. Cobb County Clerk of the Superior Court

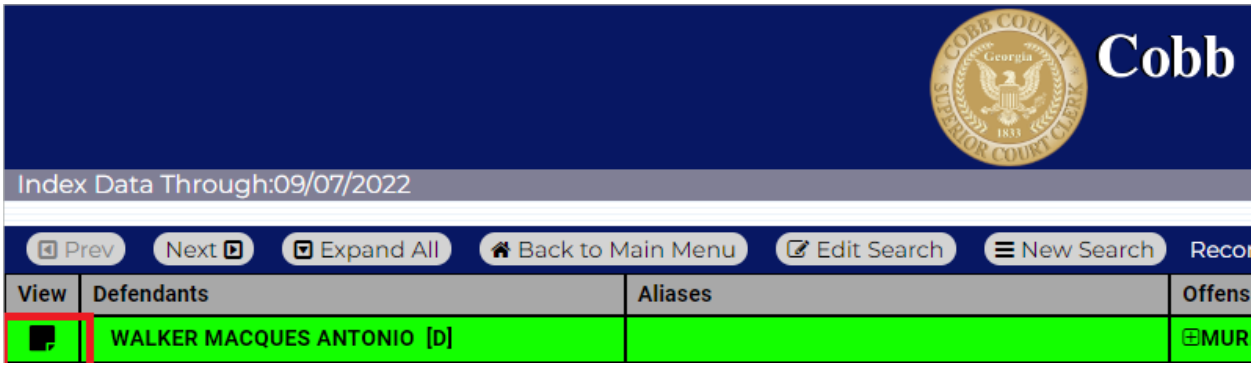


Figure 2. The view column next to the defendant's name

ISSN: 2473-4901
v8 n5731

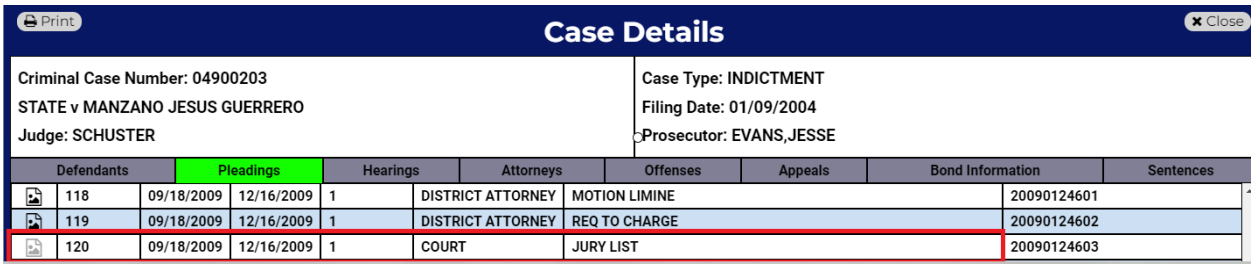


Figure 3. Jury trial case

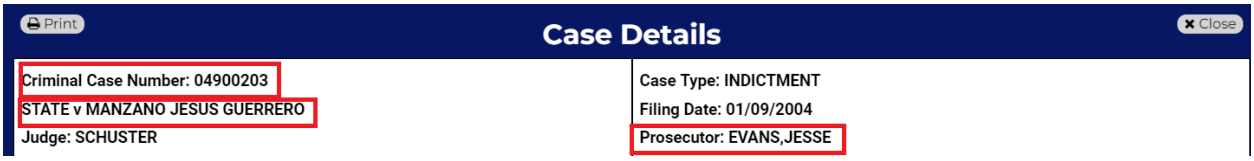


Figure 4. Prosecutor Information

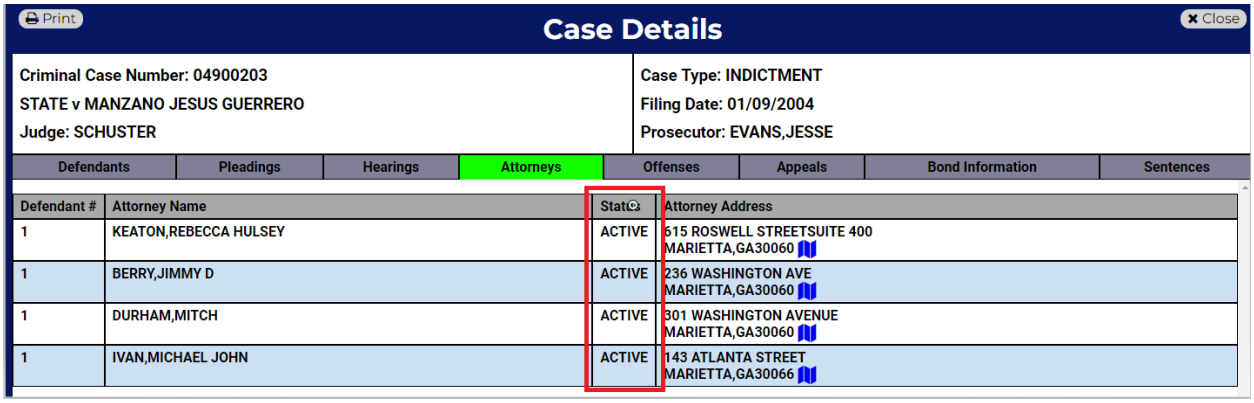


Figure 5. Attorney Informtion

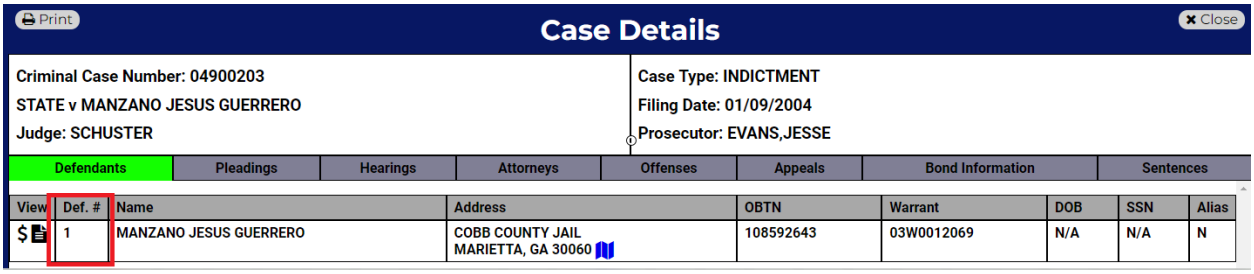


Figure 6. Codefendants

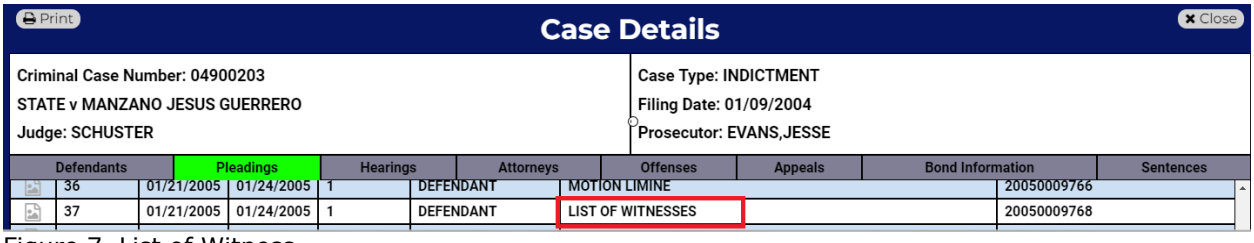


Figure 7. List of Witness

ISSN: 2473-4901
v8 n5731

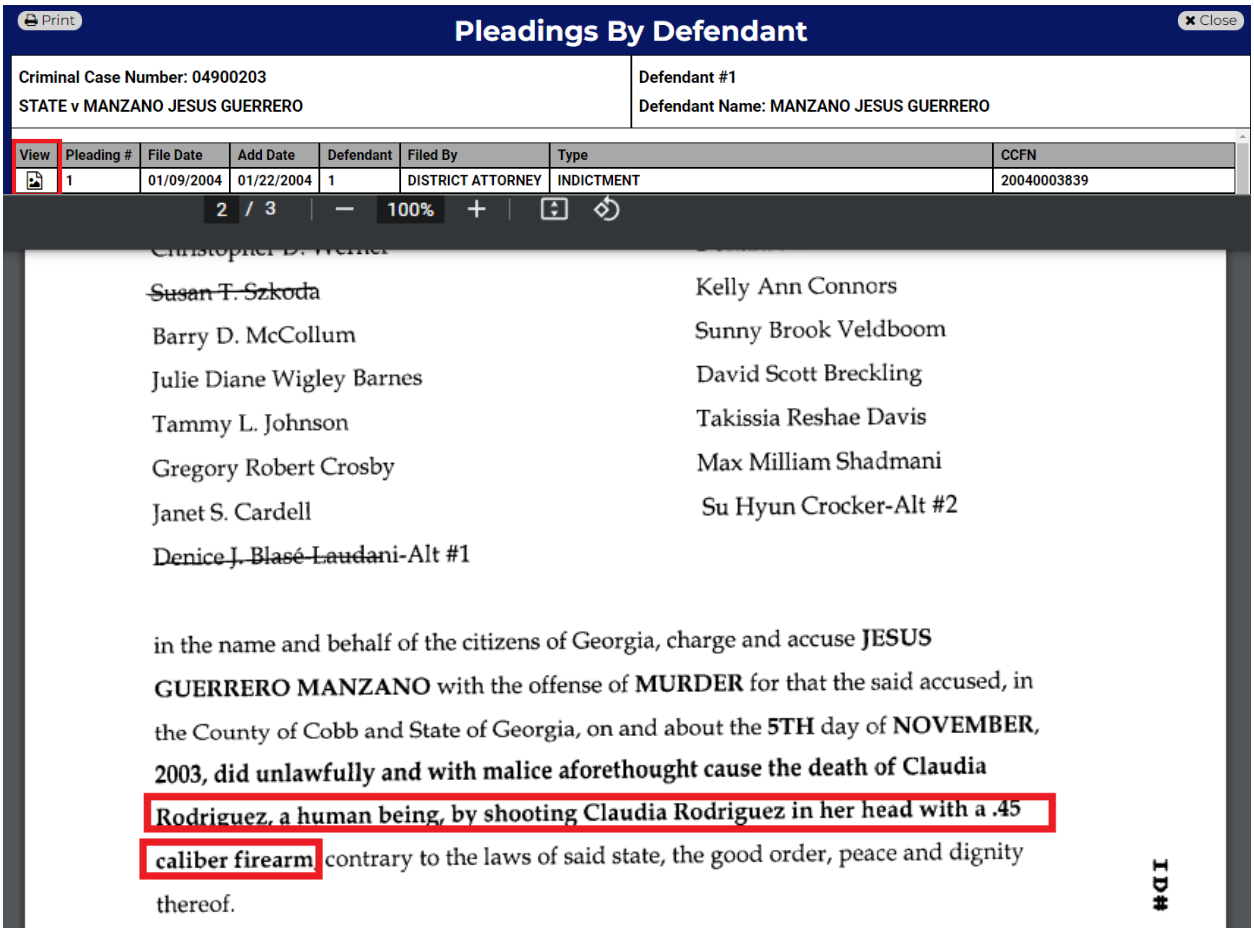


Figure 8. Murder Method

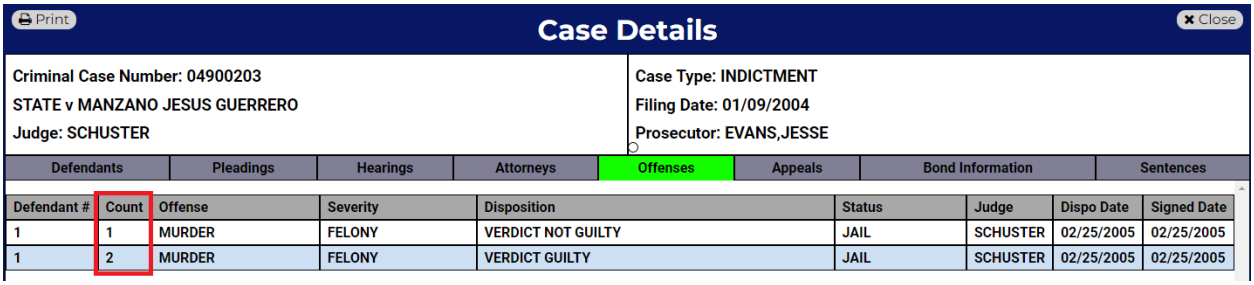
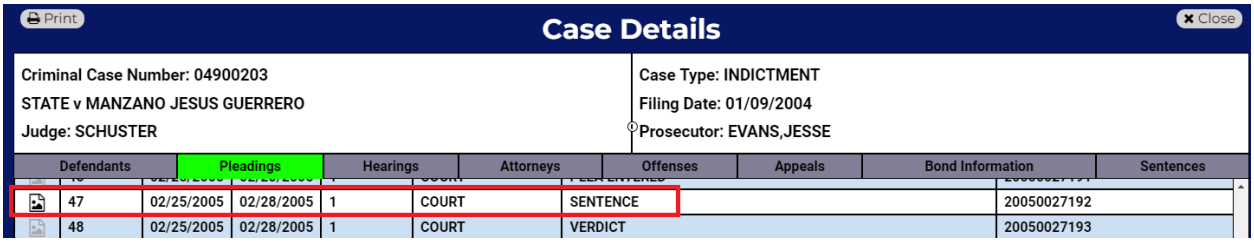


Figure 9. The number of Charges



ISSN: 2473-4901
v8 n5731

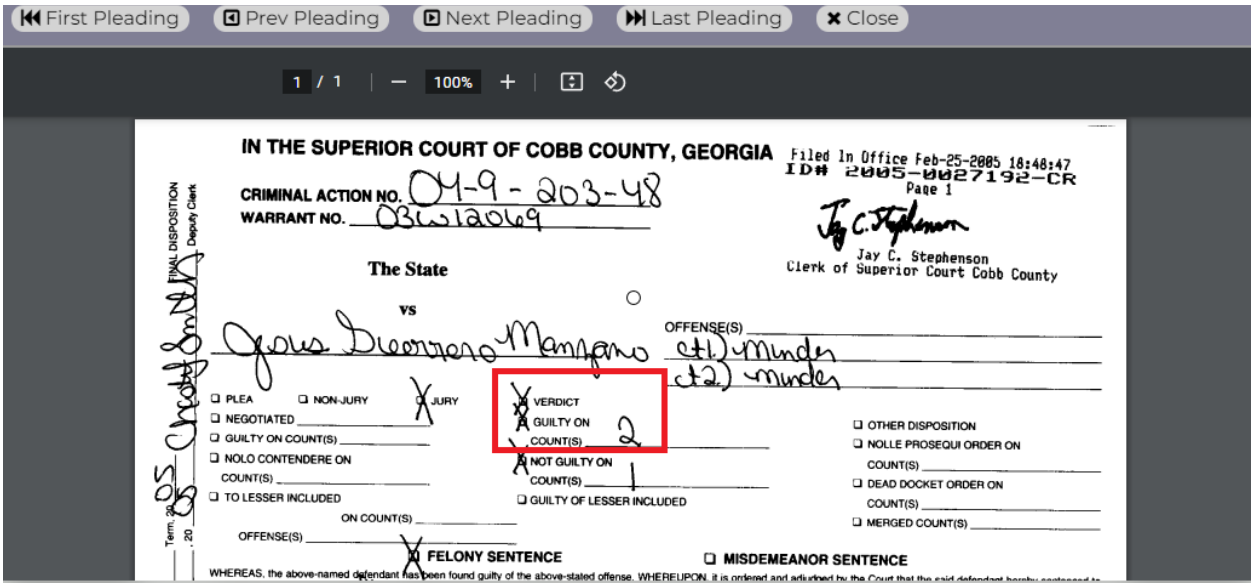


Figure 10. Verdict

ISSN: 2473-4901
v8 n5731

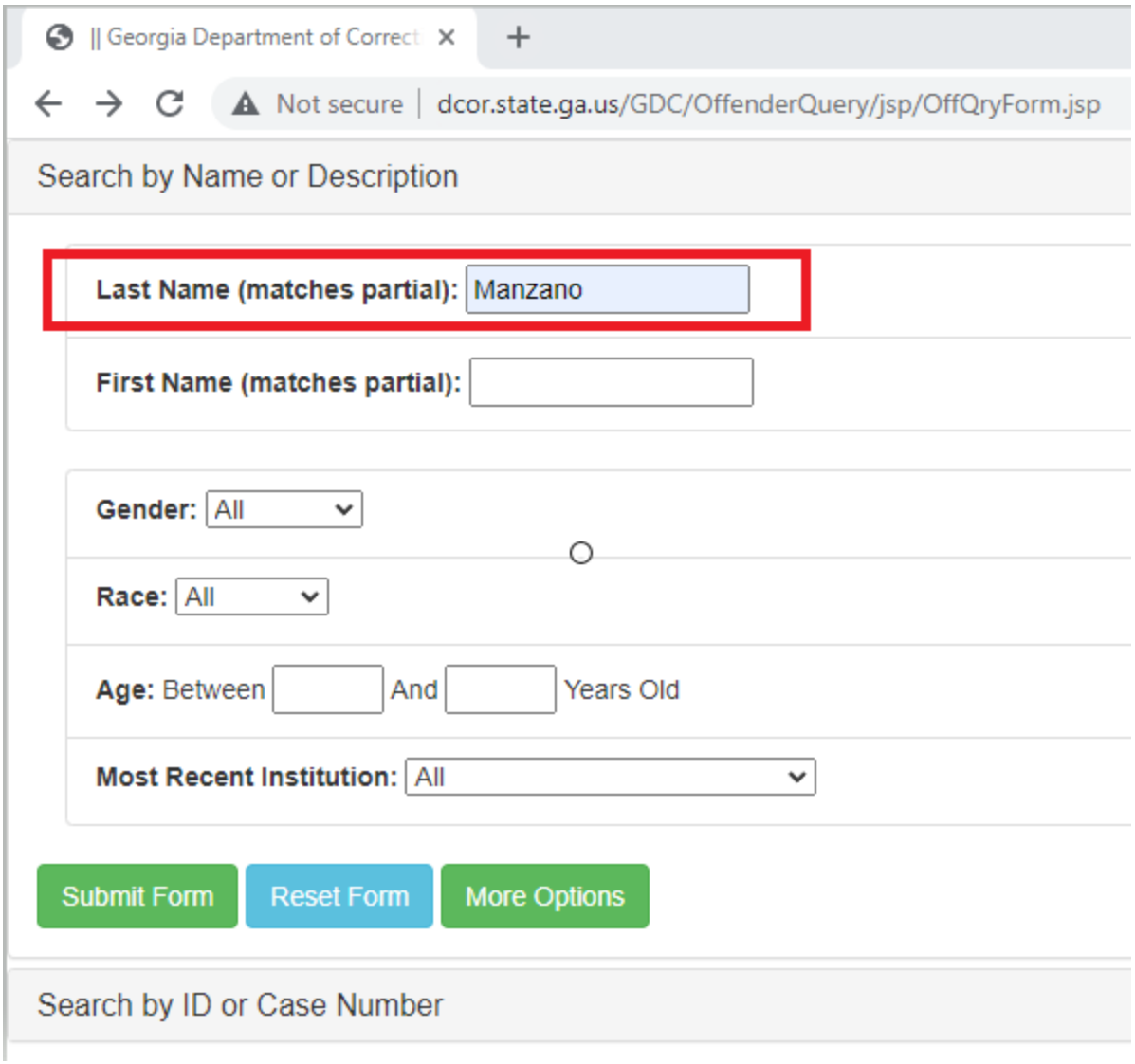


Figure 11. Public Criminal Information

ISSN: 2473-4901
v8 n5731

Case Id	Defendant Name	Num_of_Codefendants	Defendant_Age	Defendant_Sex	Num_of_Att
4900191	Antonio Walker	0	30	1	3
4900199	Andre Lawrence	1	16	1	2
4900203	Jesus Manzano	0	27	1	4
4900673	Stacey Humphreys	0	30	1	6
4901186	Rodgerick Swanson	0	42	1	2
4904319	Aaron Willis	0	21	1	3
4904974	Donovan Leger	0	33	1	2
6902974	Sonya Yvette Smith	1	39	0	4
6903363	Colton Williams	5	16	1	2
7900473	Eliot Ellerton Jeffers	1	30	1	4
7901329	William Brian Hughes	0	32	1	2
7901990	Natasha Wynetta Demery	0	41	0	2
7902004	Mario Hodges	0	40	1	2
7902824	Christian Javon Wornum	0	18	1	2

Figure 12. Data for Analysis

APPENDIX B
R Codes

```
suppressWarnings(RNGversion("3.5.3"))
install.packages(c("randomForest"))
install.packages("adabag")

library(caret)
library(gains)
library(rpart)
library(rpart.plot)
library(pROC)
library(randomForest)
library(readxl)
library(adabag)
myData_DT <- read_excel("Final_Project.xlsx", sheet = "Verdict_Data")

myData_DT$Verdict<- as.factor(myData_DT$Verdict)
myData_DT$Presecutor <- as.factor(myData_DT$Prosecutor)
myData_DT$Priors <- as.factor(myData_DT$Priors)
myData_DT$Method <- as.factor(myData_DT$Method)
myData_DT$Witness_Testimony <- as.factor(myData_DT$Witness_Testimony)
myData_DT$Defendant_Sex <- as.factor(myData_DT$Defendant_Sex)

myData_DT <- myData_DT[, 3:13]
View(myData_DT)

set.seed(1)
myIndex <- createDataPartition(myData_DT$Verdict, p=0.6, list=FALSE)
trainSet <- myData_DT[myIndex,]
validationSet <- myData_DT[-myIndex,]
View(trainSet)

set.seed(1)
default_tree <- rpart(Verdict ~., data = trainSet, method="class")
summary(default_tree)
prp(default_tree, type=1, extra=1, under=TRUE)

data.frame(imp = default_tree$variable.importance)

set.seed(1)
full_tree <- rpart(Verdict ~., data = trainSet, method="class", cp=0, minsplit=2, minbucket=1)
prp(full_tree, type=1, extra=1, under=TRUE)
printcp(full_tree)

data.frame(imp = full_tree$variable.importance)

pruned_tree <- prune(full_tree, cp=0.545454)
prp(pruned_tree, type=1, extra=1, under=TRUE)

predicted_class <- predict(pruned_tree, validationSet, type="class")
confusionMatrix(predicted_class, validationSet$Verdict, positive="1")

data.frame(actual = validationSet$Verdict, predicted = predicted_class)


predicted_prob <-predict(pruned_tree, validationSet, type="prob")
validationSet$`Verdict (1=Not Guilty)` <- as.numeric(as.character(validationSet$Verdict))
```

```
validationSet$Verdict <- as.numeric(as.character(validationSet$Verdict))
gains_table_DT <- gains(validationSet$Verdict, predicted_prob[,2])
gains_table_DT

plot(c(0, gains_table_DT$cume.pct.of.total*sum(validationSet$Verdict)) ~ c(0,
gains_table_DT$cume.obs),
    xlab="# of Cases",
    ylab ="Cumulative",
    main="Cumulative Lift Chart",
    type="l")
lines(c(0, sum(validationSet$Verdict)) ~ c(0, dim(validationSet)[1]), col="red", lty=2)
barplot(gains_table_DT$mean.resp/mean(validationSet$`Verdict (1=Not Guilty)`),
names.arg=gains_table_DT$depth,
       xlab="Percentile",
       ylab="Lift",
       ylim=c(0,3),
       main="Decile-Wise Lift Chart")
roc_object_DT <- roc(validationSet$Verdict, predicted_prob[,2])
plot.roc(roc_object_DT, print.auc = TRUE)
auc(roc_object_DT)
```