**IST 687**

**Final Project**

**December 2022**

**Joseph Davis**

**INTRODUCTION**

This research project looked at PGA tour data from the 2010-2018 seasons. The intention of this project was to analyze top level golf play and understand the elements that make these players the best in the world. This project also aimed to help others understand what is needed of top-level players to excel and to inspire amateur players to prescriptively fix their game by focusing on the most important elements of the game. The data was scraped from the PGA website by Kaggle user Jong who conveniently put all the raw data into a CSV. The data itself had 2312 rows and 18 variables. This raw data was imported, cleaned, explored, analyzed and used to generate actionable insights. Generated from this data was descriptive statistics/visuals, multiple models, ANOVA statistics, and a regression tree describing which factors are most influential. Based on the goals at hand 3 business questions were created with the hope to create better golf players in the future.

**BUSINESS QUESTIONS**

1. What factors make the best players in the world, the best?
2. How should one improve their game to become a professional?
3. How are the best players in the world improving season to season?

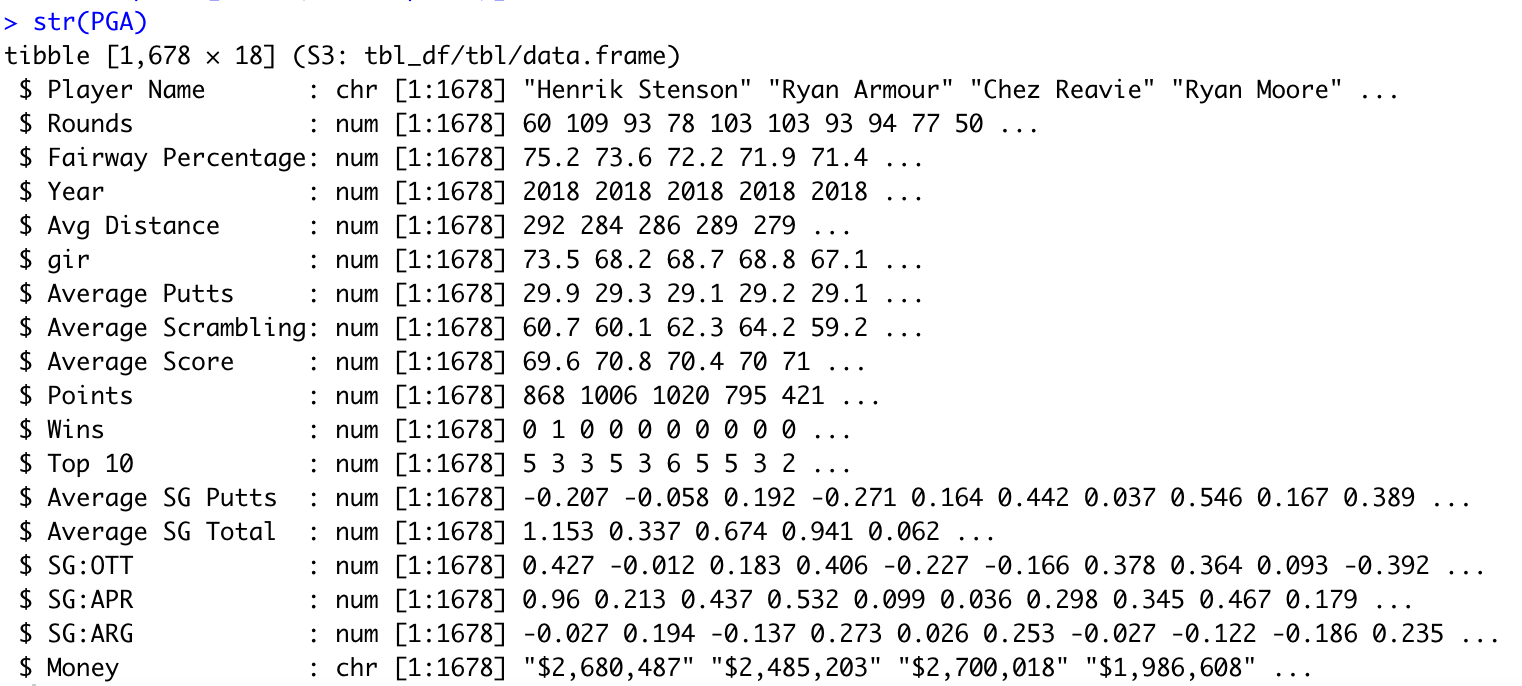
**DATA ACQUISITION/PREPARATION**

The data was acquired from Kaggle user Jong and a link can be seen at the very beginning of the raw R code at the end of this report. The raw csv file was read into the R-data frame and explored with a few commands. Of the 2312 rows within the dataframe, a significant amount of the data was incomplete. These were exclusively players who only had data on the year they played/earnings/points earned. These were lesser-known players and in terms of their contribution to the PGA tour, they were unnecessary. For this reason, they were removed, leaving the dataset with 1678 rows.

The variables within this data set included both qualitative and quantitative data. The variables were as follows:

* Player Name – Player Name
* Rounds – Rounds of golf played that year in a tournament
* Fairway Percentage – Off the tee how often do they hit the fairway
* Year – The PGA tour year these stats are from
* Avg Distance – Average driver distance
* Gir (%) – green in regulation (getting the ball onto the green and being able to 2 putt to get par)
* Average Putts – Average number of putts per round of golf
* Average Scrambling – Average numbers of times a golfer missed the green in regulation but still finished with par or better
* Average Score – Average strokes per round of golf
* Points – Ranking points, gained from doing well in tournaments
* Wins – How many tournaments did this person win this year
* Top 10 – How many finishes did this person have in the top 10 that year
* Average SG Putts - On average how many strokes a golfer gains compared to the rest of the field based on the quality of his putts
* Average SG Total - On average how many strokes a golfer gains compared to the rest of the field
* SG: OTT – On average how many strokes a golfer gains compared to the rest of the field based on the quality of his Drive
* SG: APR – Strokes Gained - On average how many strokes a golfer gains compared to the rest of the field based on the quality of his approach shot to the green
* SG: ARG – Strokes Gained approach to the green. On average how many strokes a golfer gains compared to the rest of the field based on the quality of play around the green
* Money – Earnings for that year

Next, the `Wins` and `Top 10` variables needed to be changed. If the player had 0 wins or 0 top 10 finishes, then there was an NA in their cell. Any variables containing an NA in the `Top 10` or `Wins` variable was changed to 0. This now complete main dataframe was named PGA and it was kept as the base dataset, from which other dataframes could be created. For this reason, all column variables were kept. Running a str() command on the dataframe yields the following.



Next multiple different subsets of the data were created to be used for analysis. The entire code can be seen at the end of this document in the appendix.

* 8 separate dataframes containing all variables were created, 1 for each year, in this way each specific year could be studied with depth.
* From these dataframes, the best 25 players were pulled and then combined using rbind() to have a dataframe containing the best 25 players from each year.
* A dataframe containing all the PGA data was created, however instead of an average score, the dataframe would contain a 0 if the player had 70 or above within the `Average Score` variable and a 1 if the player had below a 70 in the `Average Score` variable.
* Finally, a dataframe was created that contained the average stats of the top 25 players from each year. This yielded a dataframe that was 8 rows, 1 for each year.

**DESCRIPTIVE STATISTICS**

A list of the top 10 players from the year 2010-2018 is shown below. Notice some names appear multiple times, this is because they are from different seasons. It’s important to show players from each year as the courses change year to year so there is variation between seasons.

Table

Description automatically generated

A histogram showing the amount earned for each player in 2018 was created. This histogram showed that the earnings from the 2018 year softly followed a pareto distribution. This made me curious as to the difference in play that justified a $5million difference.

Chart, histogram

Description automatically generated

To explore what causes such a dramatic in earnings a boxplot was created to see a spread of average scores from the 2018 season. Remarkable to see that an earnings increase of $5 million can be attributed to an average score that is only 2 strokes lower.

Chart

Description automatically generated

To confirm that the people making more money were the same ones with a lower average score, a scatterplot was created. A downward trend is observed, displaying that a lower average score will result in more earnings.

Chart, scatter chart

Description automatically generated

Finally, if winning is a larger motivation for becoming better at golf, 2 scatterplots were made. Both having average score on the y-axis, 1 containing wins on the x axis, and one containing top 10 finishes on the x-axis. This again displays the trend that having a lower average score contributes to performing better in tournaments.

Chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

**MODELING**

It’s generally recognized that hitting a ball solidly and accurately is more important than driving the ball as far as you possibly can. I wanted to show this accuracy vs. distance relationship so I quickly made a scatterplot containing a trendline.

The was created using the best players from every year dataframe. There is a clear relationship, as distance increases accuracy decreases.

Chart, scatter chart

Description automatically generated

With the insight from this initial graph the first linear model was created by looking at the accuracy off the tee box from the 2018 season. For this model the variables, driving distance and strokes gained off the tees were used. This yielded a p-value of the F stastic that was less than .05, and an adjusted R2 value of .7174. Both variables were statistically significant.

Text

Description automatically generated

Next, I wanted to look at how variables within the players statistics would affect their average score for the season. Another linear model was created, this time the y-variable was average score, and the x-variables were greens in regulation, average putts, average scrambling, average distance, and fairway percentage. Additionally, the entire PGA dataset was used, not just the 2018 season. The p-value of the F statistic was less than 0.05 so we can use this model. All the variables were statistically significant and the R2 value was 0.6892. Of the variables, the average putts is the only statistic with a positive relationship to the average score. This means that as the average number of putts increases, so too does the average score. Both the average distance and fairway percentage have negative relationships with overall score, suggesting that a further and more accurate drive will decrease the overall average score for the year. Additionally, the green in regulation and average scrambling both have negative values. Indicating that if you get to green in regulation more, you will have a lower average score, and if you are good at scrambling, you will have a lower average score. These all make sense logically.

Table

Description automatically generated

Finally, I wanted to create an SVM model to predict the average score of players. First, I needed to convert the average score variable into a binary one. This was done by creating a new dataframe where the average score was 1 if the person had an average score below 70, and a 0 if their average score was above 70. 70 was chosen because only about 8.5%% of individuals in this dataset scored below a 70, meaning this was a very difficult achievement. Additionally, the name and earnings columns were removed because they were characters, and I didn’t want collinearity from the earnings category. An SVM model was created using 75% of this new dataset, and all of the remaining variables. The method used was svmRadial, and a confusion matrix with the predictions that came from the testset was created to validate it. The model had an accuracy of 96.65% with a 95% confidence interval between 94.44-98.16. Compared to the no information rate of 91.63% it can be concluded that this model performs better than having no model. The p-value of this model is less than 0.05 meaning this is a statistically significant model. The confusion matrix can be seen below.

Table

Description automatically generated

Next, I wanted to create a recursive partitioning model to explain which variables had the largest individual affect on average score. Theoretically, this would give players a guide as to what the most important parts of the game are. Again, the output would determine if the player would on average score above (0) or below (1) 70. All variables aside from golfer name and earnings were included in this model. The recursive partition tree can be seen below.

Diagram

Description automatically generated

Additionally a varImp() function was run on this model to determine which variables were the most important in determining a players average score. The variables can be seen below.

Text, letter

Description automatically generated

Interestingly the single most important variable was the strokes gained off the tees. Contrary to what many would say fundamentally, the drive seems to have the largest single impact on how well a player will score. It should be mentioned that the remaining 5 variables all have to do with the short game, and more importantly gaining strokes around and on the green. This means that if you perform better around and on the green (short game) than your peers, you’re likely to have a lower score. However, this model suggests that if you can only fix a single aspect of your game, it would be to focus on getting an advantage off the tees.

Finally, I wanted to see how these players improve year to year. If there is a larger focus on saving par for example, we would except to see an improvement in the scrambling stats. To look at the changes over the years, a new dataframe was created. The average statistics for every variable, for each individual year was calculated and placed into a new dataframe. In this way, there was a dataframe with 8 rows, 1 for each year, that would have the average stats for that year. It was my intention to look at this average stats year to year and see how the players are improving. I created Scatterplots of year vs. Fairway percentage, year vs. greens in regulation, year vs. average scrambling, and year vs. score. Unfortunately, there were no obvious trends, aside from the possibility that players are getting less accurate drives with time. The plots can be seen below.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Unfortunately there weren’t any clear trends, one hypothesis as to what could cause this could be what’s called “Tiger-proofing”. In the early 2000’s Tiger Woods dominated golf courses specifically from the tees. After a while, courses caught on and had to start changing aspects of their holes. In this way they could create a course that remained difficult for all players, even the best ones. It’s possible that because these courses change a bit year to year, the clubs constantly change their course to meet the new level of skill from the previous years. Again, this is all speculation and should be taken with a grain of salt.

To get a more complete grasp on these trends, the same graphs were made, however this time they only contained the top 25 players from each year. Perhaps, the average player isn’t improving year to year but the best players are. Again, there doesn’t seem to be any sort of a trend, the best players are just better at scrambling and have lower average scores. These graphs can be seen below.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

I then wanted to compare the top 25 players from each year to the overall PGA tour. In this way, I could use ANOVA to figure out if any of the variables between the top players and the rest of the players were signficantly different.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

To complete this ANOVA, I used a new dataframe that contained the top 100 players from the entire dataset. I compared the mean of their stats to the stats of all players. This was done by using the quantile() function. Interestingly, no stats aside from Average score, wins, and top 10 finishes were signficantly different at a 95% confidence interval. This may suggest that it’s necessary to be an all around player to compete at the very top level and that no single statistic can explain why the best players are the best.

**CONCLUSION**

In conclusion a lot was learned by investigating this PGA data. According to the recursive paritioning model, the single most important variable for determining if you will have a low average score, is strokes gained off the tees. It’s possible that this is because driving has the largest potential for adding strokes to your game. You could hit it into the hazard (+1 stroke) or out of bounds (+2 strokes). In this way, the players who are able to gain strokes off the tees, are likely the ones who are keeping themselves out of hazard danger. Again, this is speculation and would require further investigation to conclude confidently. It was also shown that the next 5 most important variables have to do with the game on and around the green. So if you want to be the best in the world, its first and foremost important to limit your errors on the teebox, however from there you still need a very solid short game. The ANOVA suggests that no single statistic differentiates the very best players in the world from the rest of the PGA tour. This could suggest that it’s simply necessary to be an all-around good player.

There are many ways that amateur players can improve their game to become a professional. Like what was just mentioned, the strokes gained off the tees is exceptionally important. However, some things can be taken less seriously. Hitting the fairway on your drive seems to have no significant relationship to how low your score is, at the top level. Additionally focusing on your approach shots and scrambling as the second most important part of your game will be important. Very surprisingly, putting didn’t have that large of an effect on determining average strokes according to the SVM model. It seems that setting yourself up for an easy putt is more important than putting itself, according to the data.

Finally, this project aimed to answer what the professionals are doing year over year to constantly improve at the game. Surprisingly there didn’t seem to be any measurable metric that professionals improved upon from year to year. It’s possible that this dataset didn’t have a large enough date range. It’s also possible that course are being “Tiger-proofed” year over year to constantly supply challenging courses. Whatever the case may be, according to this data, it seems professionals aren’t getting better year to year.

**RAW R CODE**

#Final Project IST 687- Joseph Davis

#DATA is on the pga tour from 2010-2018, and is provided by kaggle user: Jong

#url : https://www.kaggle.com/datasets/jmpark746/pga-tour-data-2010-2018?resource=download

library(tidyverse)

library(readr)

library(caret)

library(rpart)

library(rpart.plot)

library(kernlab)

library(e1071)

pgaTourData20102018 <- read\_csv("~/Desktop/Syracuse/Semester 1/Introduction to Data Science/Final Project/pgaTourData20102018.csv")

PGA <- pgaTourData20102018

#Learning a bit about the data

which.max(PGA$Wins)

PGA[998,]

str(PGA)

View(PGA)

#Begin munging

PGA <- PGA[1:1678,]

nrow(PGA)

PGA$Wins[is.na(PGA$Wins)]=0

View(PGA)

PGA$`Top 10`[is.na(PGA$`Top 10`)]=0

#Begin Creating New Data frames

PGA2018 <- PGA[PGA$Year==2018,]

View(PGA2018)

PGA2017 <- PGA[PGA$Year==2017,]

View(PGA2017)

PGA2016 <- PGA[PGA$Year==2016,]

View(PGA2016)

PGA2015 <- PGA[PGA$Year==2015,]

View(PGA2015)

PGA2014 <- PGA[PGA$Year==2014,]

View(PGA2014)

PGA2013 <- PGA[PGA$Year==2013,]

View(PGA2013)

PGA2012 <- PGA[PGA$Year==2012,]

View(PGA2012)

PGA2011 <- PGA[PGA$Year==2011,]

View(PGA2011)

PGA2010 <- PGA[PGA$Year==2010,]

View(PGA2010)

#Descriptive Statistics

#I want to take a look at how much people made in the year 2018

ggplot(PGA2018, aes(x=(parse\_number(Money)))) + geom\_histogram(bins=12, color="black", fill="green")+ggtitle("2018 Earnings")+ylab("Count")+xlab("Yearly Earnings ($)")

#lets make a boxplot describing average score

ggplot(PGA2018, aes(y=PGA2018$`Average Score`)) + geom\_boxplot() + ggtitle("2018 Average Score Boxplot") + xlab("Average Score")

#Lets make a scatterplot of Money vs. average score

ggplot(PGA2018, aes(x=(parse\_number(PGA2018$Money)), y=PGA2018$`Average Score`)) + geom\_point() + ylab("Average Score") + xlab("2018 Earnings") + ggtitle("2018 Average Score vs. Earnings")

#Lets make a scatterplot of Wins finishes vs. average score

ggplot(PGA2018, aes(x=`Wins`, y=`Average Score`))+ geom\_point() + ylab("Average Score") + xlab("2018 Wins") + ggtitle("2018 Average Score vs. Wins")

#Lets make a scatterplot of Top10 finishes vs. average score

ggplot(PGA2018, aes(x=`Top 10`, y=`Average Score`))+ geom\_point() + ylab("Average Score") + xlab("2018 Top 10 Finishes") + ggtitle("2018 Average Score vs. Top 10 Finishes")

#Making a new dataframe containing the best 25 players from each season

Best2018 <- PGA2018[order(PGA2018$`Average Score`),]

Top25From2018 <- Best2018[1:25,]

Best2017 <- PGA2017[order(PGA2017$`Average Score`),]

Top25From2017 <- Best2017[1:25,]

Best2016 <- PGA2016[order(PGA2016$`Average Score`),]

Top25From2016 <- Best2016[1:25,]

Best2015 <- PGA2015[order(PGA2015$`Average Score`),]

Top25From2015 <- Best2015[1:25,]

Best2014 <- PGA2014[order(PGA2014$`Average Score`),]

Top25From2014 <- Best2014[1:25,]

Best2013 <- PGA2013[order(PGA2013$`Average Score`),]

Top25From2013 <- Best2013[1:25,]

Best2012 <- PGA2012[order(PGA2012$`Average Score`),]

Top25From2012 <- Best2012[1:25,]

Best2011 <- PGA2011[order(PGA2011$`Average Score`),]

Top25From2011 <- Best2011[1:25,]

Best2010 <- PGA2010[order(PGA2010$`Average Score`),]

Top25From2010 <- Best2010[1:25,]

View(Top25From2010)

TopAll <- rbind(Top25From2018, Top25From2017, Top25From2016, Top25From2015, Top25From2014, Top25From2013, Top25From2012, Top25From2011, Top25From2010)

View(TopAll)

TopAll <- TopAll[order(TopAll$`Average Score`),]

head(TopAll, 10)

#Linear Model Creation

#Simple scatterplot to show the relationship for the top players

g <- ggplot(TopAll, aes(y=`Fairway Percentage`, x=`Avg Distance`)) +geom\_point(aes(color=`Average Score`))

g

g + geom\_smooth(method = "lm") + ggtitle("Driving Accuracy vs. Distance") + coord\_fixed()

#Linear Model For Driving Accuracy From the 2018 Season

Driving.lm = lm(formula=`Fairway Percentage` ~ `Avg Distance`+`SG:OTT`, data=PGA2018)

summary(Driving.lm)

#How about a linear model that describes average strokes

Strokes.lm = lm(formula=`Average Score` ~ `gir`+`Average Putts`+`Average Scrambling`

+`Avg Distance`+`Fairway Percentage`, data=PGA)

summary(Strokes.lm)

#SVM and Recursive Partition Tree

#Create a dataframe with all players and years but an additional column

mean(PGA$`Average Score`)

#New column:If they shot on average below a 70

PGAsub70 <- PGA[PGA$`Average Score`<70,]

View(PGAsub70)

nrow(PGAsub70)

#143 rows and our original PGA is 1678, so about the top 8.5%% of players

nrow(PGA)

binaryPGAsub70 <- replace(PGAsub70,"Average Score", 1)

View(binaryPGAsub70)

PGAover70 <- PGA[PGA$`Average Score`>69.9999,]

View(PGAover70)

binaryPGAover70 <- replace(PGAover70,"Average Score", 0)

View(binaryPGAover70)

BinaryPGA <-rbind(binaryPGAsub70, binaryPGAover70)

BinaryPGA$`Average Score` <- as.factor(BinaryPGA$`Average Score`)

View(BinaryPGA)

str(BinaryPGA)

BinaryPGA <- BinaryPGA[, -18]

BinaryPGA <- BinaryPGA[, -1]

#This new BinaryPGA data can be used to make an SVM and rpart

#SVM Model creation with 75% partition of the data

trainList <- createDataPartition(y=BinaryPGA$'Average Score', p=.75, list=FALSE)

testing <- BinaryPGA[-trainList,]

training <- BinaryPGA[trainList,]

#Training the model

fit.svm <- train(`Average Score`~`gir`+`Average Putts`+`Average Scrambling`

+`Average SG Putts`+`SG:OTT`+`SG:APR`+`SG:ARG`

, data=testing, method="svmRadial",preProc=c("center","scale"))

#Testing the model

predOut <- predict(fit.svm, newdata=testing)

confusion <- confusionMatrix(predOut, testing$`Average Score`)

confusion

fit.svm

#Recursive Partition Model

#Lets make a decision tree

rpart.plot(rpart(BinaryPGA$'Average Score'~BinaryPGA$Rounds+BinaryPGA$'Fairway Percentage'+BinaryPGA$`Avg Distance`

+BinaryPGA$gir+BinaryPGA$`Average Putts`+BinaryPGA$`Average Scrambling`+BinaryPGA$Wins

+BinaryPGA$`Top 10`+BinaryPGA$`Average SG Putts`+BinaryPGA$`SG:OTT`+BinaryPGA$`SG:APR`

+BinaryPGA$`SG:ARG`, data=BinaryPGA))

cartTree <- rpart(BinaryPGA$'Average Score'~BinaryPGA$Rounds+BinaryPGA$'Fairway Percentage'+BinaryPGA$`Avg Distance`

+BinaryPGA$gir+BinaryPGA$`Average Putts`+BinaryPGA$`Average Scrambling`+BinaryPGA$Wins

+BinaryPGA$`Top 10`+BinaryPGA$`Average SG Putts`+BinaryPGA$`SG:OTT`+BinaryPGA$`SG:APR`

+BinaryPGA$`SG:ARG`, data=BinaryPGA)

cartTree

t <- varImp(cartTree)

t%>% arrange(desc(Overall))%>% slice(1:6)

#I'm curious if we take out wins/top10 how it looks

#####This is our winner#####

rpart.plot(rpart(BinaryPGA$'Average Score'~BinaryPGA$Rounds+BinaryPGA$'Fairway Percentage'+BinaryPGA$`Avg Distance`

+BinaryPGA$gir+BinaryPGA$`Average Putts`+BinaryPGA$`Average Scrambling`

+BinaryPGA$`Average SG Putts`+BinaryPGA$`SG:OTT`+BinaryPGA$`SG:APR`

+BinaryPGA$`SG:ARG`, data=BinaryPGA))

cartTree <- rpart(BinaryPGA$'Average Score'~BinaryPGA$Rounds+BinaryPGA$'Fairway Percentage'+BinaryPGA$`Avg Distance`

+BinaryPGA$gir+BinaryPGA$`Average Putts`+BinaryPGA$`Average Scrambling`

+BinaryPGA$`Average SG Putts`+BinaryPGA$`SG:OTT`+BinaryPGA$`SG:APR`

+BinaryPGA$`SG:ARG`, data=BinaryPGA)

cartTree

t <- varImp(cartTree)

t%>% arrange(desc(Overall))%>% slice(1:6)

######This is our winner######

#I want a decision tree with just the top 6 biggest predictors

#This made no difference

rpart.plot(rpart(BinaryPGA$`Average Score`~BinaryPGA$`SG:OTT`+BinaryPGA$`Average Scrambling`

+BinaryPGA$`SG:APR`+BinaryPGA$`SG:ARG`+BinaryPGA$`Average SG Putts`+BinaryPGA$`Average Putts`, data=BinaryPGA))

cartTree <- rpart(BinaryPGA$`Average Score`~BinaryPGA$`SG:OTT`+BinaryPGA$`Average Scrambling`

+BinaryPGA$`SG:APR`+BinaryPGA$`SG:ARG`+BinaryPGA$`Average SG Putts`+BinaryPGA$`Average Putts`, data=BinaryPGA)

cartTree

varImp(cartTree)

#Finally I want to look at professional improvement over time

#What stats are improving year over year?

Changes2010 <- c(mean((PGA2010$`Fairway Percentage`)),mean((PGA2010$`Avg Distance`)),

mean(PGA2010$Year), mean(PGA2010$gir), mean(PGA2010$`Average Putts`),

mean(PGA2010$`Average Scrambling`), mean(PGA2010$`Average Score`))

Changes2011 <- c(mean((PGA2011$`Fairway Percentage`)),mean((PGA2011$`Avg Distance`)),

mean(PGA2011$Year), mean(PGA2011$gir), mean(PGA2011$`Average Putts`),

mean(PGA2011$`Average Scrambling`), mean(PGA2011$`Average Score`))

Changes2012 <- c(mean((PGA2012$`Fairway Percentage`)),mean((PGA2012$`Avg Distance`)),

mean(PGA2012$Year), mean(PGA2012$gir), mean(PGA2012$`Average Putts`),

mean(PGA2012$`Average Scrambling`), mean(PGA2012$`Average Score`))

Changes2013 <- c(mean((PGA2013$`Fairway Percentage`)),mean((PGA2013$`Avg Distance`)),

mean(PGA2013$Year), mean(PGA2013$gir), mean(PGA2013$`Average Putts`),

mean(PGA2013$`Average Scrambling`), mean(PGA2013$`Average Score`))

Changes2014 <- c(mean((PGA2014$`Fairway Percentage`)),mean((PGA2014$`Avg Distance`)),

mean(PGA2014$Year), mean(PGA2014$gir), mean(PGA2014$`Average Putts`),

mean(PGA2014$`Average Scrambling`), mean(PGA2014$`Average Score`))

Changes2015 <- c(mean((PGA2015$`Fairway Percentage`)),mean((PGA2015$`Avg Distance`)),

mean(PGA2015$Year), mean(PGA2015$gir), mean(PGA2015$`Average Putts`),

mean(PGA2015$`Average Scrambling`), mean(PGA2015$`Average Score`))

Changes2016 <- c(mean((PGA2016$`Fairway Percentage`)),mean((PGA2016$`Avg Distance`)),

mean(PGA2016$Year), mean(PGA2016$gir), mean(PGA2016$`Average Putts`),

mean(PGA2016$`Average Scrambling`), mean(PGA2016$`Average Score`))

Changes2017 <- c(mean((PGA2017$`Fairway Percentage`)),mean((PGA2017$`Avg Distance`)),

mean(PGA2017$Year), mean(PGA2017$gir), mean(PGA2017$`Average Putts`),

mean(PGA2017$`Average Scrambling`), mean(PGA2017$`Average Score`))

Changes2018 <- c(mean((PGA2018$`Fairway Percentage`)),mean((PGA2018$`Avg Distance`)),

mean(PGA2018$Year), mean(PGA2010$gir), mean(PGA2018$`Average Putts`),

mean(PGA2018$`Average Scrambling`), mean(PGA2018$`Average Score`))

Changes <- data.frame(Changes2010, Changes2011, Changes2012, Changes2013,

Changes2014, Changes2015, Changes2016, Changes2017, Changes2018)

View(Changes)

#Need to flip columns and rows

changesdf <- data.frame(t(Changes))

changesdf

colnames(changesdf) <- c("Fairway Percentage", "Avg Distance", "Year", "gir",

"Average Putts", "Average Scrambling", "Average Score")

#Looking at trends for Year vs. Fairway Percentage, greens in regualtion, and Scrambling

g <- ggplot(changesdf, aes(x=`Year`)) +geom\_point(aes(y=`Fairway Percentage`, color="Fairway Percentage"))+geom\_point(aes(y=`gir`, color="Green in Regulation"))+geom\_point(aes(y=`Average Scrambling`, color="Average Scrambling"))+ylab("Percentage")

g + ggtitle("Player Performance over Time")

#Don't seem to be any interesting trends but could still include this

ggplot(changesdf, aes(x=`Year`)) + geom\_point(aes(y=`Average Score`, color="Average Putts")) + ylim(70,72) + ggtitle("Player Scores over Time")

#Curious to look at the top 25 players from each year

Changes2010 <- c(mean((Top25From2010$`Fairway Percentage`)),mean((Top25From2010$`Avg Distance`)),

mean(Top25From2010$Year), mean(Top25From2010$gir), mean(Top25From2010$`Average Putts`),

mean(Top25From2010$`Average Scrambling`), mean(Top25From2010$`Average Score`))

Changes2011 <- c(mean((Top25From2011$`Fairway Percentage`)),mean((Top25From2011$`Avg Distance`)),

mean(Top25From2011$Year), mean(Top25From2011$gir), mean(Top25From2011$`Average Putts`),

mean(Top25From2011$`Average Scrambling`), mean(Top25From2011$`Average Score`))

Changes2012 <- c(mean((Top25From2012$`Fairway Percentage`)),mean((Top25From2012$`Avg Distance`)),

mean(Top25From2012$Year), mean(Top25From2012$gir), mean(Top25From2012$`Average Putts`),

mean(Top25From2012$`Average Scrambling`), mean(Top25From2012$`Average Score`))

Changes2013 <- c(mean((Top25From2013$`Fairway Percentage`)),mean((Top25From2013$`Avg Distance`)),

mean(Top25From2013$Year), mean(Top25From2013$gir), mean(Top25From2013$`Average Putts`),

mean(Top25From2013$`Average Scrambling`), mean(Top25From2013$`Average Score`))

Changes2014 <- c(mean((Top25From2014$`Fairway Percentage`)),mean((Top25From2014$`Avg Distance`)),

mean(Top25From2014$Year), mean(Top25From2014$gir), mean(Top25From2014$`Average Putts`),

mean(Top25From2014$`Average Scrambling`), mean(Top25From2014$`Average Score`))

Changes2015 <- c(mean((Top25From2015$`Fairway Percentage`)),mean((Top25From2015$`Avg Distance`)),

mean(Top25From2015$Year), mean(Top25From2015$gir), mean(Top25From2015$`Average Putts`),

mean(Top25From2015$`Average Scrambling`), mean(Top25From2015$`Average Score`))

Changes2016 <- c(mean((Top25From2016$`Fairway Percentage`)),mean((Top25From2016$`Avg Distance`)),

mean(PGA2016$Year), mean(Top25From2016$gir), mean(Top25From2016$`Average Putts`),

mean(Top25From2016$`Average Scrambling`), mean(Top25From2016$`Average Score`))

Changes2017 <- c(mean((Top25From2017$`Fairway Percentage`)),mean((Top25From2017$`Avg Distance`)),

mean(Top25From2017$Year), mean(Top25From2017$gir), mean(Top25From2017$`Average Putts`),

mean(Top25From2017$`Average Scrambling`), mean(Top25From2017$`Average Score`))

Changes2018 <- c(mean((Top25From2018$`Fairway Percentage`)),mean((Top25From2018$`Avg Distance`)),

mean(Top25From2018$Year), mean(Top25From2010$gir), mean(Top25From2018$`Average Putts`),

mean(Top25From2018$`Average Scrambling`), mean(Top25From2018$`Average Score`))

Changes1 <- data.frame(Changes2010, Changes2011, Changes2012, Changes2013,

Changes2014, Changes2015, Changes2016, Changes2017, Changes2018)

changesdf1 <- data.frame(t(Changes1))

changesdf1

colnames(changesdf1) <- c("Fairway Percentage", "Avg Distance", "Year", "gir",

"Average Putts", "Average Scrambling", "Average Score")

#Looking at trends for Year vs. Fairway Percentage, greens in regualtion, and Scrambling

g2 <- ggplot(changesdf1, aes(x=`Year`)) +geom\_point(aes(y=`Fairway Percentage`, color="Fairway Percentage"))+geom\_point(aes(y=`gir`, color="Green in Regulation"))+geom\_point(aes(y=`Average Scrambling`, color="Average Scrambling"))+ylab("Percentage")

g2 + ggtitle("Top Players Performance over Time")

ggplot(changesdf1, aes(x=`Year`)) + geom\_point(aes(y=`Average Score`, color="Average Putts")) + ylim(68,72) + ggtitle("Top Players Scores over Time")

#Lets compare these few stats, seems to be a pattern

ggplot(changesdf, aes(x=`Year`)) +geom\_point(data=changesdf, aes(y=`Fairway Percentage`, color="Fairway Percentage"))+geom\_point(data=changesdf1, aes(y=`Fairway Percentage`, color="Top Players Fairway Percentage"))+ggtitle("Fairway Percentages")

ggplot(changesdf, aes(x=`Year`)) +geom\_point(data=changesdf,aes(y=`gir`, color="Green in Regulation"))+geom\_point(data=changesdf1,aes(y=`gir`, color="Top Players Green in Regulation"))+ggtitle("Green in Regulation")

ggplot(changesdf, aes(x=`Year`)) +geom\_point(data=changesdf, aes(y=`Average Scrambling`, color="Average Scrambling"))+geom\_point(data=changesdf1,aes(y=`Average Scrambling`, color="Top Players Average Scrambling"))+ggtitle("Scrambling")

ggplot(changesdf, aes(x=`Year`)) +geom\_point(data=changesdf, aes(y=`Average Score`, color="Average Score"))+geom\_point(data=changesdf1,aes(y=`Average Score`, color="Top Players Average Score"))+ggtitle("Average Score")

#Is the difference in means of stats significant?

#Make a new dataset with the best 100 players

PGAorder <- PGA[order(PGA$`Average Score`),]

PGATop100 <- PGAorder[1:100,]

View(PGATop100)

quantile(PGA$`Fairway Percentage`, probs = c(.05, .95))

mean(PGATop100$`Fairway Percentage`)

quantile(PGA$`Avg Distance`, probs = c(.05, .95))

mean(PGATop100$`Avg Distance`)

quantile(PGA$`gir`, probs = c(.05, .95))

mean(PGATop100$`gir`)

quantile(PGA$`Average Putts`, probs = c(.05, .95))

mean(PGATop100$`Average Putts`)

quantile(PGA$`Average Scrambling`, probs = c(.05, .95))

mean(PGATop100$`Average Scrambling`)

#These 2 were signficantly different

quantile(PGA$`Wins`, probs = c(.05, .95))

mean(PGATop100$`Wins`)

quantile(PGA$`Top 10`, probs = c(.05, .95))

mean(PGATop100$`Top 10`)

#

#Lets try at just 90% CI

quantile(PGA$`Fairway Percentage`, probs = c(.1, .9))

mean(PGATop100$`Fairway Percentage`)

quantile(PGA$`Avg Distance`, probs = c(.1, .9))

mean(PGATop100$`Avg Distance`)

quantile(PGA$`gir`, probs = c(.1, .9))

mean(PGATop100$`gir`)

quantile(PGA$`Average Putts`, probs = c(.1, .9))

mean(PGATop100$`Average Putts`)

quantile(PGA$`Average Scrambling`, probs = c(.1, .9))

mean(PGATop100$`Average Scrambling`)