An Intro to Bracketology:
Predicting the Field of Teams Qualifying for the NCAA Tournament

Jeremy Thames & Josh Davis

HOD 2790: Data Science

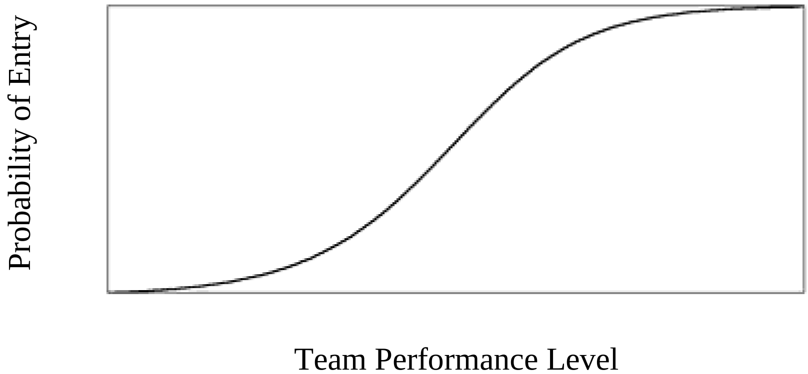
Introduction

Fifteen years ago, the term "bracketology" would have never been used on ESPN. It wasn't a term you could search for online and immediately find thousands of analysts sharing their thoughts on which teams would earn a spot in the annual NCAA men's basketball tournament. But things were changing. In 2002, bracketology was a science in its infancy as a new era of quantification was emerging in the world of sports. While Billy Beane's Oakland Athletics were defying tradition by using statistics to win baseball games, Joe Lunardi, the bracketologist himself, was making his first television appearances on ESPN. Lunardi's work primarily consisted of determining which teams would enter the tournament and what their seeds would be. Today, you can't make it past the front page of ESPN.com in February without being bombarded with predictions about the NCAA men's basketball tournament. In the month of March, Joe Lunardi becomes one of the busiest men in the world of sports, and the phrase he coined, bracketology, becomes one of the most popular searches on the internet.

Today's NCAA men's basketball tournament consists of 68 teams. Each of the 32 division 1 conferences get one automatic bid to the tournament for their conference champion and the other 36 spots are determined by a selection committee. Every year, the choices of this selection committee are analyzed by countless statisticians, sports journalists, and college basketball fans around the world. While we cannot know for sure exactly which qualities this committee looks for in selecting teams, we do know that they tend to focus on teams' RPI ranking and strength of schedule. In addition to these metrics, the committee has, in the past, stated that it uses an "eye test" as a criterion to select teams. This process all occurs in an effort to ensure that the best and most deserving teams in the country are included in the tournament. Despite the work of this committee, every year there is debate about whether or not they actually choose the correct teams. In some cases, some qualified teams are left out while seemingly inferior teams are included. The fact that these cases exist points to the qualitative nature of the process. While it will probably never be possible to perfectly model the selections of this committee, we can use what we know to develop a strong idea of what decisions they will make. By using data available to everyone, we can get a better idea of which teams will earn their way into the tournament. This is what bracketology is all about.

Hypothesis

Our initial thoughts going into this statistical analysis are that there are certain quantitative factors which influence the decisions of the selection committee every year. Among these are RPI ranking, strength of schedule, offensive efficiency, rebounding efficiency, turnover margin, and shooting percentage. We feel that every team in division 1 basketball has some non-zero probability of entering the field and that we can calculate that probability using some combination of these metrics. We do not, however, expect the relationship between these variables and the probability of tournament entry to be linear. We expect the top performing teams in a given year to have probabilities of entrance very close to 100% while the lowest performing teams should have probabilities values of close to zero. To represent this visually, we provide the following figure:



We expect that as teams become higher performing, they will at first be exponentially more likely to be selected and then will eventually experience diminishing returns from improving in performance level. We will use logistic regression models and compare them to linear regression models to test our suspicion that this is the best way to predict tournament entry. We will also try different combinations of variables in our models to attempt to find the highest performing group of predictors without overfitting to our data set.

Building the Model

Before we started blindly testing models, we wanted to really think about what we were attempting to predict. Yes, we knew we wanted to predict likelihood of tournament entry, but really, we're trying to predict likelihood of *at-large* tournament entry. The reality is, there's not much use trying to predict which teams are going to get hot at the right time and gain a spot in the NCAA tournament by winning their conference tournament. Because of this, the teams that make the tournament simply because they win their conference tournament present a problem for us. Unfortunately, the solution is not as simple as just ignoring those teams because sometimes, teams that would have made the NCAA tournament anyway win their conference tournament. We want to be predicting those teams in our model, so we cannot just blindly omit all conference tournament winners. In response to this problem, we decided to create a "deserved entry" variable that consists of all the at-large bids, and all the conference tournament winners with RPI rankings of lower than 40. We used the following logic to decide on using 40 as a break point:

1. There are 32 conferences.
2. Conference champions automatically make the tournament.
3. 68 teams make the tournament.
4. If RPI were a perfect predictor of tournament selection, the top 36 teams would make it regardless of conference tournament results.
5. Depending on the year, there are usually several conference champions from that top 36 group, so more spots are opened up for at large bids.
6. If we assume that 4 conference tournaments are won by top 36 teams, then we'll have 40 deserved entrants per year.

We recognize we're making a few major assumptions here. RPI is not a perfect predictor of tournament entry and 4 top 36 teams do not always win their conference tournaments. Because the logic we used to reach the 40 number is somewhat arbitrary, we did a sensitivity analysis of our model's coefficients to that break point.

Sensitivity Table

	30	35	40	45	50
RPI	(0.0010)	(0.0011)	(0.0011)	(0.0012)	(0.0013)
SOS	0.0210	0.0201	0.0199	0.0187	0.0180
PP100 Poss.	0.0054	0.0057	0.0053	0.0050	0.0051

Variance

	30	35	40	45	50
RPI	-10.9%	-5.8%	0.0%	8.5%	14.6%
SOS	5.3%	1.1%	0.0%	-6.1%	-9.8%
PP100 Poss.	1.2%	6.8%	0.0%	-6.0%	-4.3%

It's clear that the effect on our model is minimal, so we are comfortable using 40 as a break point.

Now that we knew what we were predicting, we gathered data on all the division 1 basketball teams for the last four years. This data is readily available on the internet. After testing linear and logistic models of prediction, it quickly became obvious that logistic regression offered the best way to look at probability of deserved entrance.

In our logistic models, we tried numerous different combinations of the following variables:

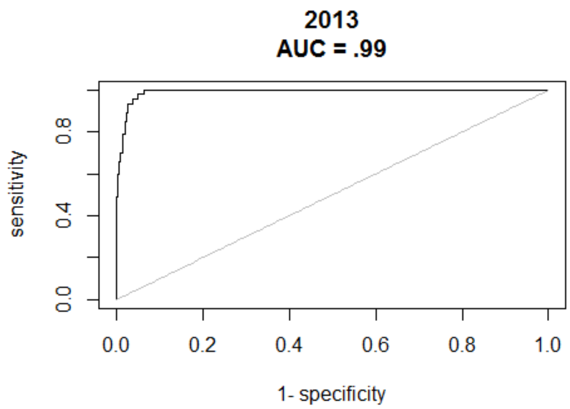
- RPI Ranking
- Strength of Schedule
- Offensive Efficiency
- Turnover Margin
- Assist Efficiency
- Rebounding Efficiency
- True Shooting Percentage (adjusts for 3 pointers and free throws)
- Pace

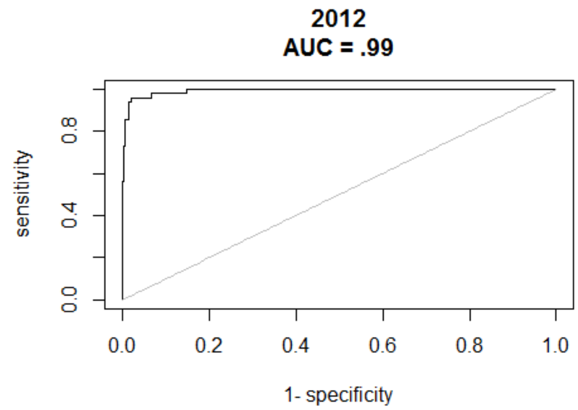
Although every single one of these variables was able to predict tournament entry in a statistically significant fashion, RPI ranking, strength of schedule, and offensive efficiency were the only ones that remained independently useful in combined models. These variables included together in a logistic regression model provide the best prediction of deserved entry.

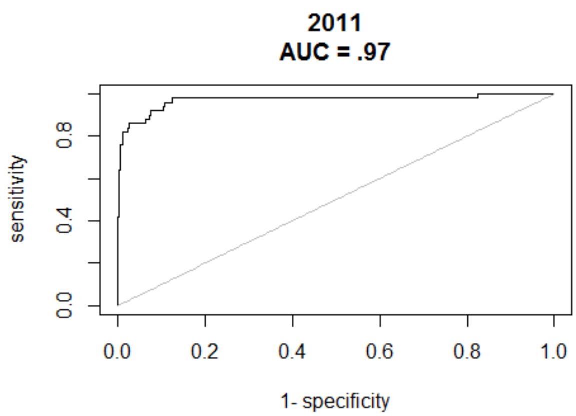
We developed our model by using a training data set of the most recent three years of college basketball. In the interest of accuracy, we used the bootstrap methodology to randomly sample from those three seasons and used the median of the coefficients that model provided. As we suspected, teams with lower RPIs, higher strengths of schedule, and higher offensive efficiency were more likely to be selected for the tournament. Applying this model back to the training data sets, we found that our model yielded AUCs of .97 or more for all three years of training data. Our next step was to test our model against a data set with which it did not have experience. Using data from the 2010-2011 college basketball season, we were able to predict at an AUC level of .975. Obviously, these variables and the coefficients we developed give us a very good idea of deserved entry into the NCAA tournament.

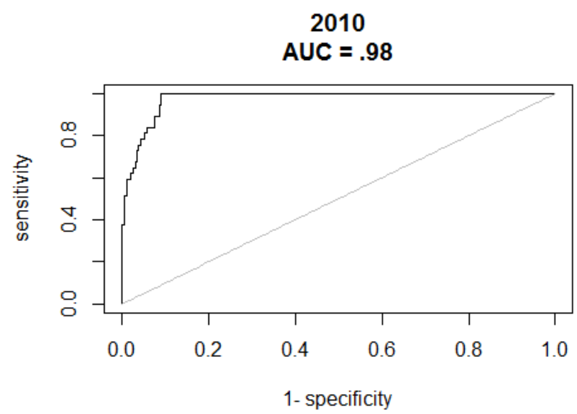
Results

Developing a model with predictive power on tournament entry was only the first step in our analysis. After we identified the most reliable variables and used the bootstrap methodology to find the most effective coefficients, we began to use this model to make some predictions. First though, here are the graphical representations of the sensitivity/specificity tradeoff in our model years.



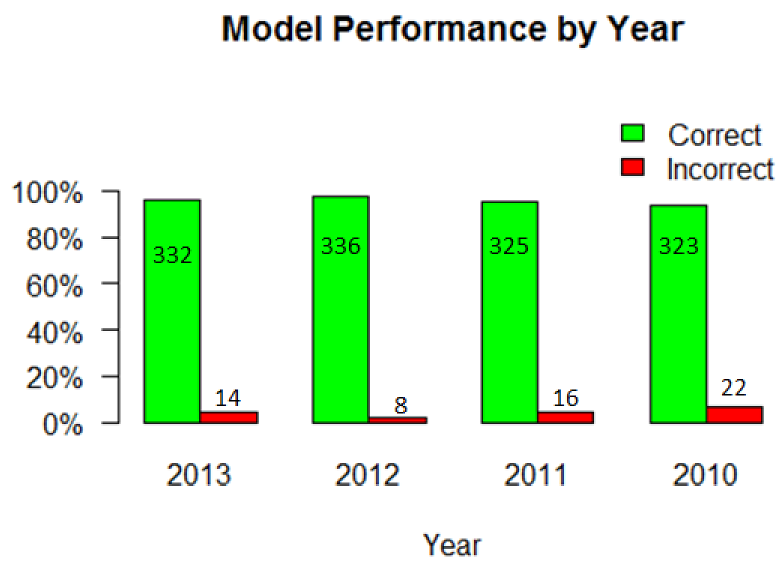






While it’s very exciting to see AUCs this high, we need to be careful not to interpret AUC as a perfect proxy for accuracy. In the NCAA tournament, only a certain amount of teams can enter the tournament. Our model assigns every team a probability but it does not realize that a fixed amount will enter. So, depending on the data, the model could severely under or overestimate the amount of teams that would enter the field if we use a fixed break point in probabilities to predict entry every year. Because of this, we are hesitant to look at AUC alone as a measure of our accuracy.

Luckily, there is a better way to review the accuracy our model. The deserved entry variable we created before tells us which teams entered the tournament without the help of winning their conference tournament. If we sum the number of deserved entrants in each year’s data set, we will, of course, have a total number of deserved entrants. We can then go over to our predictions and take the same amount of the highest probability of entry teams. With this, we will have a list of the actual deserved entrants and our model's output of the most qualified teams given the constraint of how many can be chosen. If we give every team on our predicted list an indicator value, we can compare the indicator variable to the actual deserved entry variable and have a proportion of teams correctly predicted. This gives the following breakdown:



Using Our Data

Our results suggest our model is pretty effective at predicting the decisions made by the selection committee. But everyone knows the top few teams are going to make the tournament. And everyone knows the worst few teams will not make the tournament. The value comes in predicting the teams that are on the bubble. Although our model is nowhere near as good as Joe Lunardi, who often predicts the field within a team or two, it still outperforms the RPI rankings alone. We have put together a list of bubble teams for each year in the same way that is popularly presented in sports media outlets:

	2013			2012	
Last 4 In	**First 4 Out**	**Next 4 Out**	**Last 4 In**	**First 4 Out**	**Next 4 Out**
Nebraska	UMass	VCU	Iowa	Oregon	Tennessee
Cincinnati	Wichita State	Maryland	St. Louis	Stanford	BYU
Dayton	Illinois	Utah	La Salle	Ole Miss	Southern Miss
West Virginia	St. Josephs	SMU	Cincinnati	Memphis	Maryland

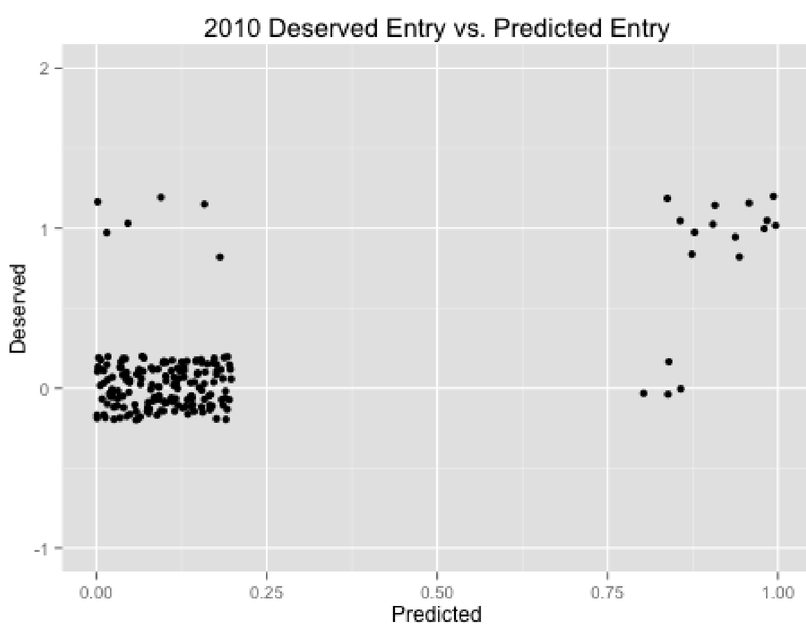
	2011			2010	
Last 4 In	**First 4 Out**	**Next 4 Out**	**Last 4 In**	**First 4 Out**	**Next 4 Out**
Pittsburgh	St. Marys	Mississippi St.	Minnesota	UNLV	Xavier
Marshall	Illinois	Oklahoma St.	Butler	California	Clemson
San Diego St.	Tennessee	Villanova	Northwestern	Texas A&M	Miami
Colorado	Seton Hall	St. Josephs	Cincinnati	Colorado	USC

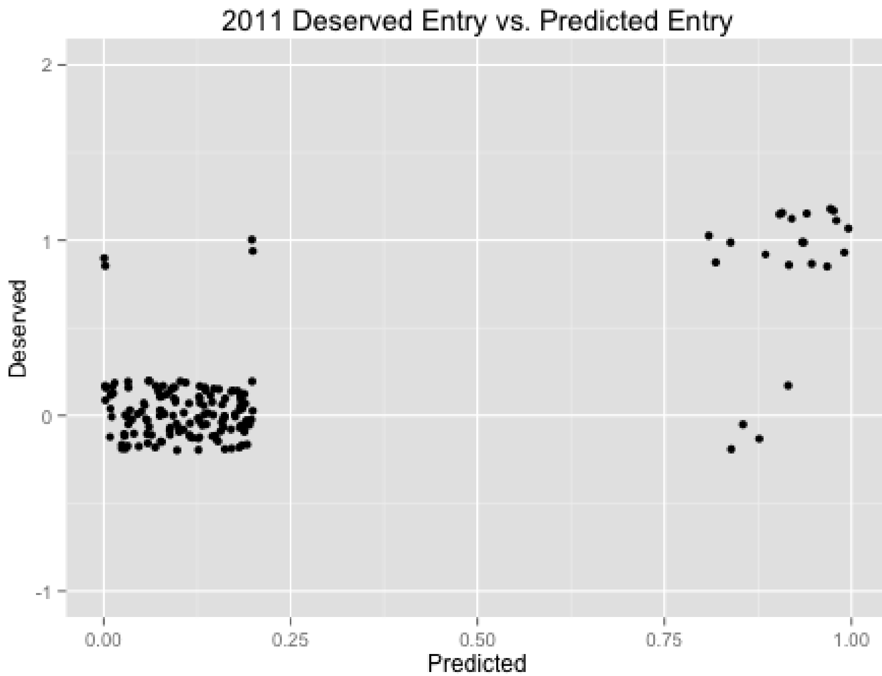
The green teams indicate correct predictions, the red indicates incorrect predictions. Clearly, picking the bubble teams is much more difficult. We would, however, like to point out that 10 of 16 of our “last four in” teams actually did manage to make the tournament and 9/16 of our “first four out” teams were left out. So, as a team, you would prefer to be on our “in” list as opposed to “first four out” list, but barely. What we’re really proud of is that 12 of 16 of our “next four out” teams were not selected for the tournament. This tells us that even though we’re not outstanding at predicting which bubble teams make it, we are outstanding at predicting which teams are on the bubble. While this may not sound like much of an accomplishment, we think this information could be incredibly valuable to coaches, players, and fans across the country.

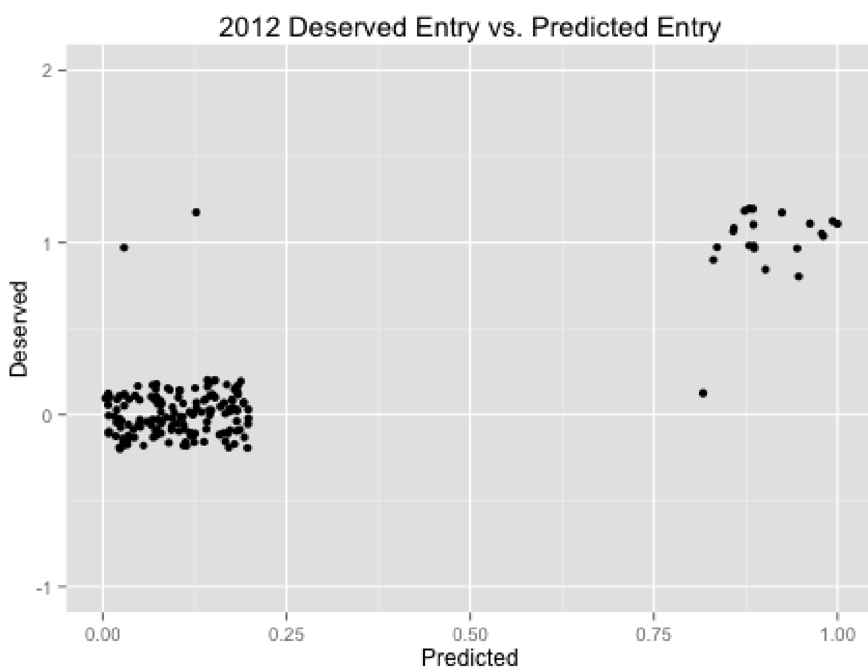
Conclusions

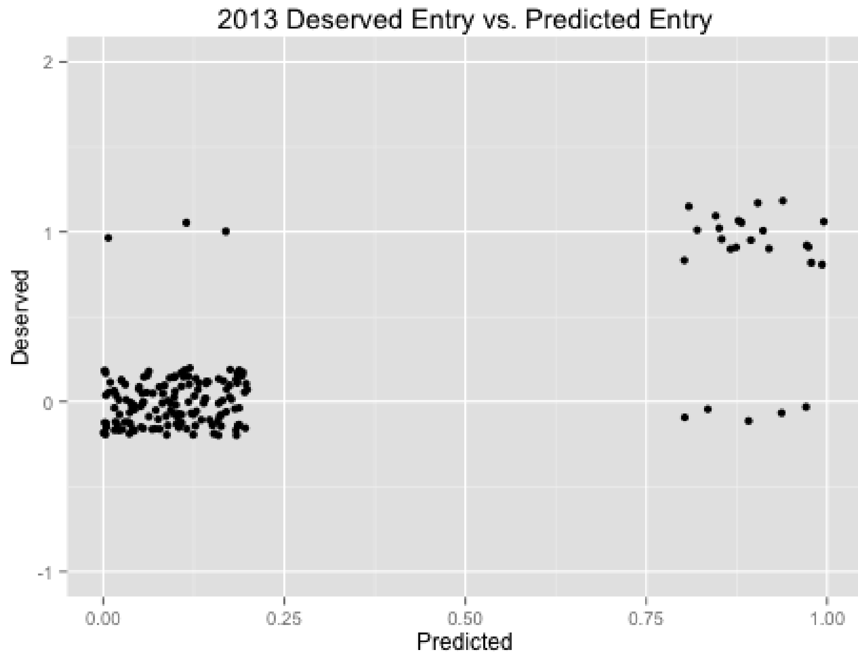
Our model, which is a result of a bootstrapped logistical regression using RPI ranking, strength of schedule, and offensive efficiency, has proven to be effective at predicting which teams will enter the NCAA tournament. It is particularly effective at predicting which teams are “on the bubble,” meaning their chances of making the tournament are close to 50/50. While we wish the model was more effective at determining which of those bubble teams would make the tournament, we believe we have made considerable progress in determining the criteria the selection committee uses to make decisions. Some of the variables we initially thought would have strong predictive power actually were not helpful at all in predicting the selection committee’s decisions. Although we may never be able to fully quantify the factors which determine entry into the NCAA tournament, we think our model proves that certain statistics are undeniably more relevant than others and can be used to effectively predict entry into the NCAA tournament. Our final conclusion is that bracketology seems to be a legitimate science and that using statistics to predict the bracket offers a valuable perspective on something countless individuals have and will continue to discuss for years.

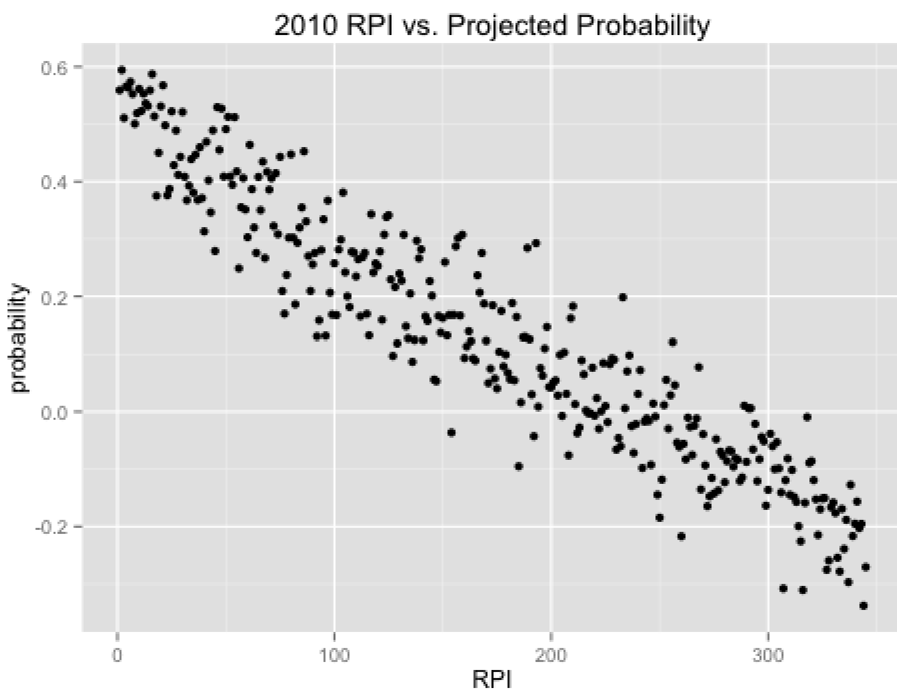
Appendix

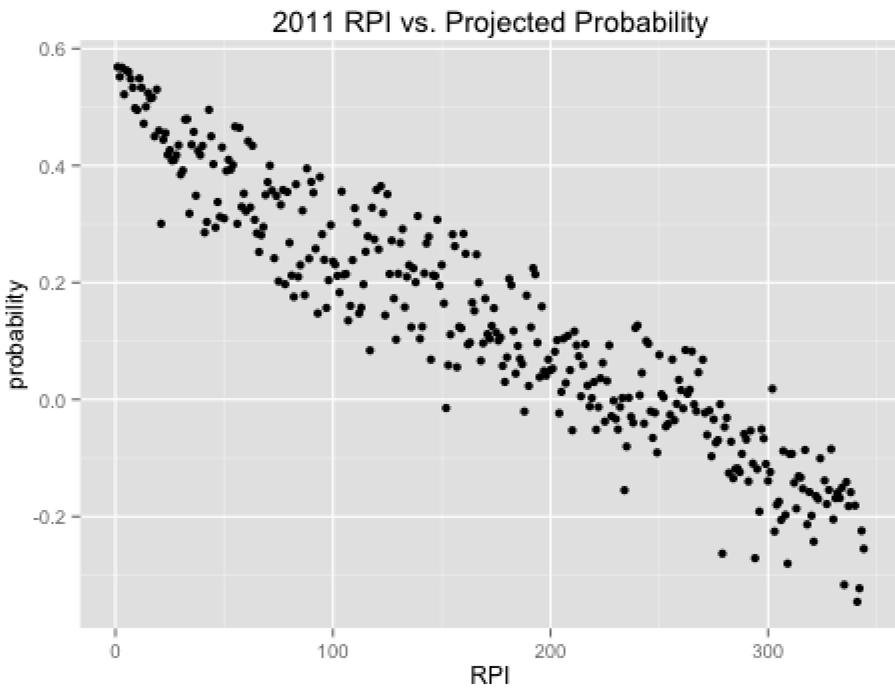


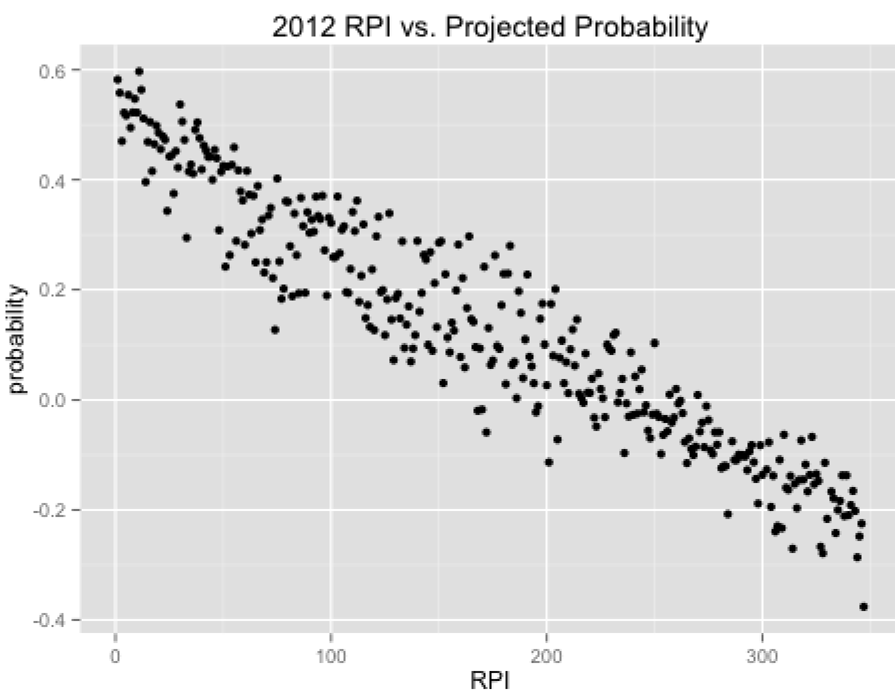

2011 Deserved Entry vs. Predicted Entry
Deserved
2
1
0
-1
Predicted
0.00
0.25
0.50
0.75
1.00

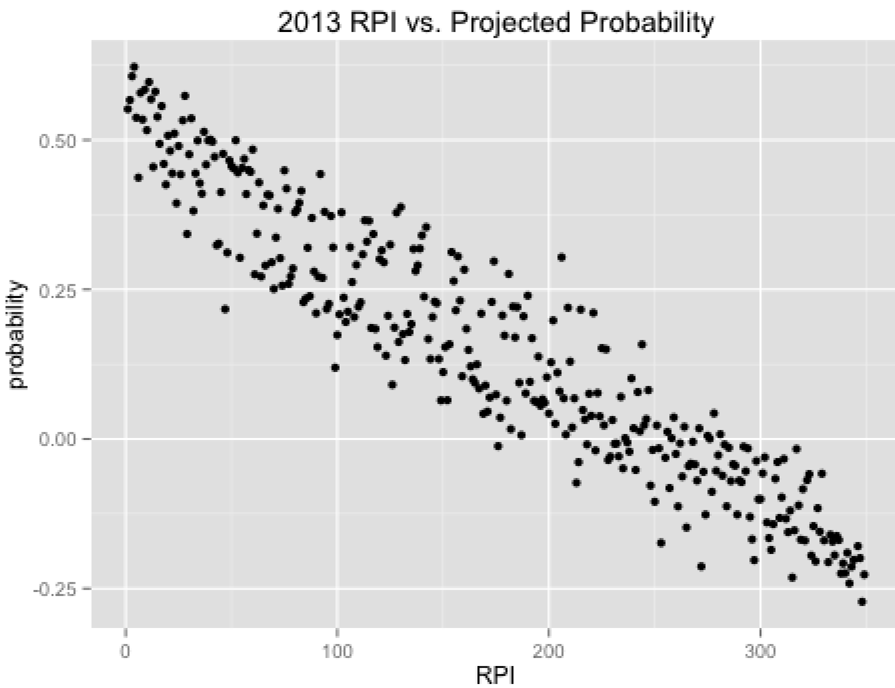

2012 Deserved Entry vs. Predicted Entry
Deserved
2
1
0
-1
Predicted
0.00
0.25
0.50
0.75
1.00


2013 Deserved Entry vs. Predicted Entry
2
1
0
-1
Deserved
0.00
0.25
0.50
0.75
1.00
Predicted


2010 RPI vs. Projected Probability
0.6
0.4
0.2
0.0
-0.2
probability
0
100
200
300
RPI


2011 RPI vs. Projected Probability
probability
0.6
0.4
0.2
0.0
-0.2
0
100
200
300
RPI


2012 RPI vs. Projected Probability
probability
0.6
0.4
0.2
0.0
-0.2
-0.4
0
100
200
300
RPI


2013 RPI vs. Projected Probability
probability
0.50
0.25
0.00
-0.25
0
100
200
300
RPI

```
################################################
# Final Project
# Jeremy Thames & Josh Davis#
# Predicting NCAA Tournament Entries #
# 9/24/14 #
################################################

library(ggplot2)
library(plyr)
library(scales)
library(AUC)
library(caret)
library(car)
library(boot)

#Pull The '13-'14,  '12-'13, & '11-'12 Data From Our CSV File

ncaa2013<-read.table("/Users/joshdavis1001/Documents/GitHub/Final-Project---
Working/2013.stats.csv",
           sep=",",
           header=TRUE,
           as.is=TRUE)

ncaa2013<-ncaa2013[,-2]
ncaa2013<-ncaa2013[,-2]
ncaa2013<-ncaa2013[,-2]

#Make a new variable that indicates the team made the tournament but not because they
won their
#conference tourney
ncaa2013$flag<-0
ncaa2013$flag[ncaa2013$Tournament.==1 & ncaa2013$RPI<40]<-1
ncaa2013$flag[ncaa2013$Tournament.==1 & ncaa2013$auto.bid==0]<-1


ncaa2012<-read.table("/Users/joshdavis1001/Documents/GitHub/Final-Project---
Working/2012.stats.csv",
                     sep=",",
                     header=TRUE,
                     as.is=TRUE)

ncaa2012<-ncaa2012[,-2]
ncaa2012<-ncaa2012[,-2]
ncaa2012<-ncaa2012[,-2]

ncaa2012$flag<-0
ncaa2012$flag[ncaa2012$Tournament.==1 & ncaa2012$RPI<40]<-1
ncaa2012$flag[ncaa2012$Tournament.==1 & ncaa2012$auto.bid==0]<-1


ncaa2011<-read.table("/Users/joshdavis1001/Documents/GitHub/Final-Project---
Working/2011.stats.csv",
                     sep=",",
                     header=TRUE,
                     as.is=TRUE)

ncaa2011<-ncaa2011[,-2]
ncaa2011<-ncaa2011[,-2]
ncaa2011<-ncaa2011[,-2]

ncaa2011$flag<-0
ncaa2011$flag[ncaa2011$Tournament.==1 & ncaa2011$RPI<40]<-1
ncaa2011$flag[ncaa2011$Tournament.==1 & ncaa2011$auto.bid==0]<-1

#Logistical Regression
log.mod.2013<-glm(flag~ RPI+
                        SOS+
                        PP100.Poss.,
                        data=ncaa2013,
                        y=TRUE)

summary(log.mod.2013)
```

```
log.pred.2013<-predict(log.mod.2013,type="response")
log.roc.2013<-roc(log.pred.2013,as.factor(log.mod.2013$y))
auc(log.roc.2013)
plot(log.roc.2013)

#Let's Test this model on 2012 and 2011

pred.new<-predict(log.mod.2013,newdata=ncaa2012, type="response")
log.roc.new<-roc(pred.new,as.factor(ncaa2012$flag))
auc(log.roc.new)

pred.new1<-predict(log.mod.2013,newdata=ncaa2011, type="response")
log.roc.new1<-roc(pred.new1,as.factor(ncaa2011$flag))
auc(log.roc.new1)

auc(log.roc.2013)
auc(log.roc.new)
auc(log.roc.new1)
#Little Bit of Overfitting

#Let's try Bootstrapping using a combined data set
#first combine the data sets

ncaa<-rbind(ncaa2013,ncaa2012,ncaa2011)

#Put a model in using ncaa data set
log.mod<-glm(flag~ RPI+
                   SOS+
                   PP100.Poss.,
                   data=ncaa,
                   y=TRUE)

summary(log.mod)

#Now let's run the bootstrap samples
thames.boot<-Boot(log.mod,R=10000)

summary(thames.boot,high.moments=F)

#Here are our new confidence intervals
confint(thames.boot,level=.9,type="norm")

#And our coefficients are
#intercept = -0.2051644
#RPI = -0.0011209
#SOS = 0.0199085
#PP100.Poss. = 0.0052917

#Let's load those coefficients into the model
log.mod$coefficients
log.mod$coefficients["(Intercept)"]<- -0.2051644
log.mod$coefficients["RPI"]<- -0.0011209
log.mod$coefficients["SOS"]<- 0.0199085
log.mod$coefficients["PP100.Poss."]<- 0.0052917

log.mod$coefficients

#Let's use these coefficients on the yearly datasets and check AUCs
predict.2013<-predict(log.mod,newdata=ncaa2013, type="response")
predict.roc.2013<-roc(predict.2013,as.factor(ncaa2013$flag))
auc(predict.roc.2013)

predict.2012<-predict(log.mod,newdata=ncaa2012, type="response")
predict.roc.2012<-roc(predict.2012,as.factor(ncaa2012$flag))
auc(predict.roc.2012)

predict.2011<-predict(log.mod,newdata=ncaa2011, type="response")
predict.roc.2011<-roc(predict.2011,as.factor(ncaa2011$flag))
auc(predict.roc.2011)

#This is awesome, we have AUCs of above .969 for all three years
#Now for the real test, let's read in 2010 data and check our model on that
```

```
ncaa2010<-read.table("/Users/joshdavis1001/Documents/GitHub/Final-Project---
Working/2010.stats.csv",
                     sep=",",
                     header=TRUE,
                     as.is=TRUE)

ncaa2010<-ncaa2010[,-2]
ncaa2010<-ncaa2010[,-2]
ncaa2010<-ncaa2010[,-2]

ncaa2010$flag<-0
ncaa2010$flag[ncaa2010$Tournament.==1 & ncaa2010$RPI<40]<-1
ncaa2010$flag[ncaa2010$Tournament.==1 & ncaa2010$auto.bid==0]<-1


#Here we go
predict.2010<-predict(log.mod,newdata=ncaa2010, type="response")
predict.roc.2010<-roc(predict.2010,as.factor(ncaa2010$flag))
auc(predict.roc.2010)

#AUC = .975

plot(predict.roc.2013,main="2013\nAUC = .99")
plot(predict.roc.2012,main="2012\nAUC = .99")
plot(predict.roc.2011,main="2011\nAUC = .97")
plot(predict.roc.2010,main="2010\nAUC = .98")


#Put the probability of tournement entry into the data sets as a variable
ncaa2013$probability<-predict.2013
ncaa2012$probability<-predict.2012
ncaa2011$probability<-predict.2011
ncaa2010$probability<-predict.2010

#Plot RPI vs. Projected Probability for each year
#2013
g2013<-ggplot(data=ncaa2013,
           aes(x=RPI,y=probability))

g2013<-g2013+geom_point() + ggtitle("2013 RPI vs. Projected Probability")
g2013

#2012
g2012<-ggplot(data=ncaa2012,
              aes(x=RPI,y=probability))

g2012<-g2012+geom_point() + ggtitle("2012 RPI vs. Projected Probability")
g2012

#2011
g2011<-ggplot(data=ncaa2011,
              aes(x=RPI,y=probability))

g2011<-g2011+geom_point() + ggtitle("2011 RPI vs. Projected Probability")
g2011

#2010
g2010<-ggplot(data=ncaa2010,
              aes(x=RPI,y=probability))

g2010<-g2010+geom_point() + ggtitle("2010 RPI vs. Projected Probability")
g2010


#Let's order these based on the probability
ncaa2013<-ncaa2013[order(-ncaa2013$probability),]
ncaa2012<-ncaa2012[order(-ncaa2012$probability),]
ncaa2011<-ncaa2011[order(-ncaa2011$probability),]
ncaa2010<-ncaa2010[order(-ncaa2010$probability),]

#Now we have a rank order of likelihood
```

```
#Different years have different amounts of spots not stolen by auto bids
#Luckily, you can use our trusty flag to figure out how many spots are legit
#This will show you how many teams should be considered "in" in each year

sum(ncaa2013$flag)
sum(ncaa2012$flag)
sum(ncaa2011$flag)
sum(ncaa2010$flag)

in.2013<-ncaa2013[1:47,c(1,16,18,19)]
in.2012<-ncaa2012[1:48,c(1,16,18,19)]
in.2011<-ncaa2011[1:50,c(1,16,18,19)]
in.2010<-ncaa2010[1:37,c(1,16,18,19)]

out.2013<-ncaa2013[48:346,c(1,16,18,19)]
out.2012<-ncaa2012[49:344,c(1,16,18,19)]
out.2011<-ncaa2011[51:341,c(1,16,18,19)]
out.2010<-ncaa2010[38:345,c(1,16,18,19)]

last.four.in.2013<-ncaa2013[44:47,c(1,16,19)]
last.four.in.2012<-ncaa2012[45:48,c(1,16,19)]
last.four.in.2011<-ncaa2011[47:50,c(1,16,19)]
last.four.in.2010<-ncaa2010[34:37,c(1,16,19)]

first.four.out.2013<-ncaa2013[48:51,c(1,16,19)]
first.four.out.2012<-ncaa2012[49:52,c(1,16,19)]
first.four.out.2011<-ncaa2011[51:54,c(1,16,19)]
first.four.out.2010<-ncaa2010[38:41,c(1,16,19)]

next.four.out.2013<-ncaa2013[52:55,c(1,16,19)]
next.four.out.2012<-ncaa2012[53:56,c(1,16,19)]
next.four.out.2011<-ncaa2011[55:58,c(1,16,19)]
next.four.out.2010<-ncaa2010[42:45,c(1,16,19)]

last.four.in.2013
first.four.out.2013
next.four.out.2013

#Subset to only teams that didn't get in

snub.2013<-ncaa2013[ncaa2013$Tournament.==0,]
snub.2013<-snub.2013[,c(1,15,19)]

snub.2012<-ncaa2012[ncaa2012$Tournament.==0,]
snub.2012<-snub.2012[,c(1,15,19)]

snub.2011<-ncaa2011[ncaa2011$Tournament.==0,]
snub.2011<-snub.2011[,c(1,15,19)]

snub.2010<-ncaa2010[ncaa2010$Tournament.==0,]
snub.2010<-snub.2010[,c(1,15,19)]


#Now let's Compare Flag vs. Predicted Probability
#first create predict in variable

ncaa2013$hit<-0
ncaa2013$hit[ncaa2013$probability>.4428]<-1

ncaa2012$hit<-0
ncaa2012$hit[ncaa2012$probability>.4143]<-1

ncaa2011$hit<-0
ncaa2011$hit[ncaa2011$probability>.3908]<-1

ncaa2010$hit<-0
ncaa2010$hit[ncaa2010$probability>.4467]<-1

#2013
g2013<-ggplot(data=ncaa2013,
              aes(x=hit,y=flag))
```

```
g2013<-g2013 + ggtitle("2013 Deserved Entry vs. Predicted Entry") + ylim(-1,2)
+xlim(0,1) + geom_point(position = position_jitter(w = 0.2, h =
0.2))+xlab("Predicted")+ylab("Deserved")
g2013

#2012
g2012<-ggplot(data=ncaa2012,
              aes(x=hit,y=flag))

g2012<-g2012+geom_point(position = position_jitter(w = 0.2, h = 0.2)) + ggtitle("2012
Deserved Entry vs. Predicted Entry")+ ylim(-1,2)
+xlim(0,1)+xlab("Predicted")+ylab("Deserved")
g2012

#2011
g2011<-ggplot(data=ncaa2011,
              aes(x=hit,y=flag))

g2011<-g2011+geom_point(position = position_jitter(w = 0.2, h = 0.2)) + ggtitle("2011
Deserved Entry vs. Predicted Entry")+ ylim(-1,2)
+xlim(0,1)+xlab("Predicted")+ylab("Deserved")
g2011

#2010
g2010<-ggplot(data=ncaa2010,
              aes(x=hit,y=flag))

g2010<-g2010+geom_point(position = position_jitter(w = 0.2, h = 0.2)) + ggtitle("2010
Deserved Entry vs. Predicted Entry")+ ylim(-1,2)
+xlim(0,1)+xlab("Predicted")+ylab("Deserved")
g2010



#Let's start making some graphics that tell us how well we did


#2013

in.2013$predict<-"Predict IN"
out.2013$predict<-"Predict OUT"
final.2013<-rbind(in.2013,out.2013)
temp.2013<-final.2013[final.2013$flag==1,]
temp.2013$flag1<-"Actual IN"
temp.1.2013<-final.2013[final.2013$flag==0,]
temp.1.2013$flag1<-"Actual OUT"
final.2013<-rbind(temp.2013,temp.1.2013)
remove(temp.2013)
remove(temp.1.2013)
table.2013<-table(final.2013$predict,final.2013$flag1)
p.table.2013<-prop.table(table.2013,margin=1)

#2012

in.2012$predict<-"Predict IN"
out.2012$predict<-"Predict OUT"
final.2012<-rbind(in.2012,out.2012)
temp.2012<-final.2012[final.2012$flag==1,]
temp.2012$flag1<-"Actual IN"
temp.1.2012<-final.2012[final.2012$flag==0,]
temp.1.2012$flag1<-"Actual OUT"
final.2012<-rbind(temp.2012,temp.1.2012)
remove(temp.2012)
remove(temp.1.2012)
table.2012<-table(final.2012$predict,final.2012$flag1)
p.table.2012<-prop.table(table.2012,margin=1)

#2011

in.2011$predict<-"Predict IN"
out.2011$predict<-"Predict OUT"
final.2011<-rbind(in.2011,out.2011)
```

```
temp.2011<-final.2011[final.2011$flag==1,]
temp.2011$flag1<-"Actual IN"
temp.1.2011<-final.2011[final.2011$flag==0,]
temp.1.2011$flag1<-"Actual OUT"
final.2011<-rbind(temp.2011,temp.1.2011)
remove(temp.2011)
remove(temp.1.2011)
table.2011<-table(final.2011$predict,final.2011$flag1)
p.table.2011<-prop.table(table.2011,margin=1)

#2010

in.2010$predict<-"Predict IN"
out.2010$predict<-"Predict OUT"
final.2010<-rbind(in.2010,out.2010)
temp.2010<-final.2010[final.2010$flag==1,]
temp.2010$flag1<-"Actual IN"
temp.1.2010<-final.2010[final.2010$flag==0,]
temp.1.2010$flag1<-"Actual OUT"
final.2010<-rbind(temp.2010,temp.1.2010)
remove(temp.2010)
remove(temp.1.2010)
table.2010<-table(final.2010$predict,final.2010$flag1)
p.table.2010<-prop.table(table.2010,margin=1)


heights<-c(.9595,.0405,.9767,.0233,.9531,.0469,.9362,.0638)
mydata<-matrix(heights,
               ncol=2,
               byrow=T,
               dimnames=list(c("2013","2012","2011","2010"),
                             c("Correct","Incorrect")))

colors<-c("green","red")

barplot(t(mydata),
        beside=T,
        horiz=F,
        col=colors,
        ylim=c(0,1.4),
        axes=F,
        )
axis(2,at = c(0,.2,.4,.6,.8,1),labels = c("0%","20%","40%","60%","80%","100%"),las=T)
title(main=("Model Performance by Year"),xlab="Year")
legend("topright",colnames(mydata),fill=colors,bty="n")
```