Upset Potential in the National Football League (NFL)

Mary J. Snyder

Department of Computer & Information Sciences   
College of Science & Mathematics   
Towson University

msnyde8@students.towson.edu

Advisor: Michael P. McGuire Ph.D.

Department of Computer & Information Sciences   
College of Science & Mathematics   
Towson University

mmcguire@towson.edu

**ABSTRACT**

Fantasy leagues, spread picks, as well as confidence points picks have become popular among fans of a variety of sports today. While the ultimate goal of these activities is to choose all winners, the place to gain a competitive edge over opponents is in picking upsets correctly. Nowhere is this more apparent than in confidence points picks. Placing too much confidence on a game that has a surprise upset could harm the overall score more than picking a few low confidence games incorrectly.

Analysis on upset potential can be used not only in these fun situation, but the information can also help teams better prepare for games. Knowing how factors such as time zone changes, number of days rest between games, overtime games played, weather, or key player injuries can affect the outcome of the game would help teams prepare themselves against an upset. This information can also be used to a team’s advantage by providing them ways to focus their training to upset other teams.

In this project, I will examine past NFL stats and data to determine the factors that contribute to upset games. Using the knowledge I gain from the analysis, I will attempt to predict upsets for games during an ongoing NFL season.

**Categories and Subject Descriptors**

H.2.8 **[Database Management]** Database Applications – *Data mining*

**Keywords**

National Football League (NFL); Confidence picks; Exploratory data analysis; Multivariate; Categorical; Histogram; Scatterplot; Scatterplot Matrix; Normalization; Binning; K-means; Regression Analysis; Classification; Decision Tree Classification; Training and Testing; Accuracy; Error Rate; Sensitivity; Specificity; Precision; Recall; F Measure; Clustering; Cluster Analysis; Partitioning Approach; k-means; Silhouette Plot;

# INTRODUCTION

## Dataset

The data required for this analysis was widespread and was not available through one source.

Injury data was particularly difficult to standardize. While every team is allowed the same number of starting players, each team has the authority to assign them as they see fit. This means not only the positions that each team starts may vary, but also the number of players in a position may also vary. For example, on offense, every team will have a quarterback (QB) and a center (C), but one team may include a single running back (RB), another may have two running backs, and a third may have three fullbacks (FB). As for defense, some teams may include a nose tackle (NT), others a left/right outside linebacker (LOLB/ROLB), while others may have multiple defensive ends (DE). This made determining critical positions across all teams almost impossible; therefore, I opted to treat all positions with equal importance. Instead, I focused on whether the player for any starting position was probable/doubtful yet still played versus did not play at all. I gave more points towards the team’s total offense and defensive injury total if a player did not play at all versus played in a limited capacity.

We chose a dataset from the UCI Machine Learning Repository whose attributes are categorical in nature. This Community and Crime dataset comprises socioeconomic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. The dataset contains information that describes or involves the community, such as the percent of the population considered urban and the median family income, as well as involve law enforcement, such as per capita number of police officers and percent of officers assigned to drug units. The issue of violent crime has been a contentious domestic political issue with sweeping implications. In this analysis, we will explore other indicators of communities along with its prevalence of violent crime to determine what relationships exist and suggest further areas of study.

### Data Attributes Selection and Values

The attributes selected for this dataset were from larger databases of census and crime information. They were included here if there was any plausible connection to crime and/or the attribute the dataset authors intended to be predicted, Per Capita Violent Crimes (ViolentCrimesPerPop). There are 1994 instances with some missing values (indicated by a ‘?’ value) and 128 attributes: 122 predictive, 5 non-predictive, and 1 goal.

Non-predictive values include:

*Team* - Name

*Score* – Number of point the team received for a game

*Ending* – Final versus overtime (OT)

*Outcome* – Win (W), Loss (L), or Tie (T)

*Date* – Day the game was or will be played

*DaysRest* – Number of days since last game played (null for first games of the season)

*Timezone* – Time zone in which the game was or will be played

*Weather+* – Other weather related information, such as high winds, rain, etc.

*Odds* – Amount by which a team is expected to win/lose

Predictive values include:

*AorH* – Away (A) or Home (H)

*Time* – Time at which the game began (home team local time)

*Weather* – Temperature at game time (or “Dome” if played indoors)

*UpsetAmt* – Magnitude of upset (0 if not an upset)

*Offense* - A representative number for a team’s offense injuries (1 or 2 points for each player injured)

*Defense* – A representative number for a team’s defense injuries (1 or 2 points for each player injured)

Goal values:

*Upset* – Yes (Y) or No (N) if the team predicted to win instead lost

### Limitations

Predictive and goal attributes (the numeric data in the dataset) was normalized into the decimal range 0.00-1.00 using an equal-interval binning method. Through this process, attributes retain their distribution and skew. Normalization also preserves the approximate ratios values within an attribute; however, it does not preserve relationships between attribute values. Hence, it is not possible/meaningful to compare the values for attributes such as whitePerCap against blackPerCap.

The values for the Per Capita Violent Crimes attribute was calculated using population data as well as the sum of crime variables considered violent crimes in the United States: murder, rape, robbery, and assault. A few states, many of which are in the Midwest, do not count rapes in the same manner as the others. This resulted in missing values for rapes and a skewed calculation of Per Capita Violent Crimes. These cities were therefore not included in the final dataset.

Information used for the LEMAS related attributes was limited to police departments with a minimum of 100 officers. A few smaller police departments were used as a random sample.

Communities that could not be found in both census and crime dataset were not included in this dataset. Since many communities in the census were missing LEMAS data, this eliminated those communities, limiting the data selection.

## Objective of Analysis

If changes in crime rates can be predicted based on changes of highly correlated indicators from the community, resources could be better allocated to help local governments in addressing crime occurring in their jurisdiction. The objective of this analysis is to explore other indicators of communities along with its prevalence of violent crime to determine what relationships exist among these indicators and suggest further areas of study. We will use utilize techniques such as classification, regression analysis, frequency item sets, and association rules, to analyze the data set.

## Risks

With a dataset this large, there is a high risk that we will not be able to filter out the “noise” to find a definite answer to our question. There are such a large number of factors contributing to any trends in the data that it may be hard to isolate which is actually the actual cause and what may be red herrings. To be clear, we are aiming to determine correlation and are not confident we will be able to determine causation without a time-lapse analysis. For example, without also utilizing FBI Uniform Crime Report datasets from previous years, we would not be able to analyze how the year-to-year changes in attributes may have resulted in a change to the violent crimes result.

Another risk we may run into is expectations of trends. While people strive to be bias free in their research there are certain ideas that seem intuitive about crime and criminal activity in general. For example, some people believe there is more crime in areas of lower income, but this may not be true among all races, geographical areas, etc. We will need to be diligent in minimizing assumptions and avoiding trying to get the data to match an assumption when no correlation exists. Another significant risk is the accuracy of the data set we will be using. The FBI uses the UCR Statistics; However, that data is first collected by local law enforcement agencies, which may not have the same reporting criteria or standards. This possible discrepancy is a point of discussion among many crime studies.

# RELATED WORK

There has been many studies and papers on the importance of sharing information among law enforcement agencies as well as why the sharing of data is lagging. However, past efforts to solve this issue have failed and the problem still plagues law enforcement. The challenge in the past has simply been how to collect, share, aggregate, and standardize data. This challenge remains today as law enforcement agencies continue to use different standards in reporting crime. The Federal Bureau of Investigation’s Uniform Crime Report data has been the more extensive effort to solve this problem to date. It is not without flaws and oversights, and the interpretation/analysis of this data is highly contentious making all the more difficult for law enforcement agencies to benefit from it. There is also a significant effort to interpret the data into the more standard UCR, which smaller stations may not have the manpower to support. Since there are still large obstacles to overcome in the sharing of data, the analysis of the data becomes cumbersome when looking in several location for the type of data represented in this dataset. Other studies have looked at single types of data or data for a limited number of communities, but not as extensive research could be accomplished for data such as that in this dataset.

# METHODOLOGY

We will begin by taking an exploratory data analysis approach. We will first visualize the data we have gathered and then begin comparing multiple variables. By comparing multiple variables through histograms, scatter plots, and various other visualization trends, we will note any obvious patterns by examining their two-way interactions. We will try to describe this relationship with a type of relationship that best fits. Once we have completed most of our exploratory data analysis, we will begin to perform testing and cross-validation to determine if classification can be successfully completed on the data. Finally, we will attempt to analyze frequent itemsets to obtain more information about trends in the dataset.

## Preprocessing

### Normalization

As previously mentioned, the numeric data in the dataset was normalized before being presented for consumption by others. The attributes retained their distribution and skew as well as preserved the approximate ratios for values within an attribute.

### Binning

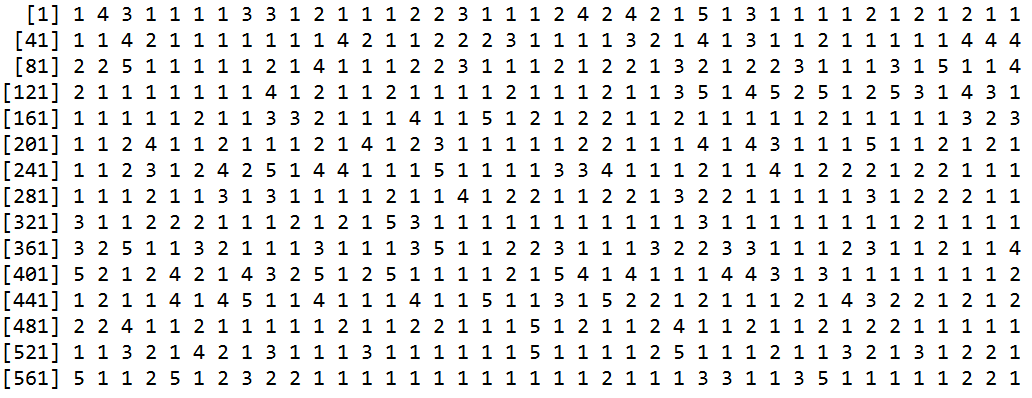
Data binning, sometimes referred to as bucketing, is a technique used in data pre-processing to accommodate algorithms that use categorical rather than continuous variables. Since in our case we wish to not only perform classification, but also frequent itemset analysis, binning will be necessary for the attributes involved. When binning, the values for the attributes are each categorized into a group or “bin” which represents that field. There are four common methods to bin field values: equal width binning, equal frequency binning, binning by clustering, and binning based on predictive value. Since many of the values we are working with are percentages and have been normalized to values between 0.00 and 1.00, we felt the most appropriate technique was equal width.

#### Upset Amount Binning

#### Weather Binning

In equal width binning, the attribute values are divided into *k* categories of equal width. While this is not one of the preferred methods of binning since outliers may influence the width of the bins, we thought it would work well with our data since our attribute values are representative of percentages. We binned the attribute values into groups of 10% (*k* of 10) as well as 20% (*k* of 5). Using the attribute ViolentCrimesPerPop as an example, with *k* of 5 (each bin representing 20%) Table 1 shows a sampling of the post-binned data and Figure 3 shows a histogram of the values binned. In addition, with *k* of 10 (each bin representing 10%) Table 2 shows a sampling of the post binned data and Figure 4 shows a histogram the values binned.

Table 1. Sample of *k*=5 Bin of ViolentCrimesPerPop Data



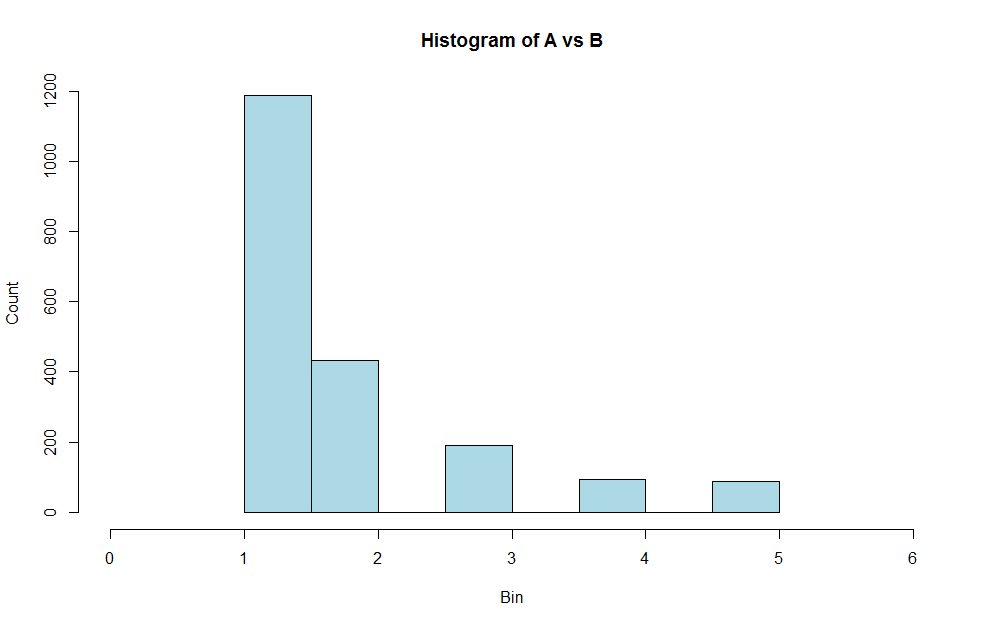
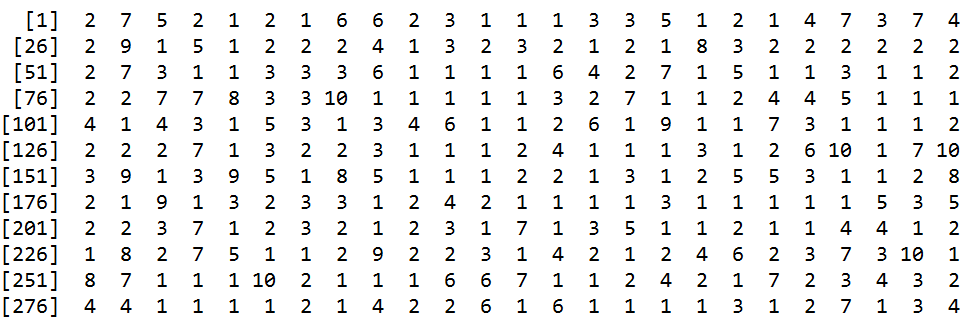


Figure 3. Histogram of Equal Width Binned Weight *k* = 5

Table 2. Sample of *k*=10 Bin of ViolentCrimesPerPop Data



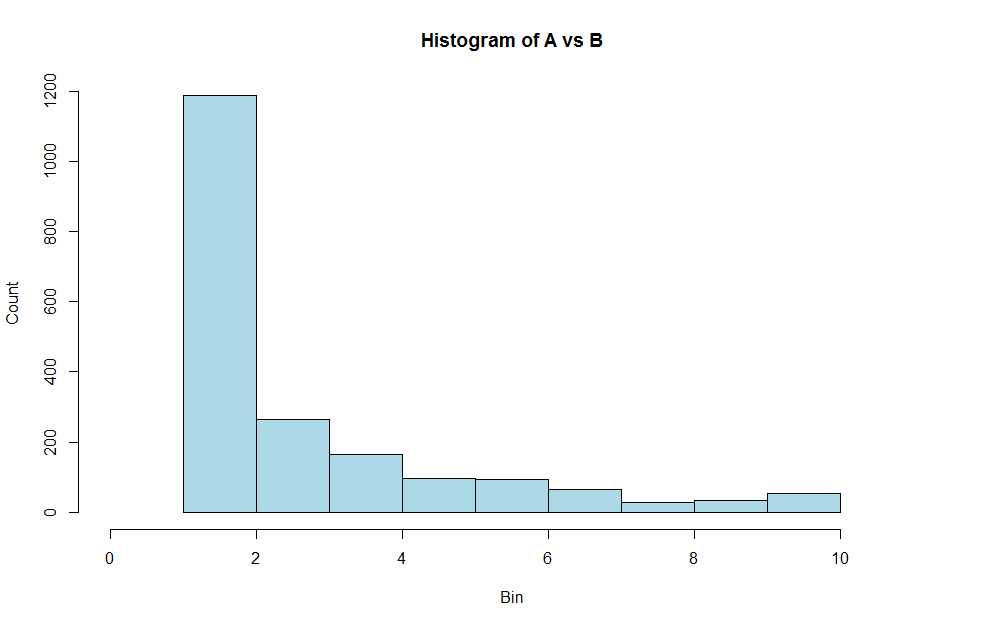


Figure 4. Histogram of Equal Width Binned Weight *k* = 10

## Experiment Design, Tools, & Approaches

### Classification

#### Training and Testing

The Holdout Method is a type of cross validation for classification. In the method, the entire data set is randomly partitioned into two independent sets of specified size: the training set and the test set. The training set is used for model construction and the test set is used to evaluate the accuracy of the constructed model. In the case of our analysis, the data was partitioned with 70% in the training set and 30% in the test set. Both sets contain examples of each classification type.

#### Decision Tree

Decision Tree classification uses a flowchart-like tree structure for classification. In the tree, each test on an attribute is represented by an internal tree node, each outcome of the attribute test is represented by a tree branch, and each tree leaf node has a classification label. A path from the root to a leaf node is a representation of a classification rule.

The tree is constructed in a divide-and-conquer manner with no backtracking. The training examples are all at the root at the start of tree construction and are partitioned reclusively based on the provided selected attributes as the construction proceeds. Partitioning ends when all the samples for a given node belong to the same class, there are no remaining attributes for further partitioning, and there are no samples left to partition.

# EVALUATION METHODOLOGY

## Evaluation Metrics

### Classification

#### Accuracy and Error Rate

Accuracy is calculated as the percentage of test samples correctly calculated (TP is true positive, TN is true negative):

Error rate is calculated as the opposite, or 1 - accuracy (FP is false positive, FN is false negative):

#### Sensitivity and Specificity

Sensitivity is calculated as the true positive (TP) recognition rate:

Specificity is calculated as the true negative (TN) recognition rate:

Accuracy can be written as a function of both sensitivity and specificity:

#### Precision and Recall

There is an inverse relationship between precision and recall.

Precision is measured as a percentage of the samples classified with a positive label that are actually positive, or exactness:

Recall is measured as a percentage of positive samples actually classified with a positive label, or completeness.

A perfect score would be 1.0 or 100%.

#### F-Measures

F-measure is a type of accuracy measurement, which takes into account both precision and recall, with the resulting score assigned is between 0 and 1.

F-measure can also be a weighted measurement as follows:

# RESULTS

## Classification

#### Decision Tree

A Decision Tree was constructed from the training set using binning of *k* of 10 (see Figure 5) the tree seemed simplistic compared to others we had seen. When examining the below Figure 6 of the relative error and complexity point (CP) the simplistic nature of the tree made sense. After the size of the tree reached three, the CP stabilized to a degree with respect to relative error. The resulting Decision Tree had very few leaf nodes despite numerous examples of multiple classification types in the training data. The Decision Tree confusion matrix shown in Table 3 also confirms the trouble the Decision Tree classification had trouble classifying half of the types, showing no classifications of 6 through 9 and few classifications of 10 for any of the test data.

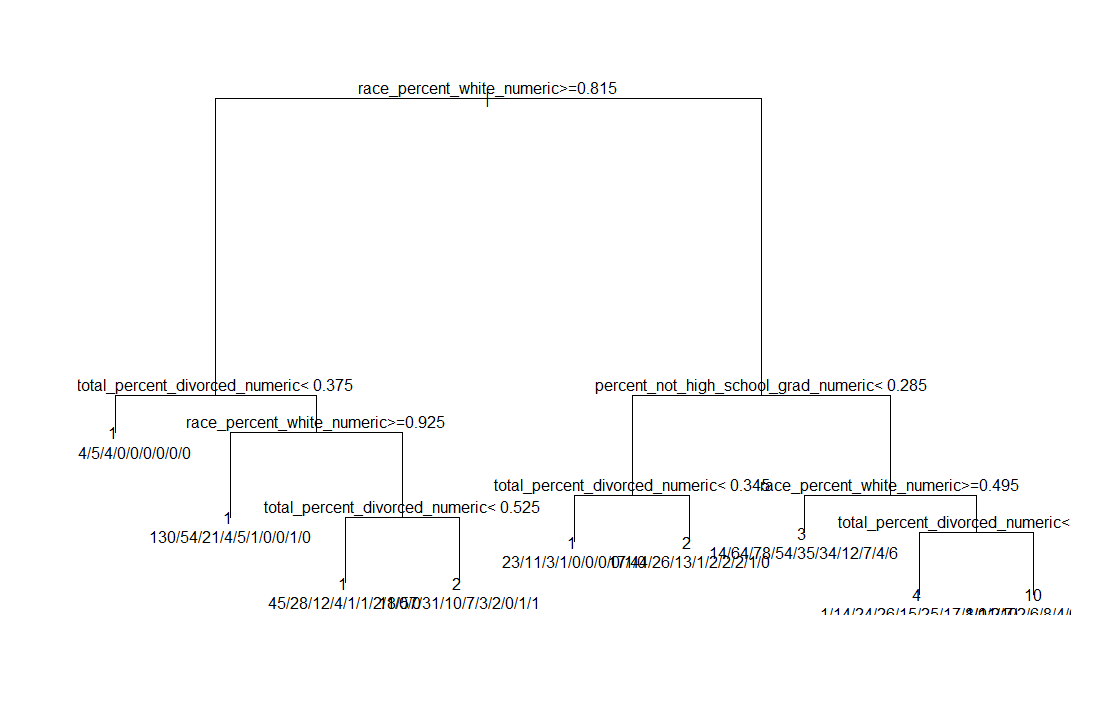


Figure 5. Classification Decision Tree – Binning *k* = 10

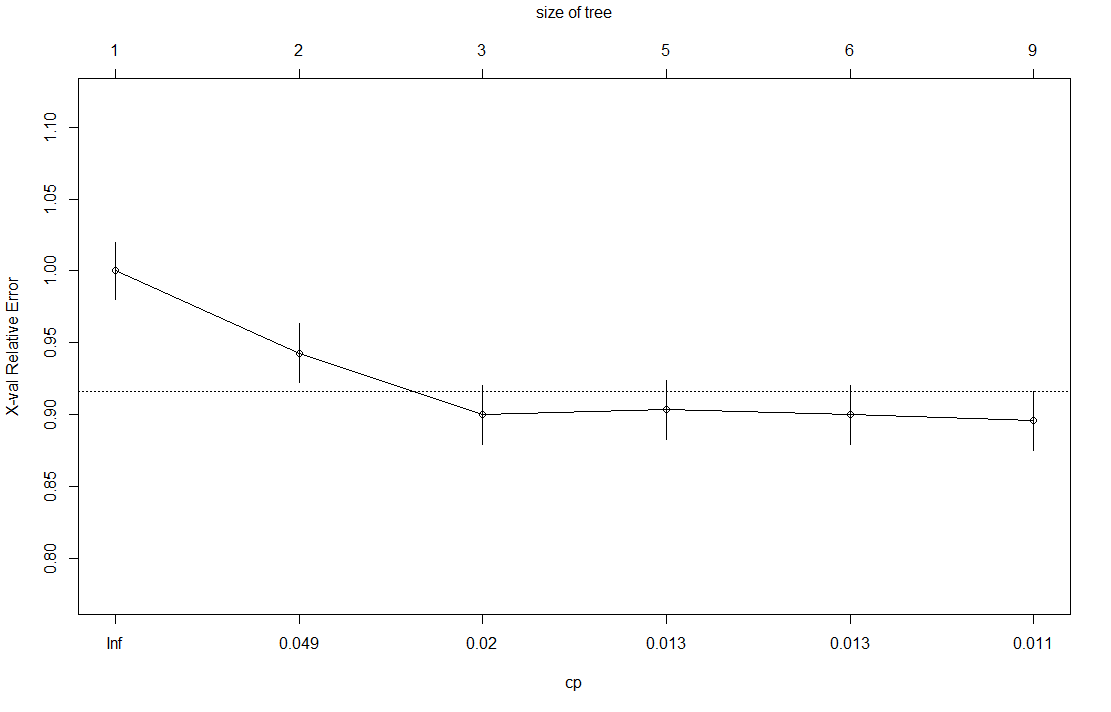
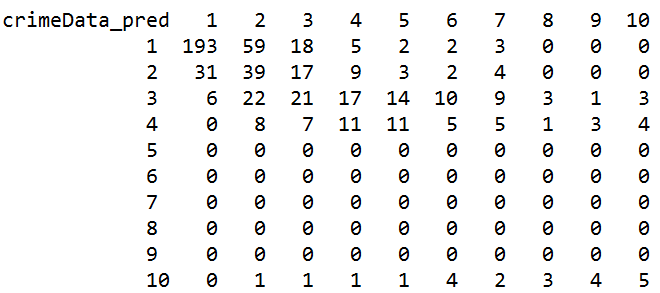


Figure 6. Relative Error and CP – Binning *k* = 10

Table 4. Confusion Matrix Results – Binning *k* = 10



Another Decision Tree was constructed from the training set using binning of *k* of 5 (see Figure 7) the tree again seemed simplistic. When we examined graph of the relative error and CP (Figure 8) the simplicity of the tree again made sense. After the size of the tree reached three, the CP stabilized with respect to relative error. The resulting Decision Tree had less leaf nodes than expected despite numerous examples of multiple classification types in the training data. The Decision Tree confusion matrix shown in Table 4 also confirms the trouble the Decision Tree classification had difficulty classifying many types, showing no classifications of 3 or 4 and few classifications of 5 for any of the test data.

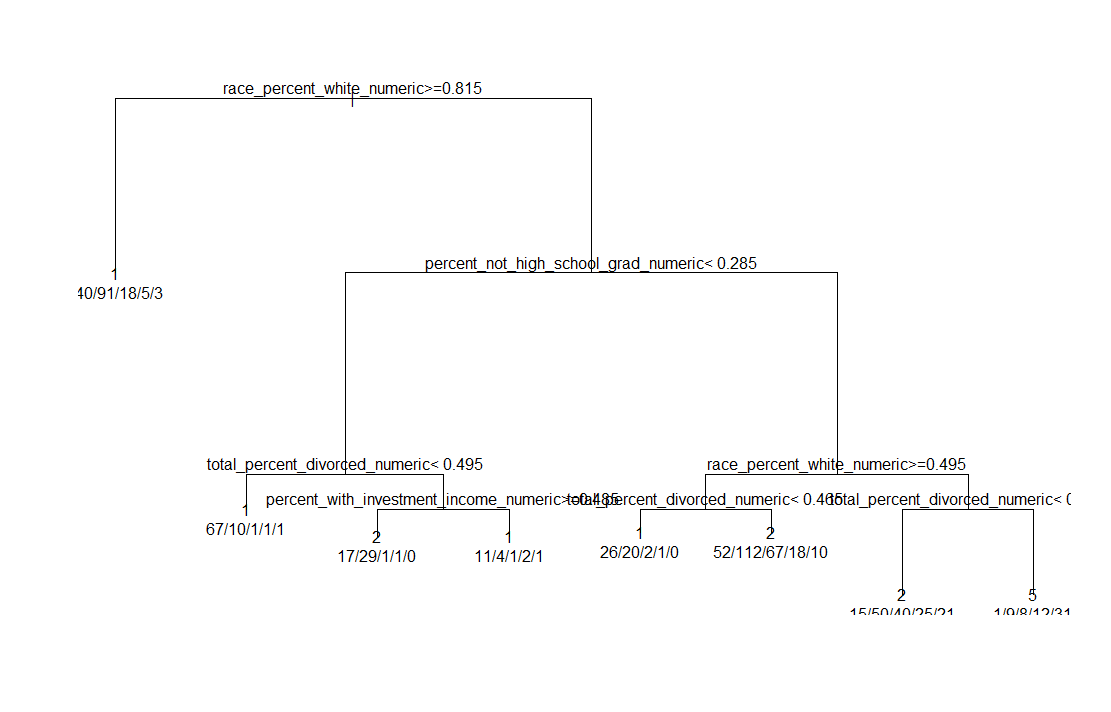


Figure 7. Classification Decision Tree – Binning *k* = 5

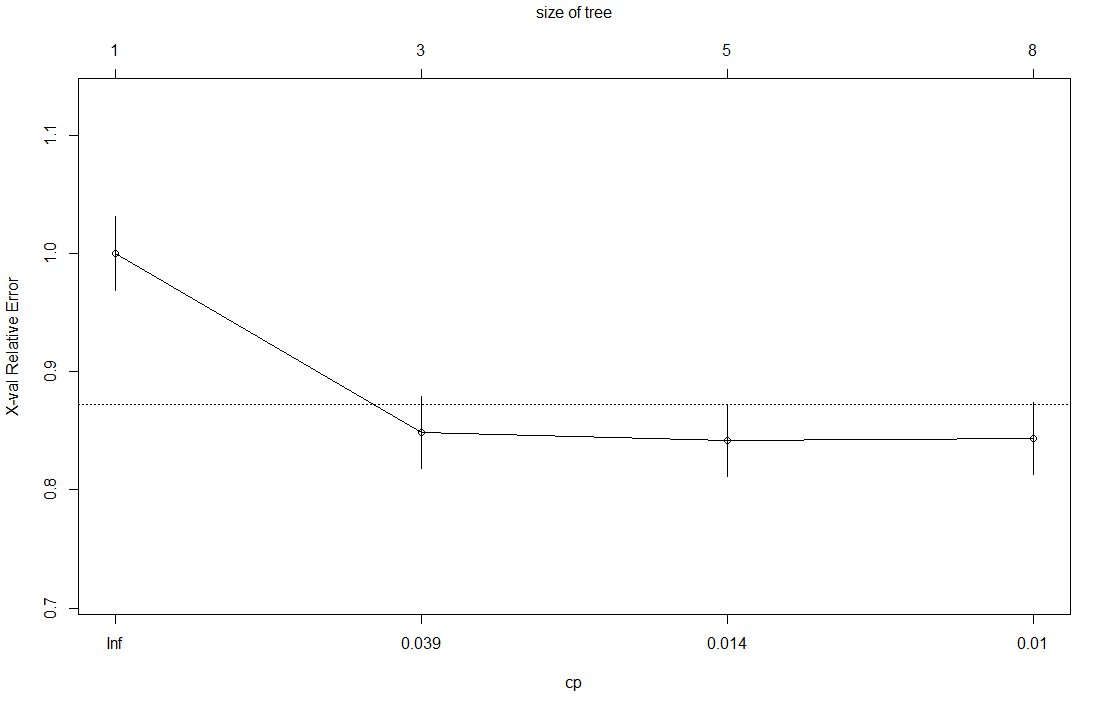
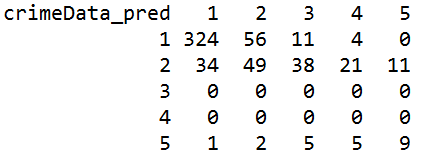


Figure 8. Relative Error and CP – Binning *k* = 5

Table 5. Confusion Matrix Results – Binning *k* = 5



#### Accuracy and Error Rate

The overall accuracy for the *k* of 10 binned data could be calculated using Table 3 as follows:

(193 + 39 + 21+ 11 + 0 + 0 + 0 + 0 + 0 + 5) / 570

= 0.467 ≈ 47%

The overall error rate for the *k* of 10 binned data could be calculated using Table 3 as follows:

(37 + 90 + 43 + 32 + 31 + 23 + 23 + 7 + 8 + 7) / 570

= 0.528 ≈ 53%

With an error rate greater than the accuracy, we determined it was futile to continuing to calculate the rest of the classification evaluation metrics for *k* of 10 binned data. As inaccurate and error prone as the *k* of 10 binned data was obviously at determining classification, it would not provide us useful results. However, we continued the evaluation with the *k* of 5 binned data.

The overall accuracy for the *k* of 5 binned data could be calculated using Table 4 as follows:

(324 + 49 + 0 + 0 + 9) / 570 = 0.67 ≈ 67%

The overall error rate for the *k* of 5 binned data could be calculated using Table 4 as follows:

(35 + 58 + 54 + 30 + 11) / 570 = 0.33 ≈ 33%

With an accuracy near 2/3 and an error rate closer to 1/3, the results for *k* of 5 binned data were much better than those of the *k* of 10 binned data.

#### Sensitivity and Specificity

The sensitivity for each value for *k* of 5 binned data classification can be calculated using Table 4 as follows:

1: 324 / 395 = 0.82 ≈ 82%

2: 49 / 103 = 0.32 ≈ 32%

3 & 4: 0 / 0 so they cannot be calculated

5: 9 / 22 = 0.41 ≈ 41%

The specificity for each value for *k* of 5 binned data classification can be calculated using Table 4 as follows:

1: 140 / 175 = 0.8 ≈ 80%

2: 359 / 417 = 0.861 ≈ 86.1%

3: 516 / 570 = 0.905 ≈ 90.5%

4: 540 / 570 = 0.95 ≈ 95%

5: 537 / 570 = 0.98 ≈ 98%

#### Precision and Recall

The precision for each value for the *k* of 5 classification can be calculated using Table 4 as follows:

1: 324 / (324 + 106) = 0.75 ≈ 75%

2: 49 / (49 + 162) = 0.23 ≈ 23%

3: 0 / (0 + 54) = 0%

4: 0 / (0 + 30) = 0%

5: 9 / (9 + 24) = 0.273 ≈ 27.3%

The recall for each value for the *k* of 5 classification can be calculated using Table 4 as follows:

1: 324 / (324 + 71) = 0.82 ≈ 82%

2: 49 / (49 + 104) = 0.32 ≈ 32%

3: 0 / (0 + 0) so it cannot be calculated

4: 0 / (0 + 0) so it cannot be calculated

5: 9 / (9 + 13) = 0.41 ≈ 41%

#### F-Measures

The F-measure for each value for the *k* of 5 classification can be calculated using Table 4 as follows:

1: (2 \* 0.75 \* 0.82) / (0.75 + 0.82) = 0.78 ≈ 78%

2: (2 \* 0.23 \* 0.32) / (0.23 + 0.32) = 0.27 ≈ 27%

3: cannot be calculated since recall could not be calculated

4: cannot be calculated since recall could not be calculated

5: (2 \* 0.828 \* 0.41) / (0.828 + 0.41) = 0.33 ≈ 33%

# DISCUSSION

We had originally hypothesized that factors such as percentage of police per capita, per capita income, and percentage of population with Bachelor’s degrees or higher would be significant in terms of violent crimes per capita. It turned out that these are indirectly related to the attributes that we found most significant.

Also in our original hypothesis was the percentage of police per capita. This too was a poor predictor of the incidents of violent crime per capita. We believe this result is biased in that the communities with police departments that had less than 100 police officers were not included in the data set minus a few randomly sampled examples.

Investment income was one of the attributes we determined to a factor in violent crimes per capita. It is indirectly related to per capita income in that investment income is usually thought of as the difference between income and expenses. It is also common for areas with high incomes per capita to be areas with a very high cost of living. Therefore, if we were to use purely per capita income the perceived wealth/income would be skewed in this scenario. Investment income is a better indicated of how well a community is doing relative to its cost of living.

In addition, there was not a significant separation between communities with a more educated resident base beyond a high school degree. It is possible that a high school degree was sufficient for a comfortable standard of living in the 1980s and 1990s. It would be interesting to explore this today and see if the threshold has increased to a bachelor’s degree.

# CONCLUSIONS

In our analysis, we saw a slight proportional relationship between high violent crime per capita and high percentage of divorcees as well as a high percentage of the population without a high school degree. We also saw a slight inverse proportional relationship between high violent crime per capita and communities with higher percentage of investment income as well as largely Caucasian communities.

When there was lower percentage of the population without a high school degree, a lower percentage of divorcees and higher percentage of Caucasians were found frequently. In addition, a higher percentage of Caucasians in the community was frequently found associated with a higher percentage of the population with investment income. High percentages of Caucasians in a community were also many times found with low to mid percentages of divorcees in the community.

The error rate for the classification was very high. There were also very few results in our frequent itemset analysis with the low percentage of violent crime per capita the only bin producing results. Even with the large size of the dataset, the fact that there was so much error and so little results does not give us confidence that the attributes we selected do have a strong case for influencing violent crime per capita.

We feel it would be useful to extend this study to analyze how change in any specific attribute is correlated with an increase or decrease in violent crime per capita. This would require additional datasets in preceding and/or succeeding years. This additional data would also help in determining outliers in the data, which could be used to improve the results of any analysis.

# REFERENCES

1. Larose, D. and Larose, C.D. *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley-Interscience, 2014.
2. Anon. 2014 NFL Weekly League Schedule | Pro-Football-Reference.com. Retrieved August 20, 2016 from <http://www.pro-football-reference.com/years/2014/games.htm>
3. Anon. 2012 Arizona Cardinals season. Retrieved August 20, 2016 from <https://en.wikipedia.org/wiki/2012_Arizona_Cardinals_season>
4. USAToday. Week-by-week 2013 NFL schedule (2013). Retrieved September 20, 2016 from <http://www.usatoday.com/story/sports/nfl/2013/04/18/week-by-week-2013-nfl-schedule/2093613/>
5. Gray, J. NFL schedule 2014: Week by Week. (2014). Retrieved September 20, 2016 from <http://www.sbnation.com/nfl/2014/4/23/5576528/nfl-weekly-schedule-2014>
6. Hirschhorn, J. B. 2015 NFL schedule released. (2015). Retrieved September 20, 1016 from <http://www.sbnation.com/2015/4/21/8341221/2015-nfl-schedule-released-seahawks-patriots-eagles-broncos>
7. Anon. Archived Closing NFL Odds, NFL Lines, NFL Point Spreads. Historical Pro Football: 2006 – Current. Retrieved September 20, 2016 from <http://www.footballlocks.com/archived_nfl_odds_lines_point_spreads.shtml>
8. Anon. National Football League Game Summary. Retrieved September 20, 2016 from <http://www.nfl.com/liveupdate/gamecenter/56954/HOU_Gamebook.pdf>
9. Anon. Weather Forecast & Reports – Long Range & Local | Wunderground | Weather Underground. Retrieved September 20, 2016 form <https://www.wunderground.com/>
10. Anon. 2011 Minnesota Vikings injuries | Pro-Football-Reference.com. Retrieved October 20, 2016 from <http://www.pro-football-reference.com/teams/min/2011_injuries.htm>
11. Anon. 2010 Minnesota Vikings Starters, Roster, & Players | Pro-Football-Reference.com. Retrieved October 20, 2016 from <http://www.pro-football-reference.com/teams/min/2010_roster.htm>
12. Mitchell, R. L. It’s criminal: Why data sharing lags among law enforcement agencies, *Computerworld*. Retrieved February 28, 2016, from Computerworld: http://www.computerworld.com/article/2486359/government-it/it-s-criminal--why-data-sharing-lags-among-law-enforcement-agencies.html