**Executive Summary & Goal**

**I. Introduction**

College basketball’s national championship tournament, or “March Madness”, is an annual national phenomenon. Over 70 million Americans fill out March Madness brackets every year, and never once has there been a perfect bracket. The elusiveness of a perfect bracket grips the American psyche so intensely that even famed investor Warren Buffet once offered $1B if anyone could predict a perfect bracket. His money has been safe – the odds of a perfect bracket have been estimated to be close to 1 in 9.2 quintillion (18 zeros).

Although there are a large number of games during the regular season for basketball, the one-game elimination system in the 64-team tournament produces untoward prediction effects. In the NBA, teams in the playoffs face each other multiple times before determining a winner, reducing the randomness and unpredictability. In college basketball, the single-game elimination introduces a large element of randomness and unpredictability. Every year, multiple “Cinderella” upsets wipe out any chance of a perfect bracket early (in 2018, 0 of 17.3mm ESPN brackets were perfect by the Sweet Sixteen round). While we do not believe we can overcome the 1 in 9.2 quintillion odds, the goal of this analysis is to explore relationships between regular season stats and tournament upsets that we can use to improving the probability of predicting upsets.

**A. Tournament Overview**

An upset happens when a lower-seeded team in the tournament knocks out a higher-seeded team. The tournament is made up of 64 teams total. 32 of those teams earned automatic bids from being their respective conference champions, and then 36 teams are granted “at-large” bids from a Selection Committee (four of the lower seeded of these teams will be eliminated in a preliminary “First Four” games, leaving 64 teams in the tournament). The Selection Committee then ranks all 68 teams from 1 to 68. It distributes the top four teams across four regions and gives each a No. 1 seeding within that region, then the next top four teams (teams ranked 5-8) get distributed across the four regions with each receiving a No. 2 seeding, etc. This process continues until all 68 teams are distributed among the four regions (typically East, West, South, and Midwest).

There exist other considerations that affect the Selection Committee’s seeding distribution. For instance, the Selection Committee tries to distribute the teams across regions so that teams in the same conference cannot meet until regional finals. The Selection Committee additionally tries to avoid regular season or tournament rematches during the early rounds. These factors, along with several others, can artificially move a team up or down a seeding, affecting the level of teams they face off against in the early rounds.

Finally, once all the teams have been distributed across all four regions and given their seeding number, the top eight teams play the bottom eight teams in each region (e.g. No. 1 plays No. 16, No. 8 plays No. 9). Once it becomes semi-finals, the regions are set up such that the winner of the top ranked region plays the winner of the bottom ranked region, and the champion of the second-ranked region plays the champion of the 3rd-ranked region.

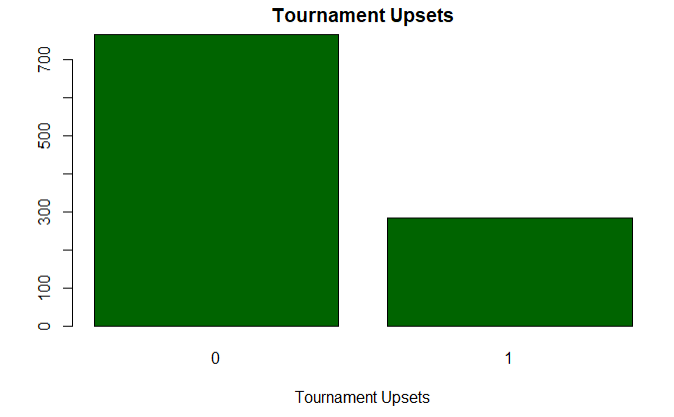
Games are played on neutral territory (no games prior to the Final Four can be played on one team’s home court), however games may be played in a team’s home state. Final Four games may happen to be played on a team’s home court as these venues are scheduled years ahead of time.

**II. Data Description**

The data was provided as a massive collaboration between the NCAA and Google. The data contained multiple files of information on coaches, players, scores, seedings, and plays in both the regular season and during the championship tournament. Because March Madness brackets can only be filled out before the tournament starts, we focus primarily on stats from the regular season. Thus, the data files used from what Google provided are the NCAA tournament result files, team name and conference name files, coach name files, and finally regular season stats files. All data begins in 2003

Across the 15 March Madness tournaments included in the dataset, there were 1,048 total games played. 283 of these games qualified as upsets. While there are certainly a smaller number of upset-games to train the data on than non-upset games, the data set is not severely imbalanced.

Exhibit <>: Total number of Tournament Upsets since 2003



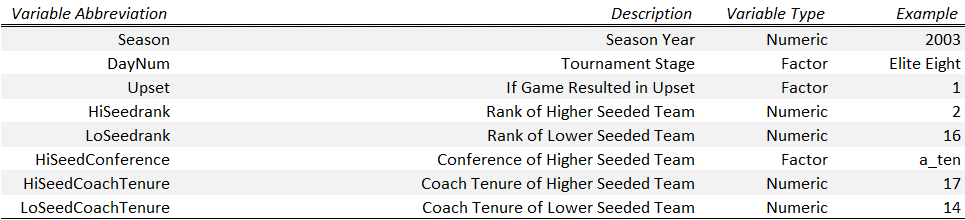
**A. Data Cleaning**

Because the data was split across 6 different files, a major portion of data cleaning was choosing relevant information from different files and concatenating it into a centralized data frame.

For tournament-related variables, we first added seed rankings to each March Madness tournament game in order to be able to track if a win during the tournament qualified as a an upset (an upset was defined as when a lower-seeded team beat a higher-seeded team). Upset takes on the value “1” if it is an upset, and a 0 otherwise. Next, we created the variable Coach Tenure that determines how long a coach was active for. A coach with a long career may be someone who has more innate talent than someone with a short career. Additionally, we added the conference for the higher seeded team to see if there are any particularly susceptible conferences. We did not add the conference for the lower seeded-team because this variable was severely imbalanced. Nearly all teams seeded No. 8-16 receive an “at-large” bid, so “at large” was the conference title for nearly all lower ranked teams. Finally, the variable tracking the “day number” of a season that a tournament game was played on was converted into a categorical variable that instead indicated which stage of the tournament a game was played during to see if there was a relationship between tournament stage and the occurrence of an upset.

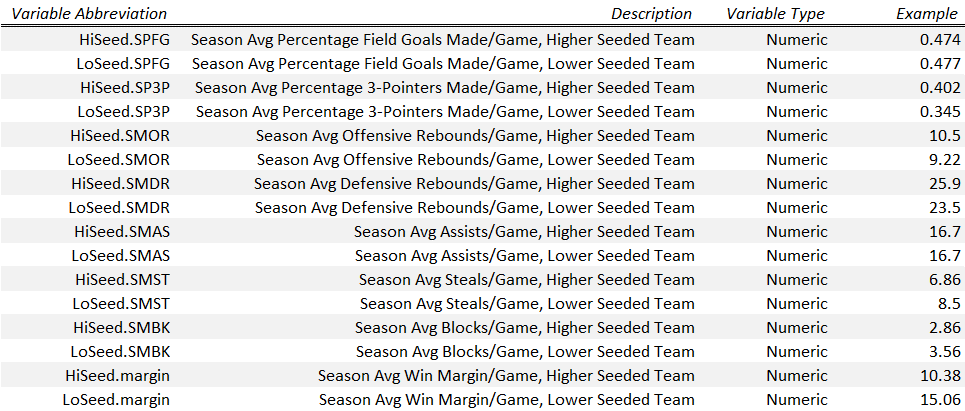
A summary of variables related broadly to the tournament data are shown in the table below.

Table <>: Variables, Tournament



Next, we created a set of variables based on performance statistics from the regular season. We took a team’s offensive and defensive statistics for their regular season games won, and averaged these across a season to determine if a team who was particularly good at one trait or set of traits could be related to creating an upset. The table of variables and their meanings specific to the regular season are shown below.

Table <>: Variables, Regular Season

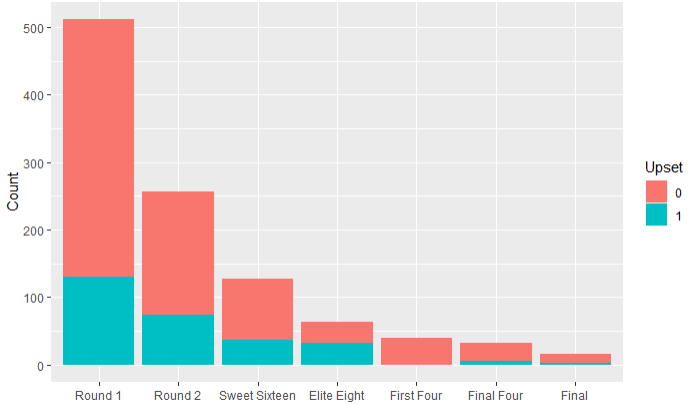


**B. Data Exploration**

In this section, we first split the data by tournament-related variables vs. regular-season related variables to assess what the profile of each set of predictor variables looks like.

For tournament-related variables, we first look at whether the number of upsets decreases the deeper we progress into the tournament. The results are shown in Exhibit <> below. Interestingly, more than 1 in 5 games in the first round are upsets. This pattern of every 1 in 5 games being an upset continues through the next several tournament rounds, with the Elite Eight round having the highest number of “upsets”. However, it is possible that by this stage, all top teams and seed rankings become less meaningful. Remember, rankings are artificially increased or decreased based on the Selection Committee trying to fulfill criteria outside of projected performance, so differences between the top rankings may not be meaningful. If the remaining Elite Eight are all top ranked, people predicting their brackets at this point based on seeding may become less meaningful.

Exhibit <>: Number of Upsets by Tournament Stage

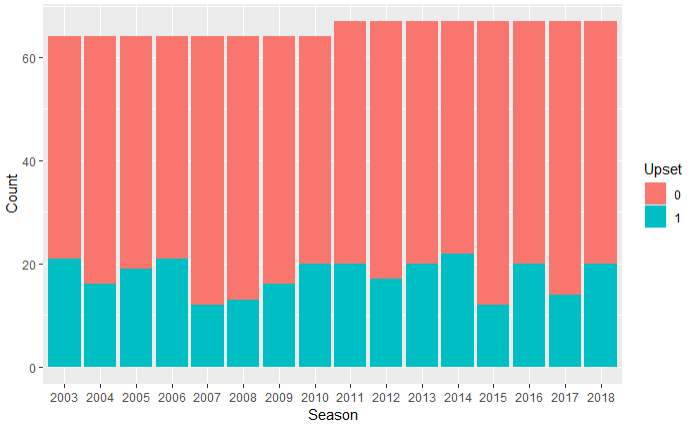


Naturally, in the next exhibit below, we look at the seed rankings for the remaining teams in each round.

**Insert seeds by rounds**

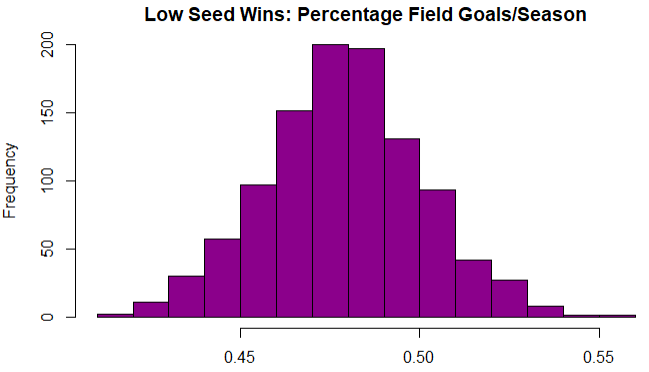
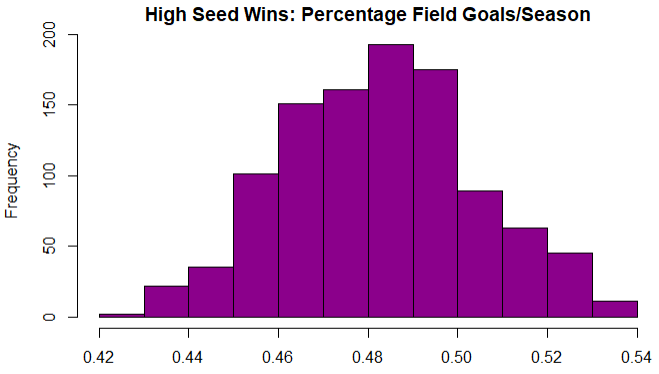
We also track the proportion of upsets by year, to see if there has been any pattern over time. Overall, it appears that the average number of upsets is slightly under 20% over time, and that it randomly fluctuates year over year.

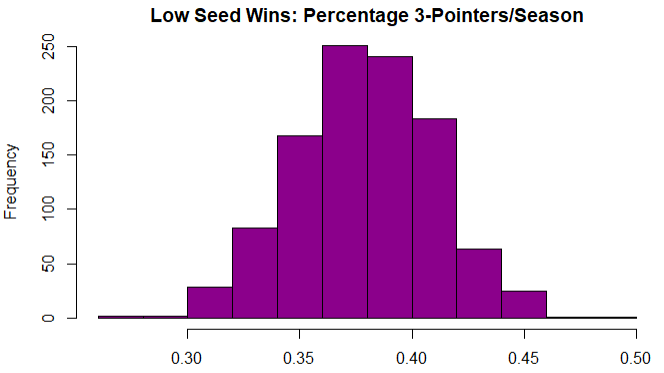
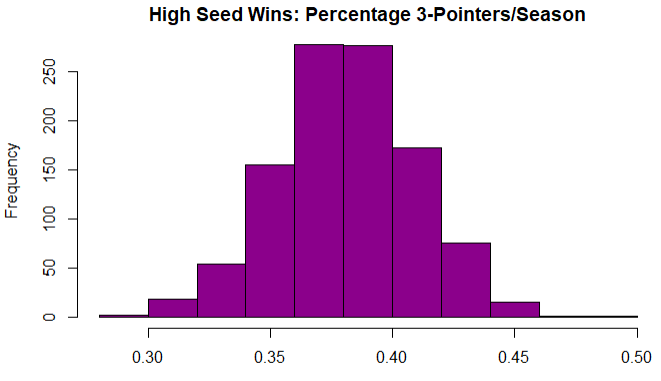
Exhibit <>: Proportion of Upsets by Season



Interestingly, when we look at performance statistics for the games that teams won during the regular season, the distribution is near normal across all offensive and defensive statistics. Additionally, the distribution is very similar between the higher seeded teams and lower seeded teams. That is, during the tournament match up, the distribution of the regular season stats for the lower seeded team is very similar to the distribution for the regular season stats for the higher seeded team. In Exhibits <> and <>, example distributions and means look very similar between the lower seeded teams and higher seeded teams. This could also help explain the high incidence of upsets in March Madness – it is possible that many teams are actually similar in skill level to one another and on average there is a not an abnormally large chasm between a higher seeded team and a lower seeded team’s ability, thus increasing the unpredictability of a one-game elimination tournament. For instance, as shown in the exhibits below, teams tend to have similar average percentage field goals and three pointers made during the regular season.

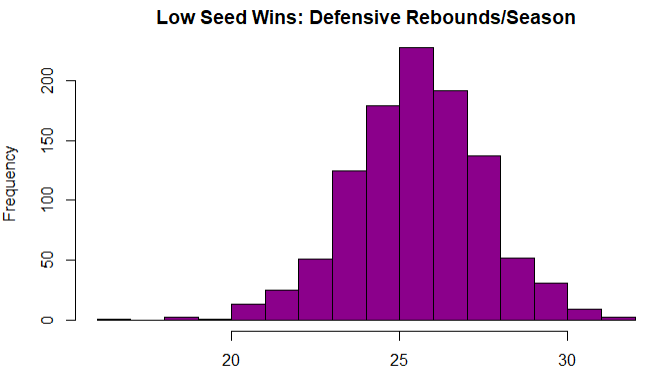
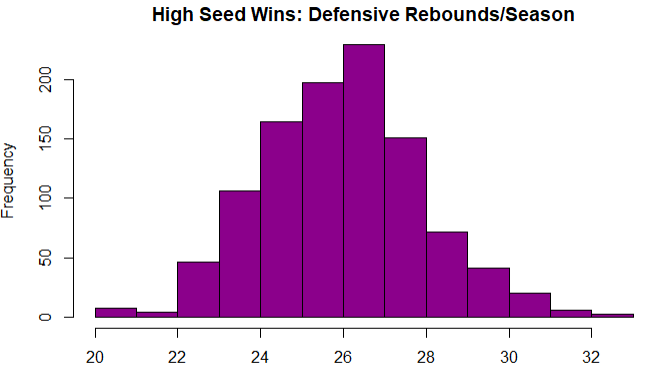
Exhibit <>: Distribution of Regular Season Offensive Stats: High vs. Low Seeds

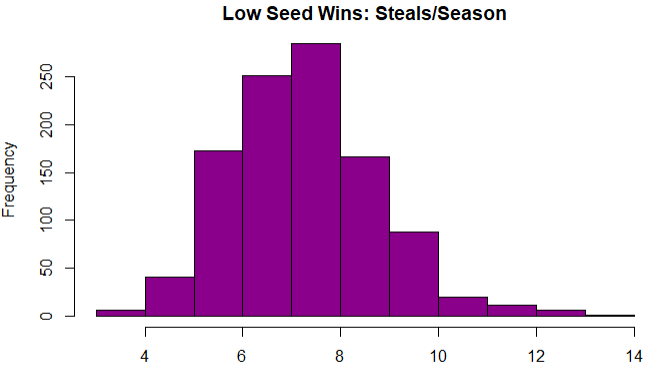
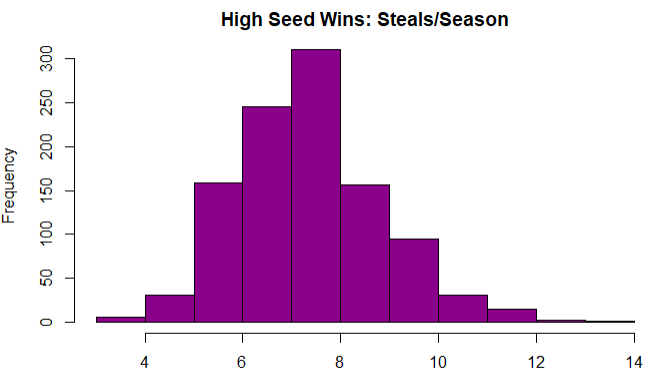


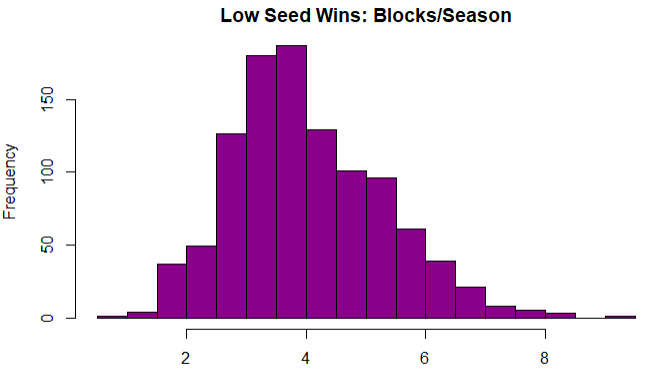
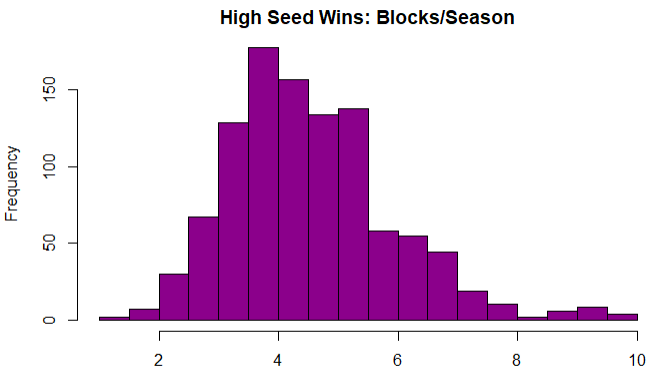


The distributions of defensive stats during the regular season split between higher seeded teams and lower seeded teams appear only slightly more skewed. However, similar to the scenario with offensive statistics, there does not appear to be a meaningful difference in the average defensive ability between the two types of teams. The only area where the average for the higher seed looks a little better is in average blocks per game per season. This could possibly indicate that a differentiator between higher seeded teams and lower seeded teams is their defensive ability.

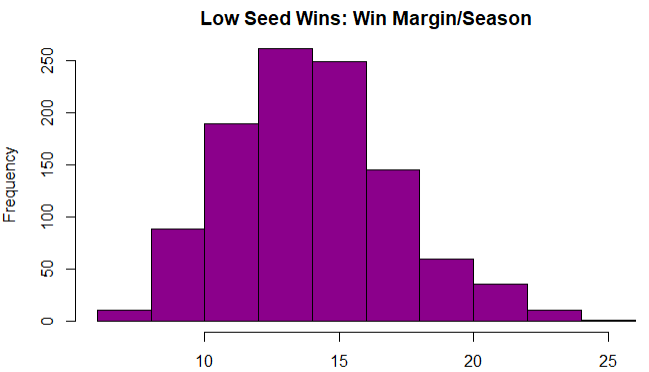
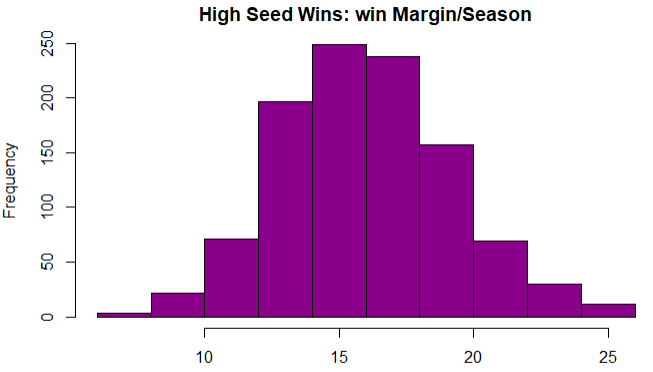
Exhibit <>: Distribution of Regular Season Defensive Stats: High vs. Low Seeds





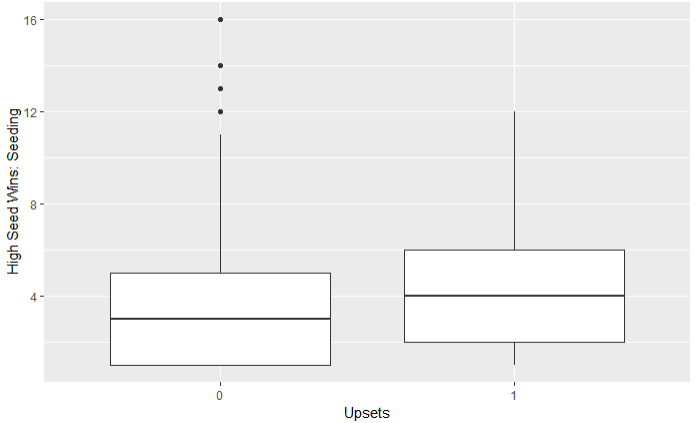
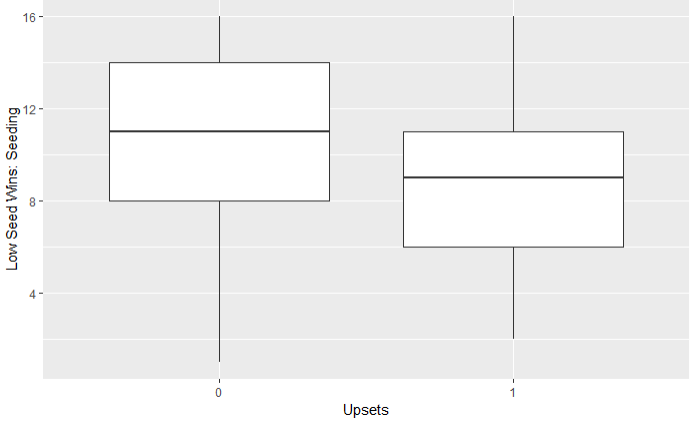


Finally, as shown in the exhibit below, teams that are the low seeds in tournament face off tend appear to have smaller average win margins relative to when high seeds in the tournament win during the regular season.



In the previous histograms, the distributions for offensive and defensive statistics for both highly and lowly ranked teams did not look meaningfully different. This was further confirmed using a box plot chart to evaluate whether these regular-season characteristics made any meaningful difference during the tournament and had a relationship with whether a lower seeded team upset a higher seeded team. Unfortunately, the vast majority of the boxplots for a given predictor variable looked identical for games that were upsets versus games that were not upsets. The box plot with the largest visible difference, however still not a significant difference, is shown below in Exhibit <>. The relationship implies that the lower a low-seeded team is ranked, the less likely an upset is. Similarly, the lower a high seed team is ranked, the more likely an upset is.

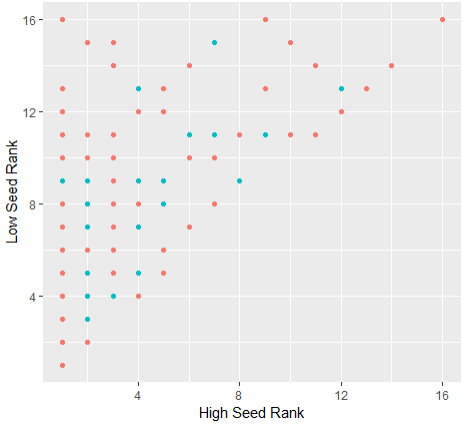
Exhibit <>: Boxplot for Seed Rank vs. Tournament Upset



We can try to dive in deeper to this relationship between seeding and likelihood of an upset in Exhibit <> below. It appears that No. 2 and 3 seeds in a region tend to be upset by the widest range of possible lower seeds. Conversely, the higher seeded team in a match up is appears less frequently upset when the team is either at the very top or very bottom of its region. For instance, if the higher seed in a match up is ranked No. 1, or is ranked near the bottom at No. 15 (e.g. it’s a “First Four” game), there appears to be far fewer upsets. The below scatterplot also shows that No. 1 seeds are typically only upset by No. 9 ranked teams (that is, the best team from the at-large conference bids).

Exhibit <>: Scatterplot of Upsets based on High Seed & Low Seed Rank

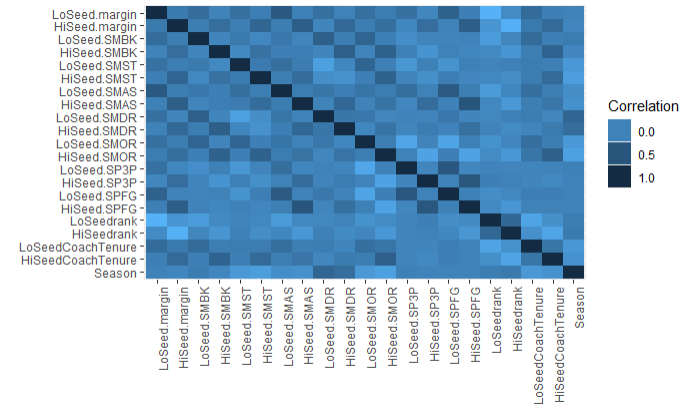




**C. Identifying Important Variables**

In this next section, we split the predictor variables by continuous and categorical variables and examine their correlations and significance. The correlation heat map of the continuous variables is shown in Exhibit <> below. Overall, there is no high collinearity or groups of variables that have a strong correlation with each other. Darker squares, indicated closer to 0.5 correlation, appear more randomly scattered.

Exhibit <>: Correlation Heatmap: Continuous Predictor Variables



Running a test for significance with just continuous variables produces the output table below. Variables significant at the 5% level are in bold. We removed the variable with the highest Pr(>Chisq) below, and then reran ANOVA. This was repeated multiple times until we were able to condense the number of continuous variables down to a more manageable 13 variables. The variables removed were Low Seed Season Mean Assists per game, Low Seed Coach Tenure, Low Seed Season Mean Offensive Rebounds per game, Low Seed Season Mean Steals per game, High Seed Season Mean Steals per game, Low Seed Season Blocks per game, and High Seed Season Mean Offensive Rebounds per game. Interestingly, more statistics related to the lower seeded team were removed than the higher seeded team, potentially indicating that it is the quality of the higher seeded team that matters more when it comes to supsets.

Table <>: Analysis of Deviance Table (Type II Test), Continuous Variables

----------------------------------------------------

ANOVA Table LR Chisq Df Pr(>Chisq)

----------------------- ---------- ---- ------------

\*\*HiSeedCoachTenure\*\* 1.306 1 0.2532

\*\*LoSeedCoachTenure\*\* 0.04908 1 0.8247

\*\*HiSeed.SPFG\*\* 1.539 1 0.2147

\*\*HiSeed.SP3P\*\* 6.102 1 0.0135

\*\*HiSeed.SMOR\*\* 0.9806 1 0.3221

**\*\*HiSeed.SMDR\*\* 5.29 1 0.02144**

\*\*HiSeed.SMAS\*\* 3.626 1 0.05687

\*\*HiSeed.SMST\*\* 0.6371 1 0.4248

\*\*HiSeed.SMBK\*\* 0.264 1 0.6074

**\*\*HiSeed.margin\*\* 29.96 1 4.412e-08**

\*\*LoSeed.SPFG\*\* 0.8527 1 0.3558

**\*\*LoSeed.SP3P\*\* 4.126 1 0.04222**

\*\*LoSeed.SMOR\*\* 0.04916 1 0.8245

**\*\*LoSeed.SMDR\*\* 7.175 1 0.007393**

\*\*LoSeed.SMAS\*\* 0.03316 1 0.8555

\*\*LoSeed.SMST\*\* 0.3754 1 0.5401

**\*\*LoSeed.SMBK\*\* 3.135 1 0.07663**

**\*\*LoSeed.margin\*\* 6.752 1 0.009363**

\*\*HiSeedrank\*\* 1.029 1 0.3105

\*\*LoSeedrank\*\* 35.36 1 2.738e-09

----------------------------------------------------

We repeat the exercise above with the set of categorical variables, and removed the variable Season.

----------------------------------------------------

ANOVA Table LR Chisq Df Pr(>Chisq)

----------------------- ---------- ---- ------------

\*\* Season \*\* 0.200 1 0.6548

**\*\* DayNum \*\* 32.272 6 1.447e-05**

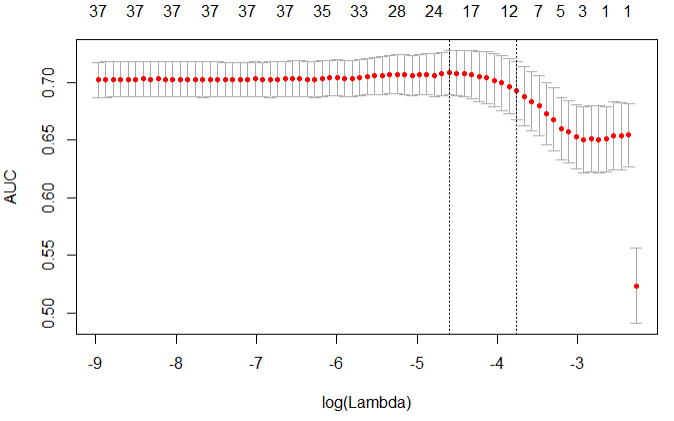
\*\* HiSeedConference \*\* 23.067 28 0.7297

----------------------------------------------------

**II. Model Build & Evaluation**

**A. Logistic with Lasso**

After got LASSO model, did backwards iteration of rerunning through glm & removing non-significant variables



------------------------------------------------------------------------

Summary Table Estimate Std. Error z value Pr(>|z|)

------------------------- ---------- ------------ ---------- -----------

\*\*(Intercept)\*\* 0.9215 1.762 0.5231 0.6009

\*\*HiSeed.SP3P\*\* 7.014 3.403 2.061 0.03928

\*\*LoSeed.margin\*\* 0.09909 0.03475 2.852 0.004345

\*\*DayNumFinal\*\* -1.811 0.8581 -2.111 0.03477

\*\*DayNumFinal Four\*\* -1.941 0.65 -2.987 0.002819

\*\*DayNumFirst Four\*\* -17.79 479.1 -0.03713 0.9704

\*\*DayNumRound 1\*\* -0.707 0.4943 -1.43 0.1527

\*\*DayNumRound 2\*\* -0.7219 0.4065 -1.776 0.07574

\*\*DayNumSweet Sixteen\*\* -0.8305 0.4064 -2.044 0.04098

\*\*LoSeed.SP3P\*\* -6.39 3.006 -2.126 0.03354

\*\*HiSeedrank\*\* 0.181 0.04767 3.796 0.0001472

\*\*LoSeedrank\*\* -0.1669 0.03914 -4.265 2.003e-05

\*\*HiSeed.margin\*\* -0.1208 0.03496 -3.454 0.0005515

------------------------------------------------------------------------

------------------------------------------------

ANOVA Table LR Chisq Df Pr(>Chisq)

------------------- ---------- ---- ------------

\*\*HiSeed.SP3P\*\* 4.295 1 0.03822

\*\*LoSeed.margin\*\* 8.186 1 0.004221

\*\*DayNum\*\* 42.33 6 1.58e-07

\*\*LoSeed.SP3P\*\* 4.559 1 0.03275

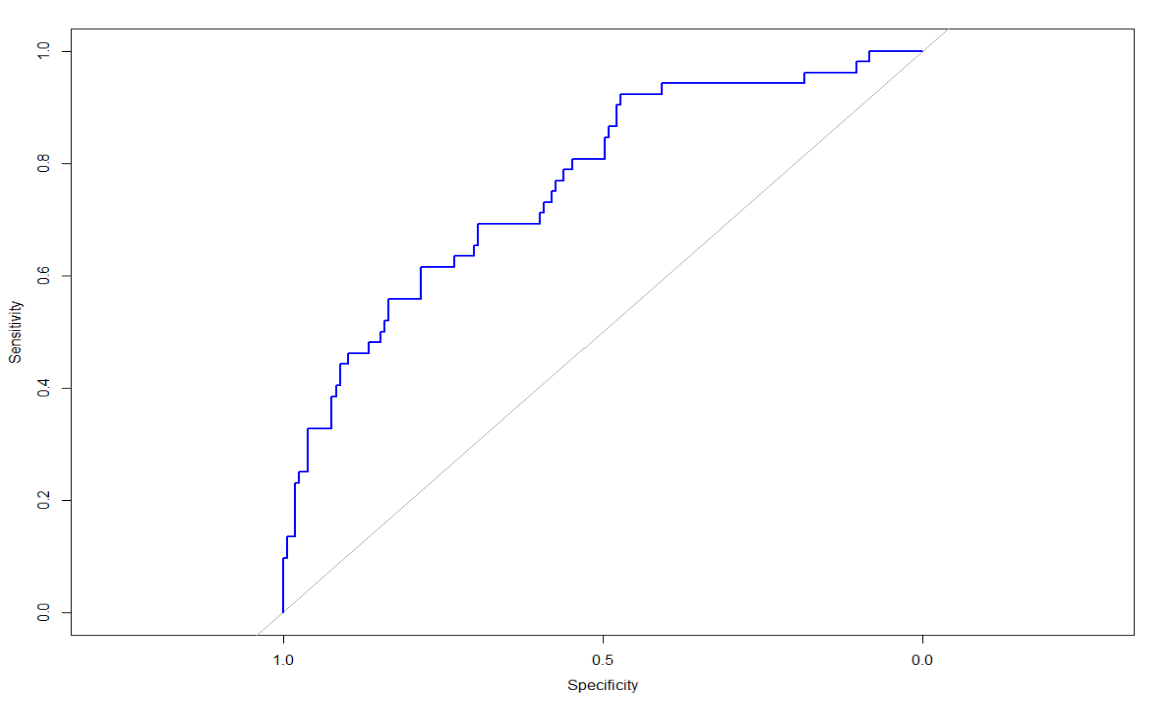
\*\*HiSeedrank\*\* 14.07 1 0.0001757

\*\*LoSeedrank\*\* 19.12 1 1.226e-05

\*\*HiSeed.margin\*\* 12.38 1 0.0004346

------------------------------------------------

Table: Analysis of Deviance Table (Type II tests)



AUC is 0.7626, high!

**B. Linear with Lasso**

---------------------------------------------------------------------------

Summary Table Estimate Std. Error t value Pr(>|t|)

------------------------------ ---------- ------------ --------- ----------

\*\*(Intercept)\*\* 1.038 0.2388 4.349 1.564e-05

\*\*LoSeed.margin\*\* 0.01443 0.005907 2.443 0.01481

\*\*factor(DayNum)Final\*\* -0.3624 0.1403 -2.582 0.01001

\*\*factor(DayNum)Final Four\*\* -0.3793 0.1085 -3.497 0.0004992

\*\*factor(DayNum)First Four\*\* -0.6917 0.1704 -4.059 5.457e-05

\*\*factor(DayNum)Round 1\*\* -0.1272 0.09061 -1.404 0.1607

\*\*factor(DayNum)Round 2\*\* -0.1602 0.0768 -2.086 0.03733

\*\*factor(DayNum)Sweet -0.1789 0.07822 -2.287 0.0225

Sixteen\*\*

\*\*LoSeed.SP3P\*\* -1.044 0.4989 -2.092 0.03676

\*\*HiSeedrank\*\* 0.03141 0.008025 3.915 9.913e-05

\*\*LoSeedrank\*\* -0.02833 0.006614 -4.284 2.088e-05

\*\*HiSeed.margin\*\* -0.01544 0.005449 -2.833 0.004734

---------------------------------------------------------------------------

---------------------------------------------------------

ANOVA Table Sum Sq Df F value Pr(>F)

-------------------- -------- ----- --------- -----------

\*\*LoSeed.margin\*\* 1.026 1 5.967 0.01481

\*\*factor(DayNum)\*\* 7.353 6 7.124 2.168e-07

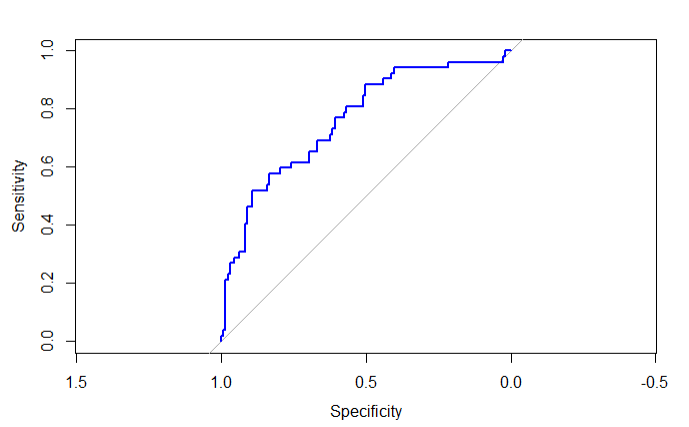
\*\*LoSeed.SP3P\*\* 0.7531 1 4.378 0.03676

\*\*HiSeedrank\*\* 2.636 1 15.32 9.913e-05

\*\*LoSeedrank\*\* 3.157 1 18.35 2.088e-05

\*\*HiSeed.margin\*\* 1.381 1 8.028 0.004734

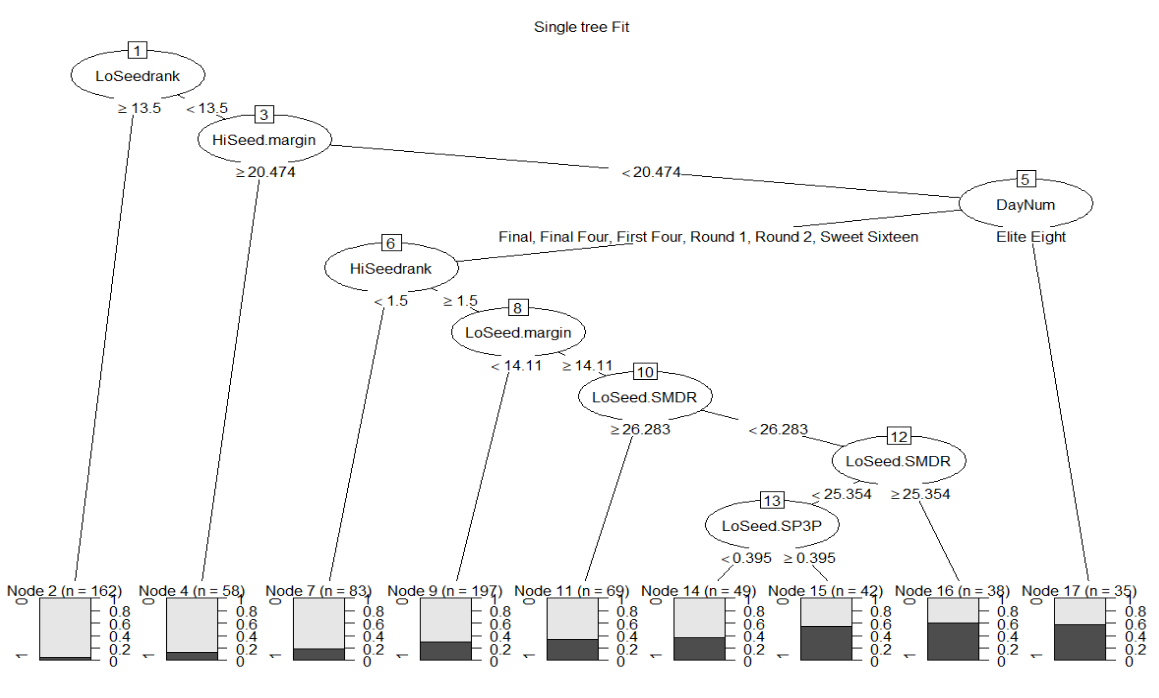
---------------------------------------------------------

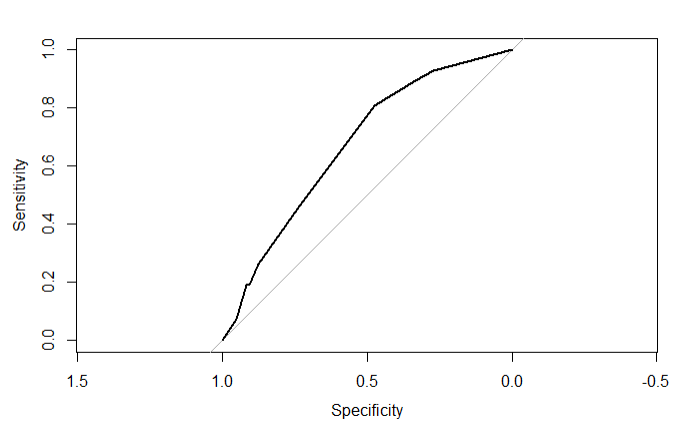


AUC is 0.7615

**C. Single Tree & Random Forest**

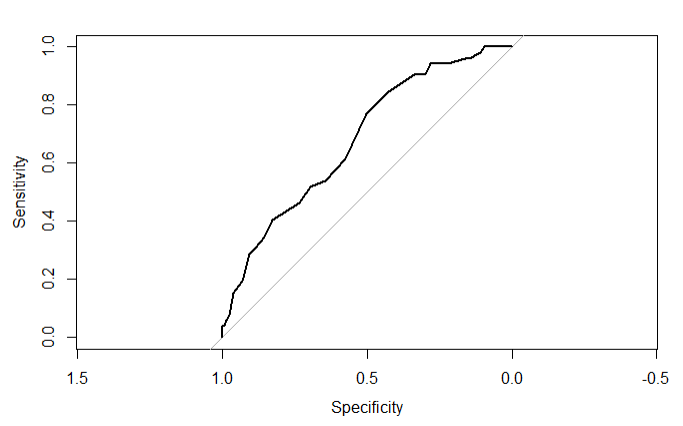
Single Tree –





AUC is 0.6915

Random forests –



Optimized mtry = 1, ntree = 50

AUC is 0.6802

**D. PCA**

-------------------------------------------------------------------

HiSeedrank LoSeedrank HiSeed.SPFG LoSeed.SPFG HiSeed.SP3P

------------ ------------ ------------- ------------- -------------

-0.3812 -0.261 0.3999 0.08524 0.2252

-------------------------------------------------------------------

---------------------------------------------------------------------

LoSeed.SP3P HiSeed.SMDR LoSeed.SMDR HiSeed.SMAS LoSeed.SMBK

------------- ------------- ------------- ------------- -------------

0.01398 0.265 0.01032 0.4093 0.05241

---------------------------------------------------------------------

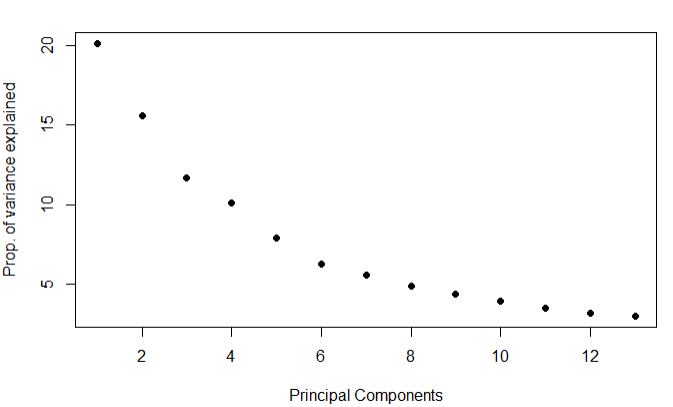
---------------------------------------------------

HiSeed.margin LoSeed.margin HiSeedCoachTenure

--------------- --------------- -------------------

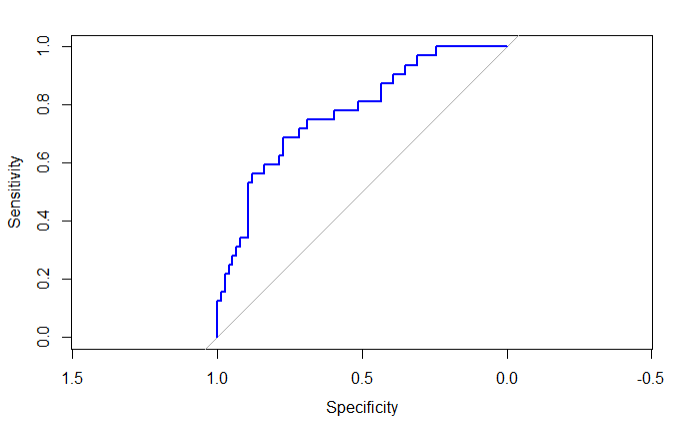
0.4834 0.1631 0.2601

---------------------------------------------------



**E. Final Model & Testing using Validation Data Set**

Model chosen: Logistic w/ Lasso



AUC = 0.7736

**IV. Conclusion**

**V. Drawbacks & Further Exploration**