What Influences the Difficulty of New York Times Crossword Puzzles?

Walter Bennette, Yuanfeng Cai, Dave Osthus and Takisha Watson

Stat 503

April 25, 2011

1 Description and Objectives

Crossword puzzles were first introduced to the readership of the New York Times after the bombing of Pearl Harbor in 1941 when editors of the newspaper deemed the game a useful distraction to the war stricken population. In the years following the first crossword puzzle the game became a staple of the newspaper, and to date, has a rich following of many active puzzlers. Editor Will Shortz intimated to the New York Times' readership in 2001 that, "The perfect level of [puzzle] difficulty, of course, differs from person to person. This is why, as editor, I vary the weekday Times crossword difficulty from easy-medium on Monday up to what the actor and puzzle aficionado Paul Sorvino calls 'the bitch mother of all crosswords' on Saturday. (He said this as a compliment.)" Therefore, this data mining project is being undertaken to further understand the New York Times crossword puzzle, and to gain insight as to what drives the difficulty level of puzzles.

The primary question to be answered by this data mining exercise is "What is the main indicator of puzzle difficulty for the different days of the week?" Crossword puzzle constructor A. J. Santora has been quoted as saying, "You can take a simple puzzle with ordinary words and search in libraries for clues that can make a puzzle extraordinarily difficult", but can all of the difficulty of a particular crossword be simply from the wording of the puzzle's clues? As a group, we believe that the difficulty of a crossword puzzle may also lie in the composition of the puzzle's answers, meaning that puzzles may include longer and rarer answers as the week progresses.

With the exception of Sunday, New York Times' crossword puzzles are built in a 15 cell by 15 cell grid. Puzzles on Monday, Tuesday, Wednesday and Thursday are allowed a maximum of 78 answers. Puzzles on Friday and Saturday are allowed a maximum of 72 answers, an indication that the Friday and Saturday puzzles are more difficult because of the inclusion of clues with longer answers. These maximum answer totals are not the cut and dry rule, but rather strongly suggested. The Sunday puzzle is built on a larger grid, frequently 23 cells by 23 cells, and is designed to have the difficulty of a Thursday puzzle but with more clues. Are there reasons in addition to the number of clues and clue difficulty that make a particular crossword puzzle more difficult?

To answer the above question and the primary question posed earlier, "What is the main indicator of puzzle difficulty for the different days of the week?", we assume that the difficulty of the crossword puzzle does increase as the week progresses, and therefore, attempt to build a classifier to classify the puzzles into one of three categories. The first category will be easy puzzles and will contain the puzzles of Monday, Tuesday, and Wednesday. The second category will be puzzles of medium difficulty and this group will contain the puzzles for Thursday and Sunday. Then finally, the last group will be the hard puzzles and will contain the puzzles found in the Friday and Saturday edition of the newspaper. We have collected data from www.xwordinfo.com to describe the rudimentary components of the New York Times' crossword puzzles for all of 2010. The same information was collected for the 2011 crossword puzzles, between January 1st and April 14th. This will be treated as our test data set.

A listing of the variables are shown below. The percentage variables, along with the variables in italics, were created by us. Percentage variables are created by dividing the individual amounts by the total puzzle amount. For example, the percentage of word lengths equal to three is the number of three letter words divided by the total number of puzzle words for a particular crossword puzzle.

- **Blocks:** Number of blocks. A block is a cell in the crossword puzzle grid that cannot contain a letter.
- **Letter Count:** Letter count
- **Word Count:** Word Count
- **A-Z:** The number/percentage for each letter in each puzzle (26 variables)
- **X3-X27:** The number/percentage of words of length 3 to 27 in each puzzle (25 variables)
- ***BvWC:*** Words per block. Calculated as the total words in a puzzle divided by the number of blocks in that puzzle.
- ***AWL:*** Average Word Length
- ***Vowels_y:*** The number/percentage of letters that are A, E, I, O, U or Y

- ***Vowels:*** The number/percentage of letters that are A, E, I, O or U
- ***Consonants_y:*** The number/percentage of letters that are consonants, including Y
- ***Consonants:*** The number/percentage of letters that are consonants, excluding Y
- ***Rare:*** The number/percentage of letters including J, Q, X and Z
- ***Common:*** The number/percentage of letters excluding J, Q, X and Z

It is believed that if we can build accurate classifiers from this available information, then we can surely gain partial insight into what drives the complexity of a crossword puzzle. This leads us to the goals of our analysis.

Goals of Data Analysis:

- Understand why a puzzle becomes more difficult aside from the number of answers required to solve the puzzle.
- Confirm claims made by the New York Times in regards to puzzle structure and composition.

  - Claim 1: Do our findings suggest that puzzles do indeed increase in difficulty throughout the week?

  - Claim 2: Do Thursday and Sunday puzzles have similar difficulty, or at least components?

2 Suggested Analysis

The summary of our analysis is shown in Table **??**.

3 Results

3.1 Summary Statistics

There are 365 crossword puzzles in our training set (2010 puzzles) and 104 crossword puzzles in our test set (2011 puzzles). In the training data set, there are either 52 or 53 puzzles for each day of the week. In the testing data set, there are either 14 or 15 puzzles for each day of the week. Summary statistics can be found in Table **??**.

Summary Statistics Findings:

- Every puzzle had an A, D, E, I, L, M, N, O, R, S, and T.
  - 50% of the puzzles had no J, Q and Z.
  - The median for X was 1, so 50% of the puzzles had at least 1 X.
- Letter count for Sunday is much hinger than all of the other days. This is due to the larger sized grid on which the puzzle is built. The letter counts vary between the other days of the week.
- Word counts are similar for Monday through Wednesday, there is a drop in word count from Thursday through Saturday, and a drastic increase for Sunday.

Approach	**Reason**	**Type of questions addressed**
Data Restructuring: Make % of letters in puzzle and % word lengths in puzzle. Make variables for vowels, consonants, rare letters, common letters, average word lengths and words per block.	Sunday crossword puzzles are larger than the other days of the week - this is not a difference we want to consider when determining what makes crossword puzzles different. By making percentage variables, we remove the differences in crossword puzzle size. The indicator variables allow us to potentially reduce the dimensionality of our data.	
Summary Statistics	Extract location and scale information.	"Do crosswords have the same number of words all throughout the week?" "Do more consonants get used in puzzles later in the week?"
Histograms	Explore univaraite distributions of block counts and average word lengths.	"Do word lengths get longer as the week progresses from Monday to Sunday?" "Are there more words per block in puzzles from earlier in the week than later?"
Line Plots	Explore how the distribution of letters and word lengths change by day.	"Does the distribution of letters change over the week?" "Does the distribution of word lengths differ by day?"
Box Plots	Explore the relationship between day of week and other quantitative variables.	"Do puzzles early in the week have lower average word lengths than puzzles later in the week?"
Classification Trees, SVM, Random Forests, LDA, QDA . . .	Classify crossword puzzles into day of week. The claim is that crossword puzzles become more difficult throughout the week. Is this because the word length or word composition changes? If they do, classification should be able to pick up on these systematic differences.	"Is puzzle composition different as the week progresses, i.e. word lengths?" "Can we identify the day of week the crossword puzzle comes from based on the collected rudimentary variables?"

Table 1: Suggested Analysis

—Blocks—	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Min	32	32	32	32	19	18	66
Median	37	36	36	37	30	29	72.5
Maximum	43	42	42	48	42	38	96
Mean	36.96	37	36.85	37.08	30.28	29.5	74.23
SD	2.22	2.08	2.43	3.35	4.44	3.91	6.63
—Letter Count—							
Min	182	176	178	177	182	187	343
Median	188	189	189	188	195	196	369
Maximum	193	193	202	206	206	218	448
Mean	188	187.7	188.4	188.2	194.5	196	374
SD	2.22	2.65	4.11	5.42	4.75	5.04	20.49
—Word Count—							
Min	74	70	72	62	58	56	136
Median	77	78	76	74	70	70	140
Maximum	78	78	81	81	76	72	171
Mean	76.92	76.9	76.38	74.79	68.77	68.25	142.6
SD	1.15	1.93	1.81	3.57	3.48	3.77	8.91
—Rare—							
Min	0	0	0	0	0	0	0
Median	2	1	2	2	2	1.5	2
Maximum	10	7	7	17	9	9	9
Mean	1.92	1.64	2.19	2.87	2.32	2.192	2.89
SD	1.74	1.53	1.78	3.04	2.25	2.33	1.90
—Consanants (%)—							
Min	0.56	0.56	0.55	0.54	0.56	0.56	0.57
Median	0.59	0.59	0.59	0.59	0.59	0.59	0.59
Maximum	0.63	0.63	0.65	0.62	0.64	0.63	0.62
Mean	0.59	0.59	0.59	0.59	0.59	0.59	0.59
SD	0.02	0.02	0.02	0.02	0.02	0.02	0.01

Table 2: Summary statistics: The number of blocks are similar for Monday through Thursday. Sunday has the highest letter and word count compared to the other days of the week; this is most likely due to the size of Sunday puzzles. The maximum number of rare letters varies between the different days of the week. The majority of the days have a maximum of 7 or 9 rare letters. The percentage of consonants is roughly the same for everyday of the week.

- The median number of rare letters is between 1 and 2 throughout the week. Thursday has the highest number of rare letters in a puzzle with 17. The majority of the days have a maximum number of rare letters of 7 or 9.
- Every puzzle has more consonants than vowels. The mean percentage of constants in the same for each day of the week.

3.2 Histograms

Histograms were created to explore the univariate distribution of multiple varaibles and are shown in Figures **??** and **??**.

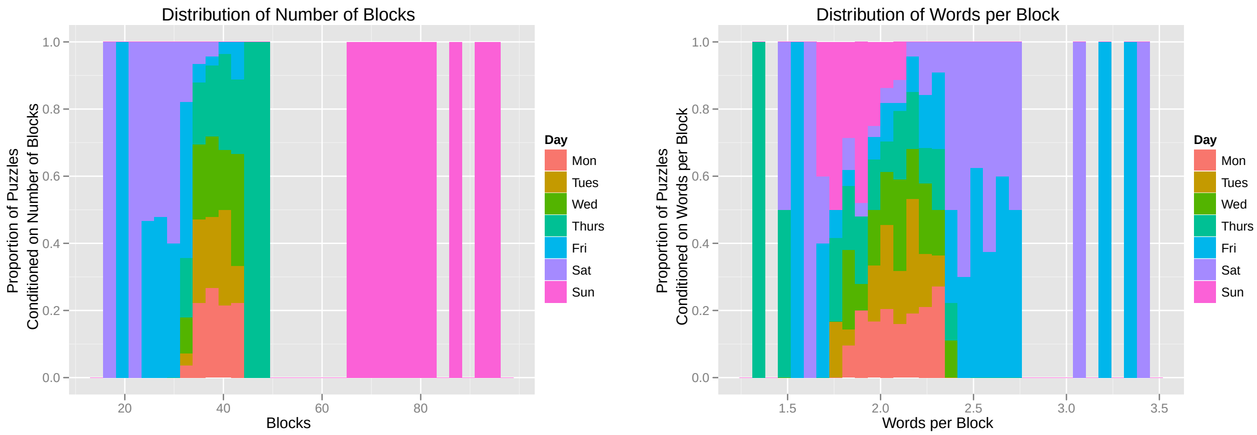

Figure 1: Left: Histogram of blocks. There are obviously more blocks on Sunday (pink) compared to those on any other days. Right: Histogram of words per block. Friday (blue) and Saturday (purple) have relatively large values for BvWC.

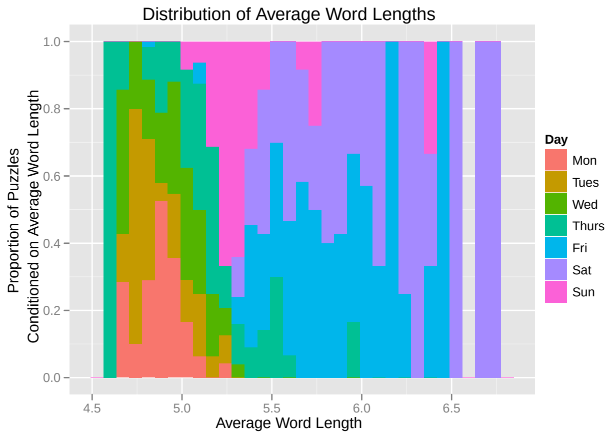

Figure 2: The histogram of average word length (AWL) shows that proportion of the days of the week that have an AWL for the values between 4.5 and 7. Saturday and Friday have the largest AWL values while Monday, Tuesday, and Wednesday have low AWL values.

Summary of Histogram Findings:

- Using counts makes it very easy to separate Sunday puzzles from all other days, by virtue of its size. This is not information we want to use, so the percentage data will be used instead of counts throughout the analysis.

- Average word lengths are higher for Friday and Saturday than all other days.
- We get a sense that days seem to be cluster in the words per block measurement. We noticed a group of similar values for Monday, Tuesday, and Wednesday, a group for Thursday and Sunday, and then a final group for Friday and Saturday. This supports our decision to classify into these three groups of days, rather than days themselves.

3.3 Line Plots

Figures **??** and **??** show the distribution of letters and word lengths of crossword puzzles by day of week.

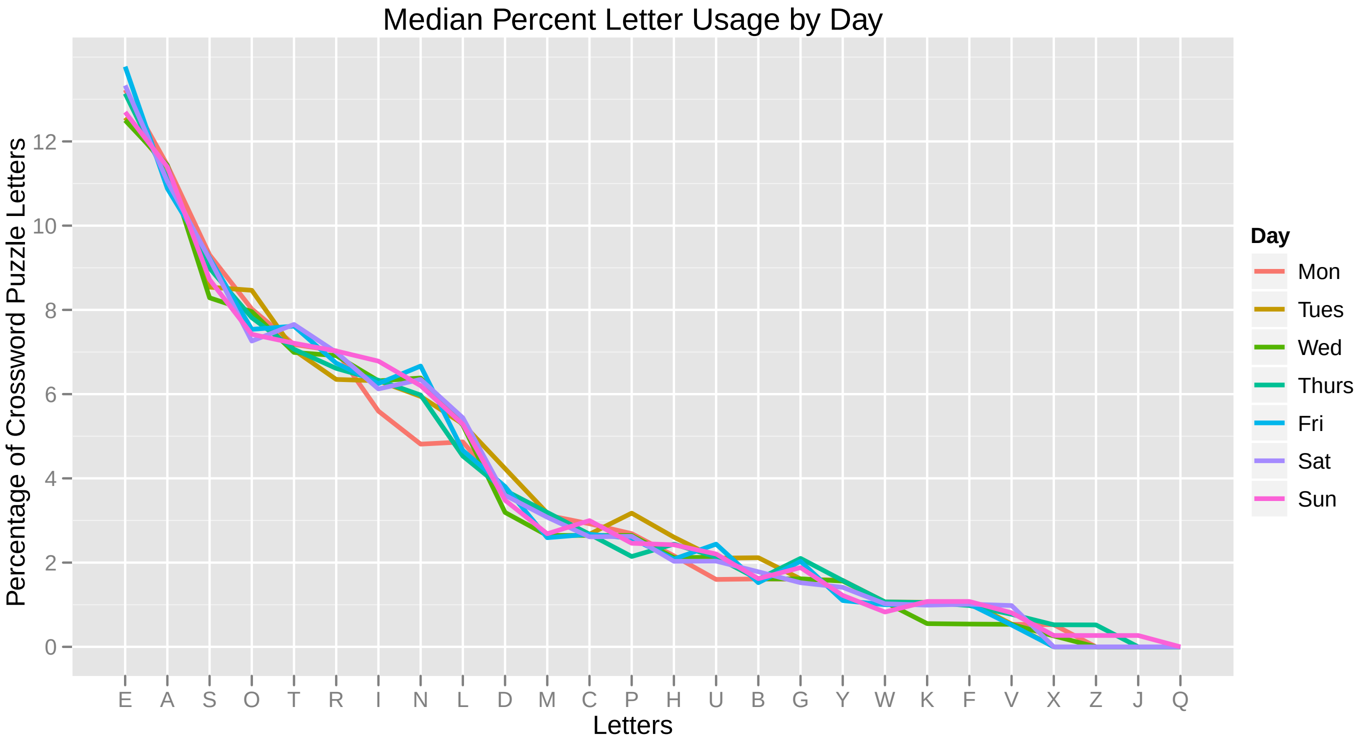

Figure 3: Visual for the median usage of letters by day.

Line Plot Findings:

- The most common letters are, in order of prevalence, E, A, S, O, T, R, I, N and L.
- There is little difference in median letter percentages between day.
- There appear to be three groups of days, especially when considering four letter words.
- Thursday and Sunday puzzles do appear to be similar in terms of word length distributions, as suggested by the editor of the New York Times.

3.4 Boxplots

Figure **??** shows some of the more interesting letter distribution distinctions by day. Figure **??** shows some of the most telling differences between days of the week, in terms of average word length, percent vowels, and words per block.

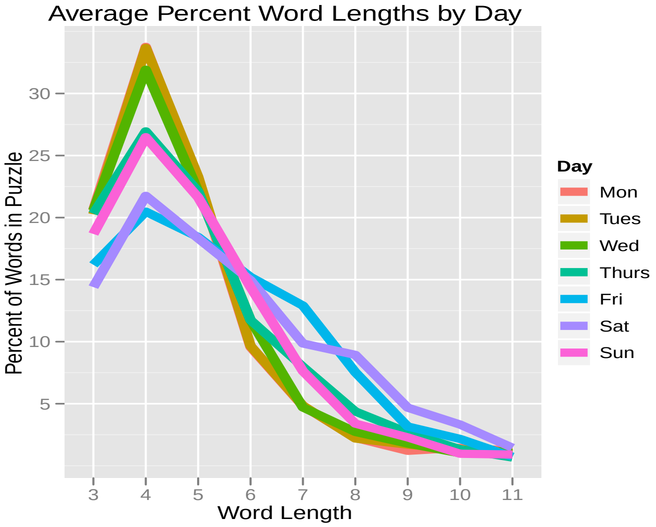

Figure 4: Average percent of word lengths between 3 and 11 letters, grouped by day. There appear to be three groups of days. Monday through Wednesday are all similar, with high rates of shorter words (3-5 letters), Friday and Saturday have lower rates of short letter words but high rates of 7-9 letter words. Thursday and Sunday are similar in that they are middle of the road in both categories.

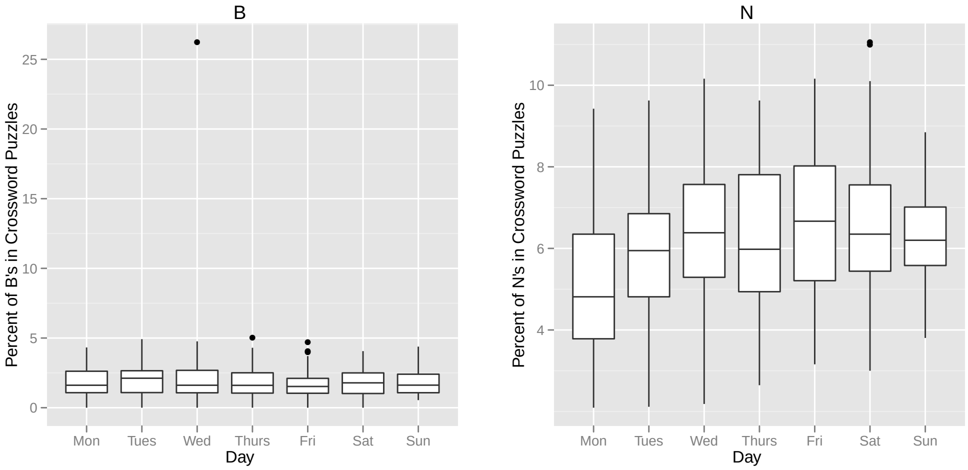

Figure 5: Percentage of the usage of the letters B and N by day of week. B: There was a 'B' themed puzzles that actually did have around 25% of all letters as 'B's. N: There seems to be an increasing usage of the letter 'N' as the week progresses.

Box Plot Findings

- There appears to be little difference in the distribution of letters between days, indicating they may not be very useful when it comes to classifying days.
- Average word length does increase throughout the week, if we exclude Sunday.
- In terms of average word length, Sunday is most similar to Thursday, as indicated by the crossword editor.
- There is a difference in words per block between Friday and Saturday compared to all other day.
- The percentage of vowels including 'Y' seems to decrease throughout the week, though only slightly and with a lot of overlap.

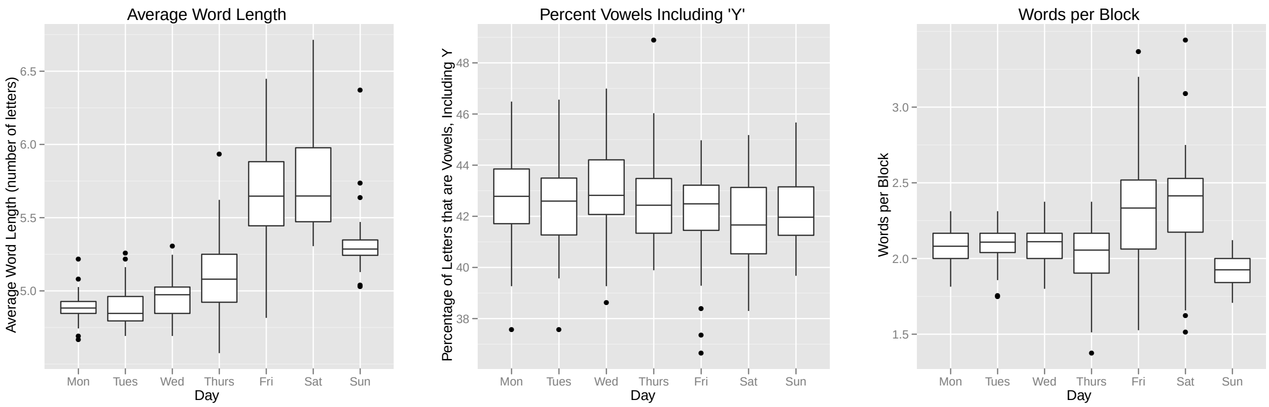

Figure 6: Left: Average word length increases from Monday to Saturday, then falls off on Sunday to Thursday levels. Middle: Percent vowels including ‘Y’ gradually decreases from Monday to Sunday, but the scale is fairly tight. Right: Words per block are constant between Monday and Thursday, shoot up for Friday and Saturday, then fall to their lowest levels on Sunday.

4 Classification Results

After many failed attempts to classify crossword puzzles into one of seven days, we decided a more reasonable goal would be to classify crossword puzzles into three groups of days: Monday through Wednesday, Thursday and Sunday, and Friday and Saturday. The training and test accuracy from our classification methods are shown in Table **??**.

Five approaches were used to classify the data into the three difficulty levels. Linear discriminate analysis (LDA), quadratic discriminate analysis (QDA), multinomial regression, a decision tree, and support vector machines (SVM). Multinomial regression is an extension of logistic regression that allows for the prediction of more than two class values. The best classification method will be used to find the important variables to distinguish between difficulties, but finding these variables may be a complicated task, so interpretability will be taken into consideration when choosing a best classification method.

Method	**Training Accuracy (%)**	**Test Accuracy (%)**	**Attributes Used**
LDA	80.8	74.0	X4, X5, X6, BvWC, AWL, Rare
QDA	85.5	82.7	X4, X5, X6, BvWC, AWL, Rare
Multinomial Regression	84.7	77.9	X4, X5, X6, BvWC, AWL, Rare
Decision Tree	88.2	82.7	X3, X7, X8, BvWC, Rare, AWL
SVM	84.7	76.9	X3, X4, X7, BvWC, AWL

Table 3: Accuracy of training and test data set.

The models shown in Table **??** have been built from intelligently selected subsets of the following percent variables; N, S, X3, X4, X5, X6, X7, X8, X9, Vowels, and Rare, as well as BvWC and AWL. The variables selected for the multinomial regression model were identified using a backwards stepwise procedure where AIC was utilized as the selection criteria. This subset of variables was also used for LDA and QDA. For the model selection of SVM, we randomly selected 2000 of the possible $2^{13} = 8192$ combinations of variables and calculated the accuracy of the SVM on the training data set. The combination of variables that resulted in the highest accuracy on the training data set was selected as our best SVM model. Finally, all variables were allowed to be included in the construction of the decision tree, but only a select few were actually utilized, as decided by the decision tree algorithm. It is interesting to see that all of the methods chose to include the structural information of the puzzle shown in words per block, the average word length, and that most included the frequency of rare letters and some combination of the percents of short and medium word lengths. One interesting result of Table **??** was the poor relative performance of the SVM classifier. Through our experience, SVM has almost always been one of the most successful classifiers, but not in this case. One reason for this may be due to the quality of tuning of this SVM classifier to this particular data set. It is possible that a different kernel, i.e. radial rather than polynomial, would yield better results. We do not make the claim that that the classifiers in Table **??** are the best possible classifiers. They are, however, the best of the ones we examined. Given that interpretability is high on our requirements wish list of a classifier, unless SVM produces significantly better accuracy, it is not guaranteed that the SVM method would be considered "best."

QDA and the decision tree models have the best accuracy when predicting the testing data, each with an accuracy of 82.7%. Unique to the construction of the decision tree was the use of an instance selection procedure to help identify "useful" crossword puzzles from which to build the tree, resulting in a simple and general decision tree. The two top classification methods most frequently confuse the category relating to the puzzles of medium difficultly from Thursday and Sunday, as can be see in Figure **??**. Both methods utilize the variables of words per block, average word length, the occurrence of rare letters, and the percentages of different length words, but neither are able to obtain a perfect separation of the crossword puzzles. Obviously the selected variables are helpful in separating the puzzles, but some information is missing. We believe that some of this missing information is contained in the complexity of the puzzle's clues. Our selected variables are, however, at least part of the reason for the different difficulty levels of crossword puzzles.

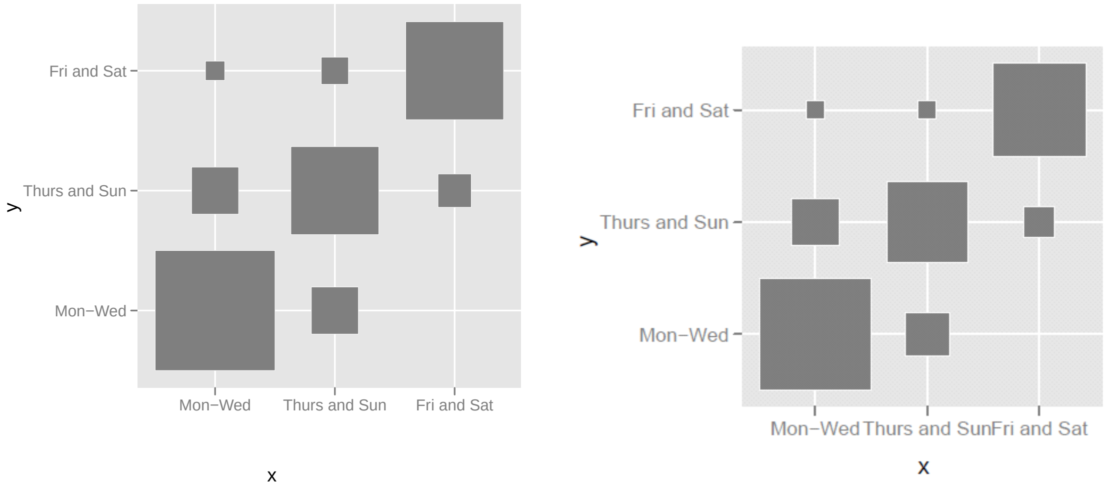

Figure 7: Accuracy of QDA (left) and Decision Tree (right). The actual groups are on the y-axis; the predicted, on the x-axis. We see heavy diagonals, indicating both classifiers did a good job. In both puzzles, no Monday-Wednesday puzzle was classified as a Friday or Saturday. Also with both classifiers, the most difficulty was in classifying Thursday or Sunday puzzles.

Viewing the decision tree displayed in Figure **??** gives good insight to the difficulty drivers of different crosswords throughout the week. The decision tree stresses the importance of average word length in determining the day from which the crossword puzzle belongs, and relays the notion that the puzzles from later in the week have longer answers. The structure associated with the average word length makes sense, in that puzzles with longer answers would be more difficult to solve, and thusly are from later days of the week. The final decision tree also utilizes the percent of rare letters to distinguish between different days of the week and a lower percent of rare letters lead to earlier in the week puzzles. The rare letter structure is intuitive, and indicates that the use of rare letters increases the difficulty of the puzzle. Finally, in one branching, the decision tree incorporates the measure of words per block into its structure, the higher value of words per block indicates more work for the puzzler, and the high end of this split leads to later in the week, or more complicated, puzzles.

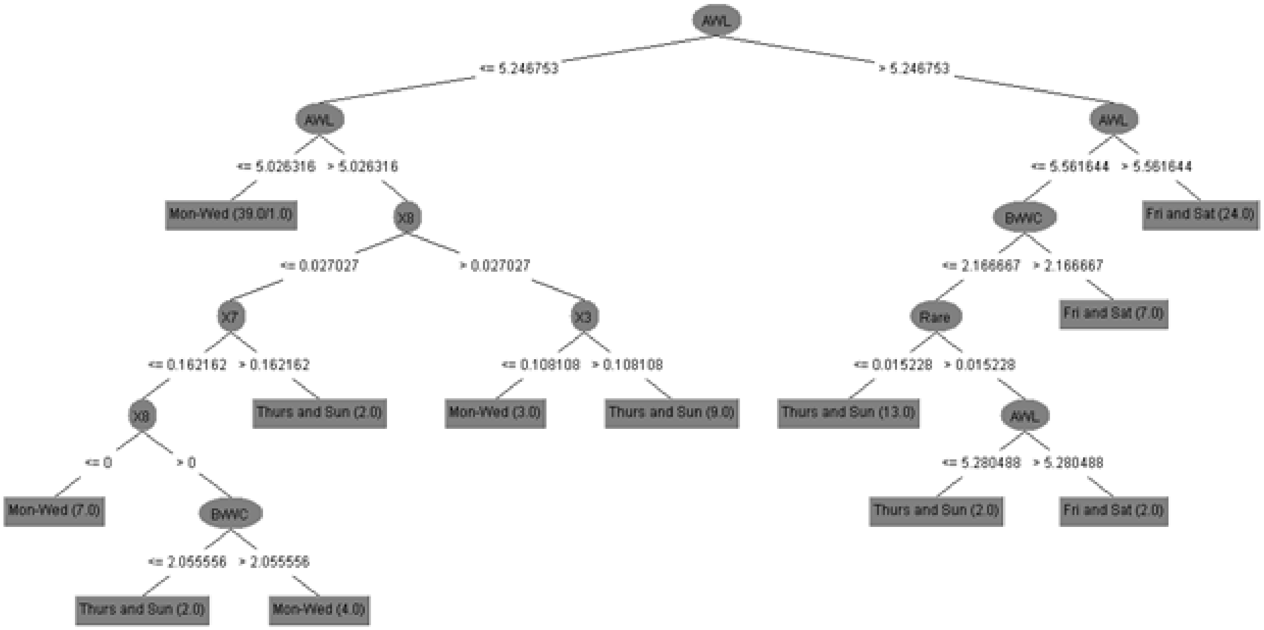

Figure 8: Best decision tree for crossword classification.

5 Conclusion

From the summary statistics we see that the average number of blocks in a puzzle decreases as the week progress, which allows for more spaces for the puzzler to fill in if we exclude Sunday. From the box plots of average word length we can see an increase in the average word length as the week progress if we exclude Sunday. These suggest that the puzzle does get harder as the week progresses.

We believe that our classification schemes confirm our suspicion that the difficulty of a crossword puzzle is not solely bound to the complexity of the crossword clue, and the classifications appear to back up the claim made by the New York Times that the complexity of the crossword puzzle increases as the week progresses. From the variables we analyzed, average word length is the most important indicator of puzzle complexity, but other variables such as the percent of rare letters and the number of words per block also add to the difficulty of the different puzzles.

References

Horne, J. (n.d.). Fun facts about the New York Times Crossword Puzzle. XWord Info – All about the New York Times Crossword Puzzle. Retrieved April 14, 2011, from www.xwordinfo.com/default.aspx

McCann, K. (n.d.). New York Times Specification Sheet. CRUCIVERB.COM - Crossword Constructors Community Center. Retrieved April 9, 2011, from http://www.cruciverb.com/index.php/index.php?action =ezportal;sa=page;p=16

SHEPARD, R. F. (1992, February 16). Bambi Is a Stag and Tubas Don't Go ?Pah-Pah?: The Ins and Outs of Across and Down. The New York Times - Breaking News, World News & Multimedia. Retrieved April 9, 2011, from http://www.nytimes.com/1992/02/16/magazine/crosswords-fiftyyears.html?ei=5070&en=eb226 eeaee1ab061&ex=1237089600&pagewanted=all

SHORTZ, W. (2001, April 8). How to Solve the New York Times Crossword Puzzle - NYTimes.com. The New York Times - Breaking News, World News & Multimedia. Retrieved April 9, 2011, from http://www.nytimes.com/2001/04/08/magazine/08PUZZLE.html?ex=1236916800&en=c04d7b2df8ca7806&ei=5070

Appendix: R Code

```
#Chileans
#Code
#Final Project

#Create new variables and clean data set
d <- read.csv("P:/503/Final Project/2010 Data.csv",header=T)
dp <- read.csv("P:/503/Final Project/2010 Data Percentages.csv",header=T)

for (i in 2:ncol(dp)){
dp[,i] <- as.numeric(as.character(dp[,i]))
}

#Make WordCount / Block number variable
d$BvWC <- d$WordCount/d$Blocks
dp$BvWC <- dp$WordCount/dp$Blocks

#Average Word Length
 d$AWL <- NA
 for (j in 1:nrow(d)) {
counter <- 0
  for (i in 31:(29+21)) {
counter = counter + (d[j,i]*(i-30))
 }
d$AWL[j] <- counter
 }
 d$AWL <- d$AWL/d$WordCount
 dp$AWL <- d$AWL


#Number of Vowels (with y)
d$Vowels_y <- d$A + d$E + d$I + d$O + d$U + d$Y
dp$Vowels_y <- dp$A + dp$E + dp$I + dp$O + dp$U + dp$Y

#Number of Vowels (without y)
d$Vowels <- d$A + d$E + d$I + d$O + d$U
dp$Vowels <- dp$A + dp$E + dp$I + dp$O + dp$U

#Number of Consanants (with y)
d$Consanants_y <- d$LetterCount - d$Vowels
dp$Consanants_y <- (d$LetterCount - d$Vowels)/d$LetterCount

#Number of Consanants (without y)
d$Consanants <- d$LetterCount - d$Vowels_y
dp$Consanants <- (d$LetterCount - d$Vowels_y)/d$LetterCount

#Number of "Rare" Letters (J,X,Q,Z) and "Common" ones
d$Rare <- d$J + d$X + d$Q + d$Z
dp$Rare <- dp$J + dp$X + dp$Q + dp$Z

d$Common <- d$LetterCount - d$Rare
```

```
dp$Common <- 1-dp$Rare

setwd("P:/503/Final Project")
write.csv(d,"2011_Data_Counts.csv")
write.csv(dp,"2011_Data_Percents.csv")


d <- read.csv("P:/503/Final Project/2011 Data.csv",header=T)
dp <- read.csv("P:/503/Final Project/2011 Data Percentages.csv",header=T)

for (i in 2:ncol(dp)){
dp[,i] <- as.numeric(as.character(dp[,i]))
}

#Make WordCount / Block number variable
d$BvWC <- d$WordCount/d$Blocks
dp$BvWC <- dp$WordCount/dp$Blocks

#Average Word Length
 d$AWL <- NA
 for (j in 1:nrow(d)) {
counter <- 0
  for (i in 31:(29+21)) {
counter = counter + (d[j,i]*(i-30))
 }
d$AWL[j] <- counter
 }
 d$AWL <- d$AWL/d$WordCount
 dp$AWL <- d$AWL


#Number of Vowels (with y)
d$Vowels_y <- d$A + d$E + d$I + d$O + d$U + d$Y
dp$Vowels_y <- dp$A + dp$E + dp$I + dp$O + dp$U + dp$Y

#Number of Vowels (without y)
d$Vowels <- d$A + d$E + d$I + d$O + d$U
dp$Vowels <- dp$A + dp$E + dp$I + dp$O + dp$U

#Number of Consanants (with y)
d$Consanants_y <- d$LetterCount - d$Vowels
dp$Consanants_y <- (d$LetterCount - d$Vowels)/d$LetterCount

#Number of Consanants (without y)
d$Consanants <- d$LetterCount - d$Vowels_y
dp$Consanants <- (d$LetterCount - d$Vowels_y)/d$LetterCount

#Number of "Rare" Letters (J,X,Q,Z) and "Common" ones
d$Rare <- d$J + d$X + d$Q + d$Z
dp$Rare <- dp$J + dp$X + dp$Q + dp$Z
```

```
d$Common <- d$LetterCount - d$Rare
dp$Common <- 1-dp$Rare

setwd("P:/503/Final Project")
write.csv(d,"2011_Data_Counts.csv")
write.csv(dp,"2011_Data_Percents.csv")



######### READ IN CLEANED DATA #####
library(ggplot2)
d <- read.csv("P:/503/Final Project/2010_Data_CountsT.csv",header=T)
dp <- read.csv("P:/503/Final Project/2010_Data_PercentsT.csv",header=T)

names(d)[29]<-"X"
names(dp)[29]<-"X"

d <- d[,-1]
dp <- dp[,-1]

d$Day <- factor(d$Day, levels=levels(d$Day)[c(3,6,7,2,1,5,4)])
dp$Day <- factor(dp$Day, levels=levels(dp$Day)[c(3,6,7,2,1,5,4)])

levels(d$Day) <- c("Mon","Tues","Wed","Thurs","Fri","Sat","Sun")
levels(dp$Day) <- c("Mon","Tues","Wed","Thurs","Fri","Sat","Sun")


##### Histograms ######
puzzle=dp

setwd("P:/503/Final Project/EDA")
pdf("Hist_Blocks.pdf",height=5,width=7)
qplot(Blocks, data=puzzle, fill=Day,geom="histogram",
ylab="Proportion of Puzzles \n Conditioned on Number of Blocks",
main="Distribution of Number of Blocks",xlab="Blocks")+
geom_bar(position = "fill")
dev.off()
pdf("Hist_AWL.pdf",height=5,width=7)
qplot(AWL, data=puzzle, fill=Day,geom="histogram",
ylab="Proportion of Puzzles \n Conditioned on Average Word Length",
main="Distribution of Average Word Lengths",xlab="Average Word Length")+
geom_bar(position = "fill")
dev.off()
pdf("Hist_BvWC.pdf",height=5,width=7)
qplot(BvWC, data=puzzle, fill=Day,geom="histogram",
ylab="Proportion of Puzzles \n Conditioned on Words per Block",
main="Distribution of Words per Block",xlab="Words per Block")+
geom_bar(position = "fill")
dev.off()
```

```
##### Boxplots ######
setwd("P:/503/Final Project/EDA")
#qplot(Day, Rare, geom="boxplot", data=dp, ylab="Percent",
main="Rare Letters (J,Q,X,Z)")
#qplot(Day, Common, geom="boxplot", ylab="Percent", data=dp,
main="Common Letters (not J,Q,X,Z)")
#qplot(Day, Vowels, geom="boxplot", ylab="Percent", data=dp,
main="Percent Vowels Less ’Y’")
pdf("Vowels_y.pdf",height=5,width=5)
qplot(Day, Vowels_y*100, geom="boxplot",
ylab="Percentage of Letters that are Vowels, Including Y",
data=dp, main="Percent Vowels Including ’Y’")
dev.off()
pdf("AWL.pdf",height=5,width=5)
qplot(Day, AWL, geom="boxplot",
ylab="Average Word Length (number of letters)",
data=dp, main="Average Word Length")
dev.off()
pdf("WpB.pdf",height=5,width=5)
qplot(Day, BvWC, geom="boxplot", ylab="Words per Block",
data=d, main="Words per Block")
dev.off()
#qplot(Day, A, geom="boxplot", data=dp)
pdf("B.pdf",height=5,width=5)
qplot(Day, 100*B, ylab="Percent of B’s in Crossword Puzzles",
main="B", geom="boxplot", data=dp) #I can’t find the puzzle that has 25% B’s.
dev.off()
#qplot(Day, C, geom="boxplot", data=dp)
#qplot(Day, D, geom="boxplot", data=dp)
#qplot(Day, E, geom="boxplot", data=dp)
pdf("F.pdf",height=5,width=5)
qplot(Day, 100*F, ylab="Percent of F’s in Crossword Puzzles",
main="F",geom="boxplot", data=dp) #Wednesday is lower than the others
dev.off()
#qplot(Day, G, geom="boxplot", data=dp)
#qplot(Day, H, geom="boxplot", data=dp)
#qplot(Day, I, geom="boxplot", data=dp)
pdf("J.pdf",height=5,width=5)
qplot(Day, 100*J, geom="boxplot", ylab="Percent of J’s in Crossword Puzzles",
main="J",data=dp) #Sunday uses J even less often than the other days
dev.off()
#qplot(Day, K, geom="boxplot", data=dp)
#qplot(Day, L, geom="boxplot", data=dp)
#qplot(Day, M, geom="boxplot", data=dp)
pdf("N.pdf",height=5,width=5)
qplot(Day, 100*N, geom="boxplot", ylab="Percent of N’s in Crossword Puzzles",
main="N",data=dp) #Increasing over time
dev.off()
#qplot(Day, O, geom="boxplot", data=dp)
#qplot(Day, P, geom="boxplot", data=dp)
#qplot(Day, Q, geom="boxplot", data=dp)
```

```
#qplot(Day, R, geom="boxplot", data=dp)
#qplot(Day, S, geom="boxplot", data=dp)
#qplot(Day, T, geom="boxplot", data=dp)
#qplot(Day, U, geom="boxplot", data=dp)
#qplot(Day, V, geom="boxplot", data=dp)
#qplot(Day, W, geom="boxplot", data=dp)
#qplot(Day, X.1, geom="boxplot", data=dp)
#qplot(Day, Y, geom="boxplot", data=dp)
#qplot(Day, Z, geom="boxplot", data=dp)


#Line plot of Letters
letters <- data.frame(Day=d$Day, dp[,6:30],indx=factor(1:nrow(dp)))
l.melt <- melt(letters)
l.avg <- ddply(l.melt, .(Day,variable),summarise,
med = median(value),
avg = mean(value),
max=max(value),
min=min(value))
l.all.avg <- ddply(l.melt, .(variable),summarise,
med = median(value),
avg = mean(value),
max=max(value),
min=min(value))
l.avg$variable <- factor(l.avg$variable,
levels=levels(l.avg$variable)[c(order(l.all.avg$avg,decreasing=T))])

setwd("P:/503/Final Project/EDA")
pdf("Letters.pdf",height=5,width=9)
qplot(variable, med*100, color=Day, group=Day, geom="line",
size=I(1),data=l.avg,
xlab="Letters", ylab="Percentage of Crossword Puzzle Letters",
main="Median Percent Letter Usage by Day")
dev.off()
#qplot(variable, avg, color=Day, group=Day, geom="line", size=I(1),data=l.avg)
#qplot(variable, max, color=Day, group=Day, geom="line", size=I(1),data=l.avg)
#qplot(variable, med, color=Day, group=Day, geom="line", size=I(1),data=l.avg)


#### Line Plot of Word Lenghts #######
#Words Between 3 and 8 letters long
word3_8 <- data.frame(Day=d$Day,dp[,34:42])
w.melt <- melt(word3_8)
w.avg <- ddply(w.melt, .(Day,variable), summarise, count = mean(value), med=median(value))
w.melt <- w.avg

w.melt$VarNum<-3
w.melt$VarNum[w.melt$variable=="X4"]<-4
w.melt$VarNum[w.melt$variable=="X5"]<-5
w.melt$VarNum[w.melt$variable=="X6"]<-6
w.melt$VarNum[w.melt$variable=="X7"]<-7
```

```
w.melt$VarNum[w.melt$variable=="X8"]<-8
w.melt$VarNum[w.melt$variable=="X9"]<-9
w.melt$VarNum[w.melt$variable=="X10"]<-10
w.melt$VarNum[w.melt$variable=="X11"]<-11
setwd("P:/503/Final Project/EDA")
#pdf("Word_Length.pdf",height=5,width=5)
qplot(factor(VarNum),count*100,group=Day,geom="line",color=Day,
size=I(2.5),data=w.melt,xlab="Word Length",
ylab="Percent of Words in Puzzle",
main="Average Percent Word Lengths by Day")
#dev.off()



#############################################################
#############################################################
#############################################################

#### Classification Techniques #########
test.std<-read.csv("P:/503/Final Project/2011_Data_Percents.csv",header=T)
train.std<-read.csv("P:/503/Final Project/2010_Data_PercentsT.csv",header=T)

test <- test.std[,-c(1,3:18,20:23,25:33,41:52,55,57,58,60)]
train <- train.std[,-c(1,3:18,20:23,25:33,41:58,61,63,64,66)]

test$Category <- "Mon-Wed"
test$Category[(test$Day == "h") | (test$Day == "n")] <- "Thurs and Sun"
test$Category[(test$Day == "f") | (test$Day == "s")] <- "Fri and Sat"
test$Category <- factor(test$Category)
test$Category <- factor(test$Category, levels=levels(test$Category)[c(2,3,1)])

testfull <- test
test <- testfull[,-1]

train$Category <- "Mon-Wed"
train$Category[(train$Day == "h") | (train$Day == "n")] <- "Thurs and Sun"
train$Category[(train$Day == "f") | (train$Day == "s")] <- "Fri and Sat"
train$Category <- factor(train$Category)
train$Category <- factor(train$Category, levels=levels(train$Category)[c(2,3,1)])

trainfull <- train
train <- trainfull[,-1]

#write.csv(test,"ClassTest.csv")
#write.csv(train,"ClassTrain.csv")

# Logistic regression
library(nnet)
puzzle.log.reg<-multinom(Category~.,data=train)
table(train$Category,predict(puzzle.log.reg,train,type="class"))
table(test$Category,predict(puzzle.log.reg,test,type="class"))
```

```
puzzle.log.reg

step(puzzle.log.reg,direction="backward")
puzzle.log.reg.b<-multinom(Category~X4 + X5 + X6 + BvWC + AWL + Rare,data=train)
table(train$Category,predict(puzzle.log.reg.b,train,type="class"))
table(test$Category,predict(puzzle.log.reg.b,test,type="class"))
multi <- data.frame(Actual=test$Category,
Predicted = predict(puzzle.log.reg.b,test,type="class"))
puzzle.log.reg.b

setwd("P:/503/Final Project/EDA")
#pdf("Multi.pdf",height=5,width=5)
ggfluctuation(xtabs(~Actual + Predicted,data=multi))
#dev.off()

#Accuracy Rates Test
sum(diag(table(test$Category,predict(puzzle.log.reg.b,test,type="class"))))/
sum(table(test$Category,predict(puzzle.log.reg.b,test,type="class")))
#Accuracy Rates Train
sum(diag(table(train$Category,predict(puzzle.log.reg.b,train,type="class"))))/
sum(table(train$Category,predict(puzzle.log.reg.b,train,type="class")))



###### LDA #########
library(MASS)
lda.train <- data.frame(train$Category,train$X4,train$X5,
train$X6,train$BvWC,train$AWL,train$Rare)
lda.test <- data.frame(test$Category,test$X4,test$X5,
test$X6,test$BvWC,test$AWL,test$Rare)
puzzle.lda<-lda(lda.train[,-1],lda.train[,1])
table(lda.train[,1],
predict(puzzle.lda,lda.train[,-1],dimen=1)$class)
table(lda.test[,1],
predict(puzzle.lda,lda.test[,-1],dimen=1)$class)
puzzle.lda
lda <- data.frame(Actual=lda.test$test.Category,
Predicted=predict(puzzle.lda,lda.test[,-1],dimen=1)$class)

setwd("P:/503/Final Project/EDA")
pdf("LDA.pdf",height=5,width=5)
ggfluctuation(xtabs(~Actual + Predicted,data=lda))
dev.off()


#Accuracy Rates Test
sum(diag(table(lda.test[,1],predict(puzzle.lda,
lda.test[,-1],dimen=1)$class)))/sum(table(lda.test[,1],
predict(puzzle.lda,lda.test[,-1],dimen=1)$class))
#Accuracy Rates Train
sum(diag(table(lda.train[,1],predict(puzzle.lda,
lda.train[,-1],dimen=1)$class)))/sum(table(lda.train[,1],
predict(puzzle.lda,lda.train[,-1],dimen=1)$class))


# Variable importance
for (i in c(2:9))
cat(cor(train[,i],
predict(puzzle.lda,train[,2:9],dimen=1)$x),"\n")
var(train)
# Calculate constant
apply(puzzle.lda$means,2,mean)%*%puzzle.lda$scaling


####### QDA #########
qda.train <- data.frame(train$Category,train$X4,train$X5,
train$X6,train$BvWC,train$AWL,train$Rare)
qda.test <- data.frame(test$Category,test$X4,test$X5,test$X6,
test$BvWC,test$AWL,test$Rare)
puzzle.qda<-qda(qda.train[,-1],qda.train[,1])
table(qda.train[,1], predict(puzzle.qda,qda.train[,-1])$class)
table(qda.test[,1],predict(puzzle.qda,qda.test[,-1],dimen=1)$class)

qda <- data.frame(Actual=qda.test$test.Category,
Predicted=predict(puzzle.qda,qda.test[,-1],dimen=1)$class)

setwd("P:/503/Final Project/EDA")
#pdf("QDA.pdf",height=5,width=5)
ggfluctuation(xtabs(˜Actual + Predicted,data=qda))
#dev.off()

#Accuracy Rate Test
sum(diag(table(qda.test[,1],predict(puzzle.qda,qda.test[,-1],dimen=1)$class)))/
sum(table(qda.test[,1],predict(puzzle.qda,qda.test[,-1],dimen=1)$class))
#Accuracy Rate Train
sum(diag(table(qda.train[,1],predict(puzzle.qda,qda.train[,-1],dimen=1)$class)))/
sum(table(qda.train[,1],predict(puzzle.qda,qda.train[,-1],dimen=1)$class))


#### SVMs and Decision Trees were created in Weka
```