Report: Predicting Scoring, Play Outcomes, and Yards

Introduction

College football is one of the United States' favorite sports. Most schools are members of conferences and spend most of the season competing against other schools in their conference. The Power 5 Conferences, the Southeastern Conference (SEC), Big Ten Conference, Atlantic Coastal Conference (ACC), Pacific-12 Conference (Pac-12), and Big 12 Conference, dominate Division I college football.

The purpose of this project is to examine a few questions important to football: 1) which factors predict if a play will result in a score, 2) which factors predict whether the outcome of a play will be a gain of yards, a loss of yards, or no yards, and 3) which factors predict the number of yards gained on a play?

Data

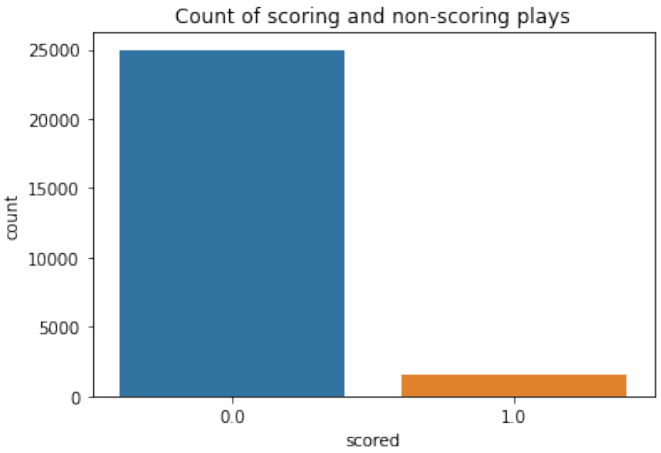
The main dataset for this project was play-by-play data from the 2019 college football season. I downloaded the dataset from SportsDataStuff.com, who scraped the data from internet sites such as ESPN. The data included every play from every game in the 2019 season. I combined the play-by-play data with a dataset of weekly top 25 rankings. From these rankings, I created 6 groups to use in analysis: top 5, 6-10, 11-15, 16-20, 21-25, and unranked. I also used a dataset of Power 5 conference teams to add a conference membership column to the dataset.

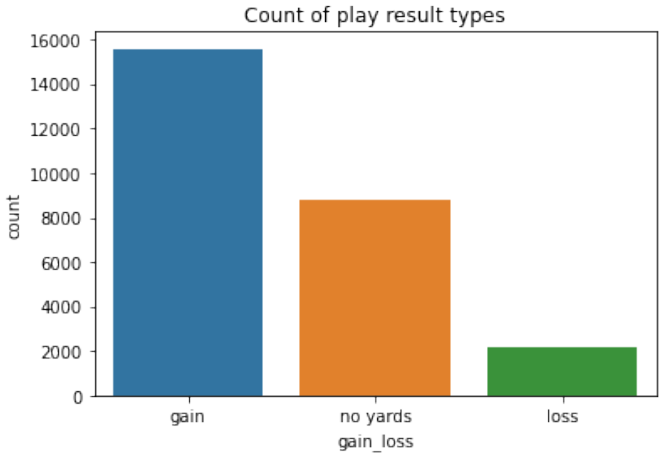
For the purposes of this project, I reduced the data to focus on Big 10 and SEC teams' offensive plays during home games. I dropped all other plays, reducing the dataset from almost 150,000 plays to about 34,000 plays.

The target variables in the dataset are named 'scored', 'gain_loss', and 'yards'.

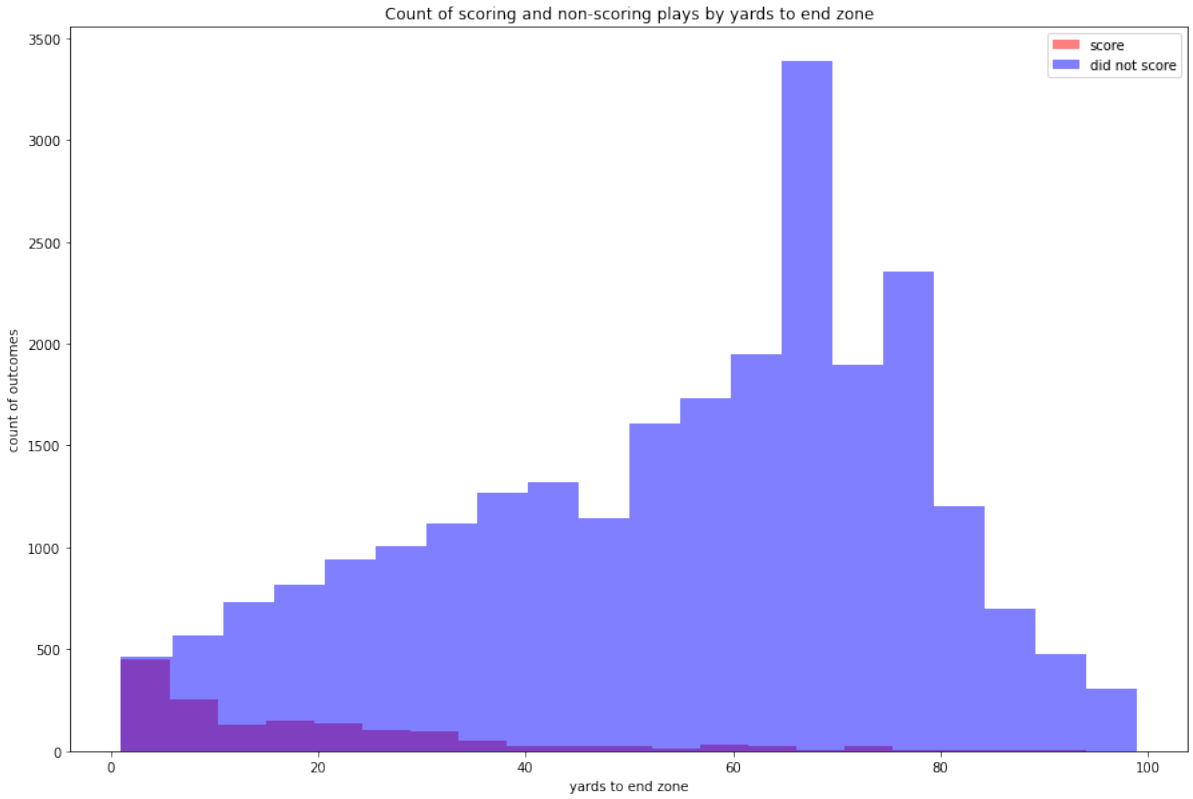
Exploratory Data Analysis

The classes are imbalanced for both the 'scored' target variable and the 'gain_loss' target variable. This imbalance is predictable, as scoring is relatively rare, but teams are usually moving forward.

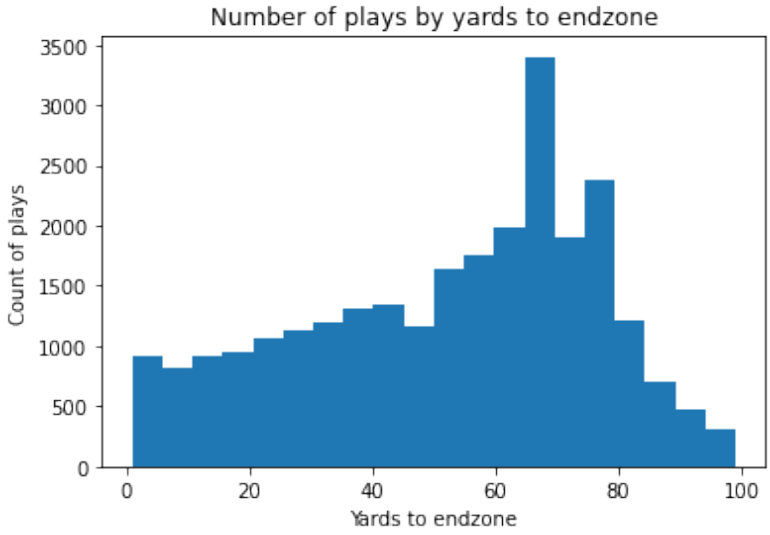




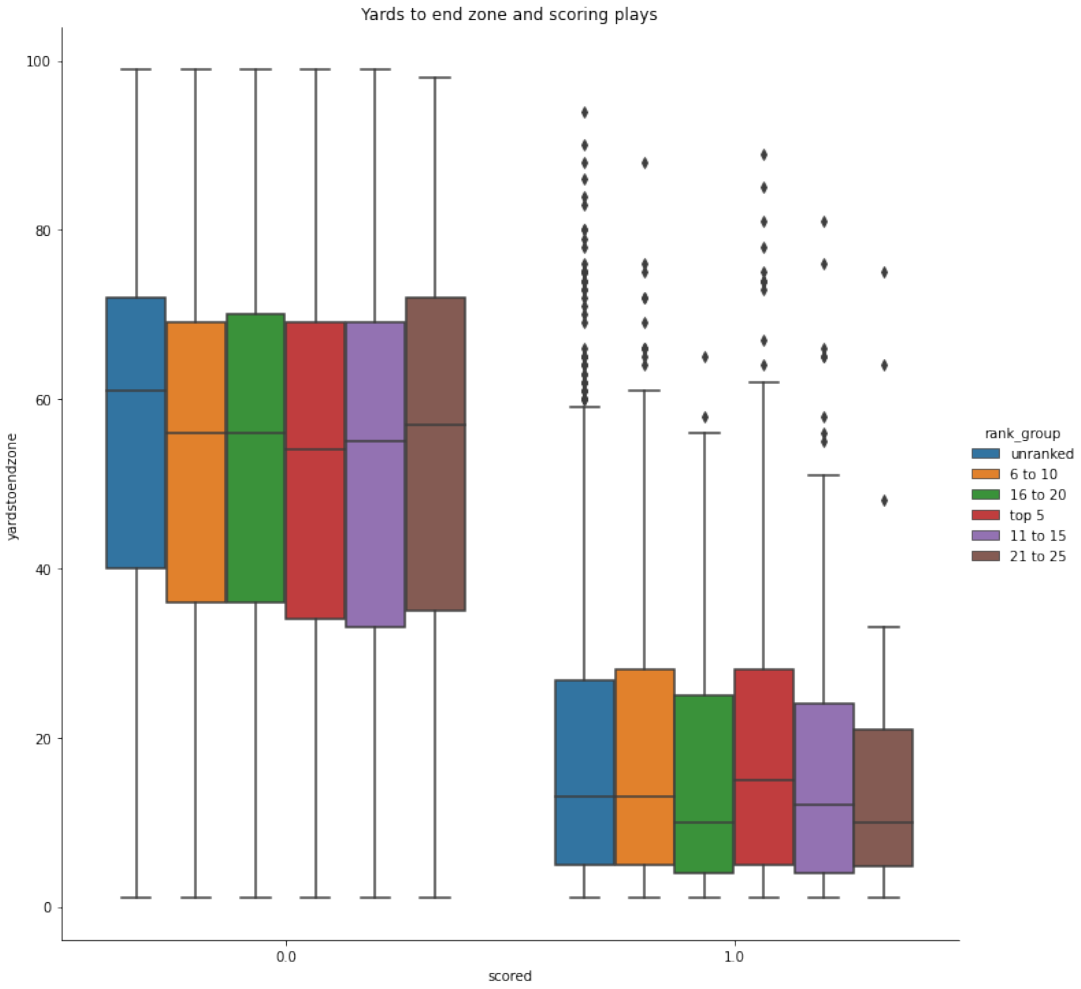
Typically, teams score from closer to the end zone than from other locations on the field. Scoring from more than 40 yards away from the end zone is rare.



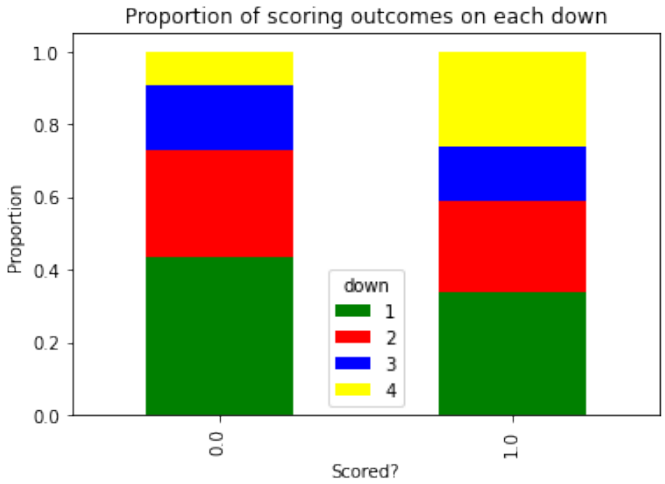
A quick look at this graph indicates a few spikes in non-scoring plays between 60 and 80 yards. The histogram of play counts and yards from the end zone shows that there are simply more plays within these yard ranges, likely because drives usually begin in this area after a kickoff or a punt. The distribution of yards to end zone on non-scoring plays largely mirrors the distribution of yards to end zone overall.



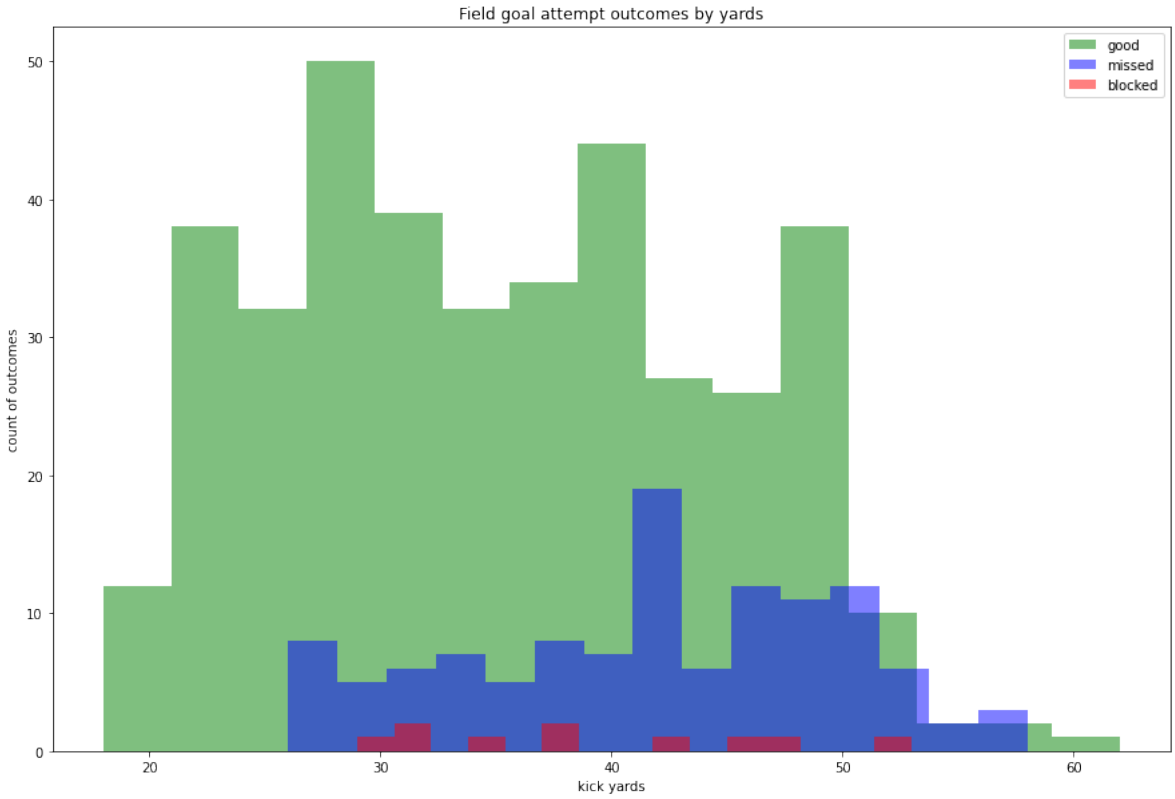
Top 5 teams had a slightly higher average yards to end zone from which they scored, and the group of 16-20 had a slightly lower average, even below that of unranked teams.



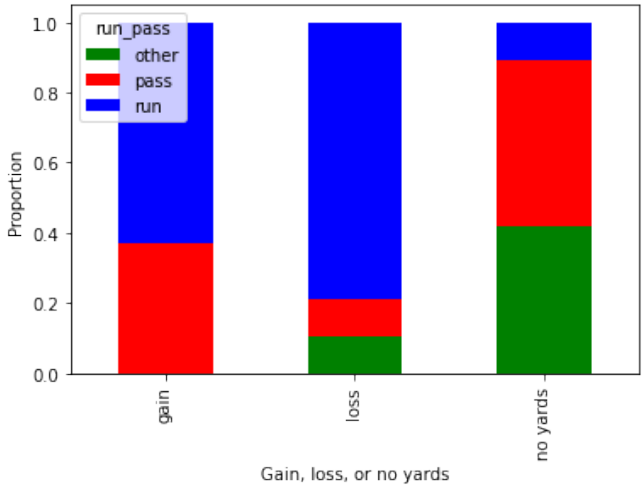
Teams scored most frequently on first down. The higher proportion of scoring plays on fourth down, compared to non-scoring plays on fourth down, is likely due to field goals.



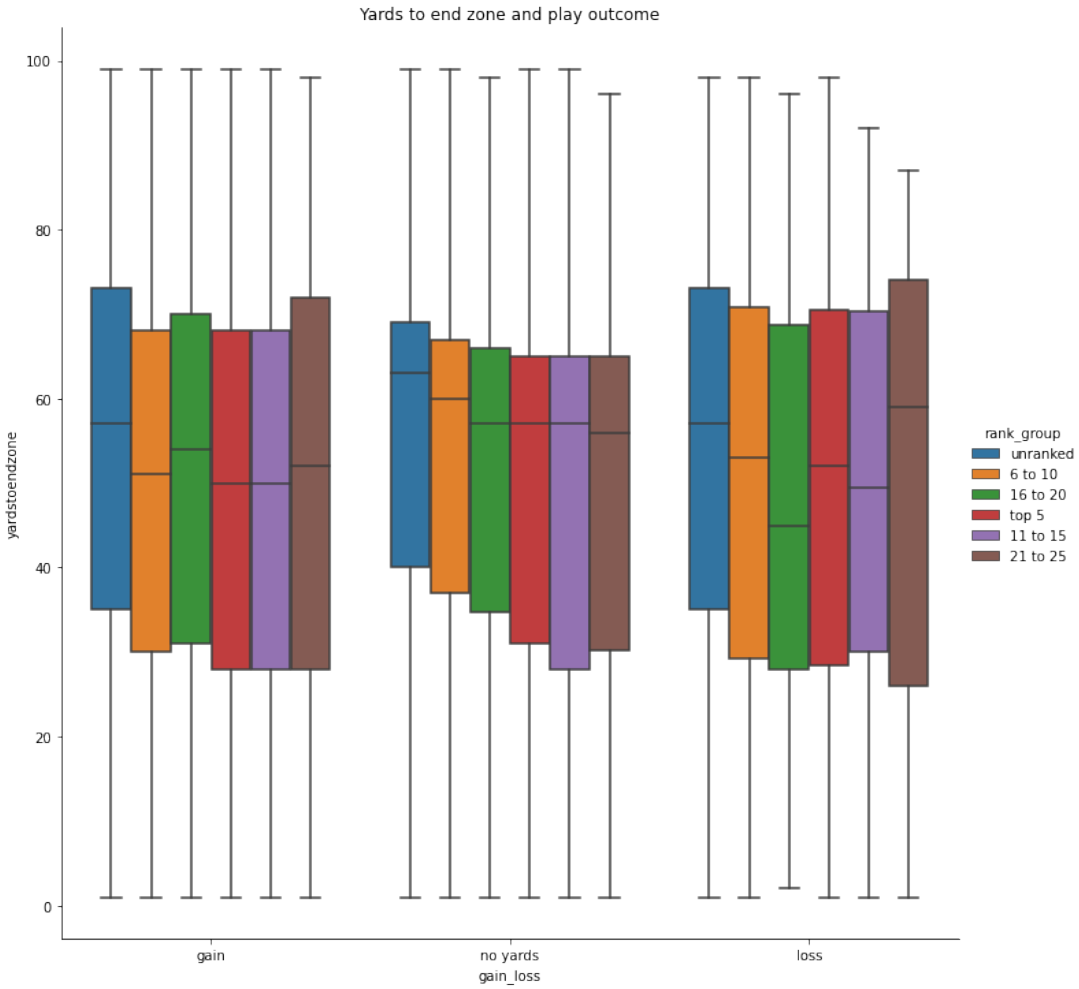
Unsurprisingly, field goal attempts were missed more often the longer the kick. Some teams have exceptionally good kickers and could successfully kick field goals of over 60 yards.



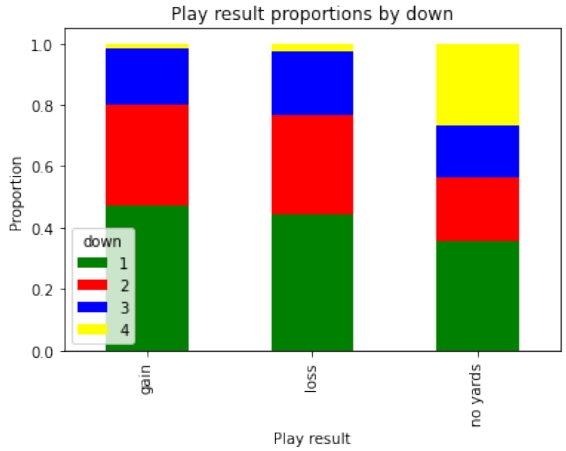
The vast majority of plays that lost yards were run plays. This discrepancy is likely due in part to the different risks associated with run and pass plays. On a pass play, a team is unlikely to lose yards because an incomplete pass will result in no yards, but they risk an interception. On a run play, the team might lose yards but they will not throw an interception. “Other” plays resulting in no yards are largely punting plays.



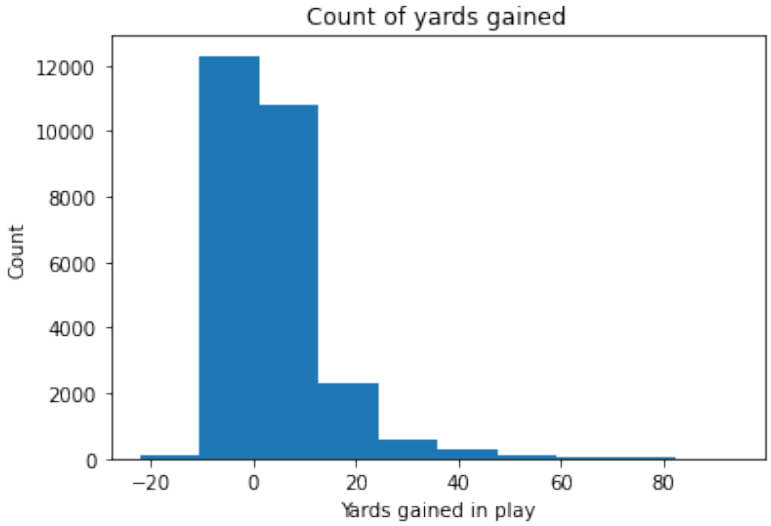
Rank group did not appear to be a big factor in whether the play resulted in a gain, loss, or no yardage, although unranked teams tended to lose yards throughout more of the field than ranked teams.



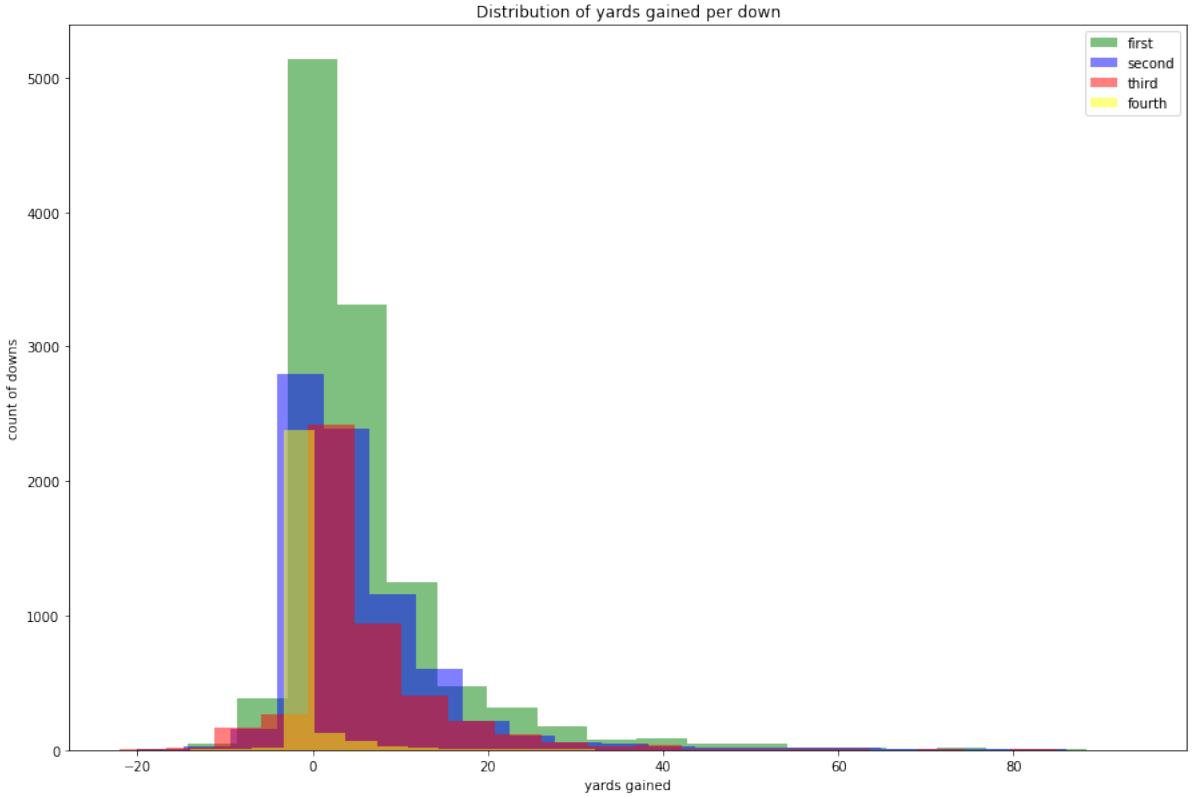
The proportion of gain and loss plays on all downs are extremely similar, nearly identical. A higher proportion of plays resulting in no yards take place on fourth down, again due to punting.



The numeric target variable, ‘yards’, is not normally distributed. It is centered around 0, with most plays resulting in between -10 and 10 yards. There are more extreme positive yardage plays than negative.



Teams had the most first downs, as a first down is required on every series of downs while second, third, and fourth can be skipped upon successful gain of 10 yards. The distribution of yards gained is similar for each down except fourth. This difference is almost certainly due to the propensity for punting the ball on fourth down to move the other team further away from the end zone.

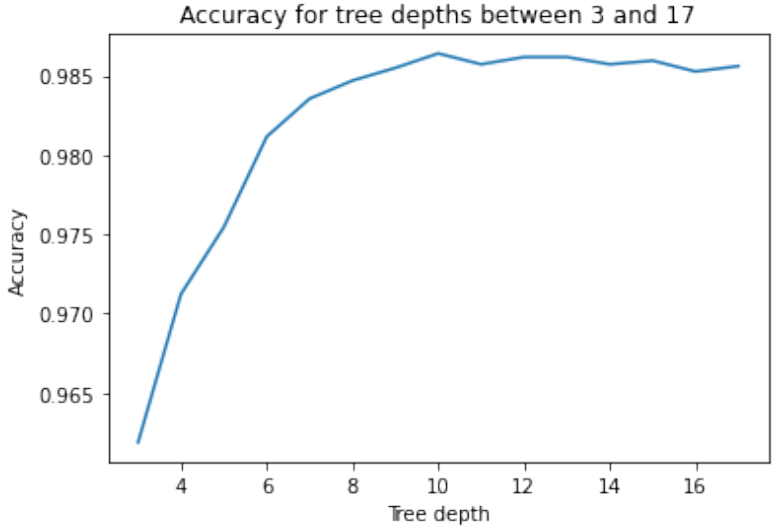


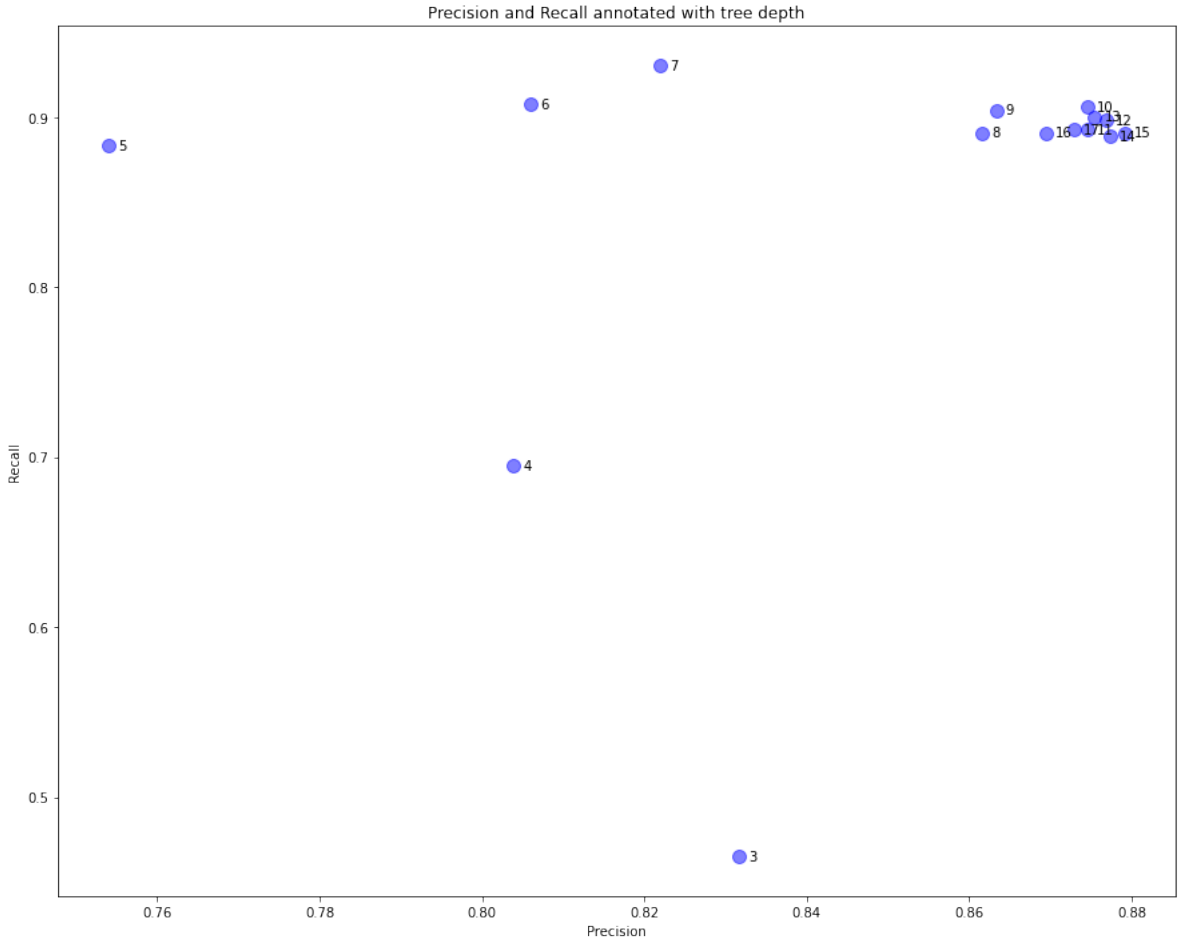
Modeling / Findings

To prepare the data for modeling, I scaled the yards variable, setting the lowest to 0 and the highest to 100 (variable name is 'yards_sc'). I then created a train/test split for each of the target variables: 'scored', 'gain_loss', and 'yards_sc'. I removed the 'yards_sc' variable from the feature set for 'gain_loss' because a play's classification as a gain or loss depends entirely on yards gained or lost during the play. I also removed 'gain_loss' from the features for predicting 'yards_sc'.

Predicting scoring plays ('scored')

To predict whether a play is a scoring play, I ran a decision tree, a random forest model, and a logistic regression. The best model for scored was a decision tree model with a depth of 10. The model's accuracy was highest at 10, and 10 also provided what I judged to be the best combination of precision and recall.



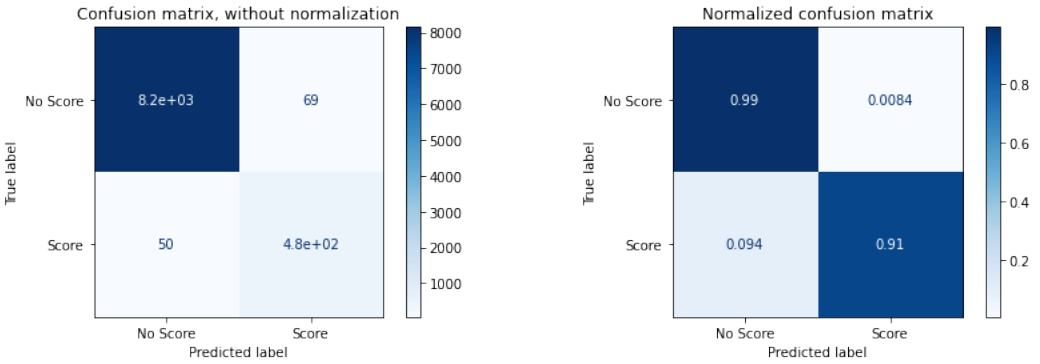


10 is not the only potential choice, as the accuracy is not vastly different for the trees ranging between 3 and 17 leaves, nor is precision and recall vastly different for trees over 9. I selected it because it had the highest accuracy and was the smallest in the cluster of trees with similar recall and precision.

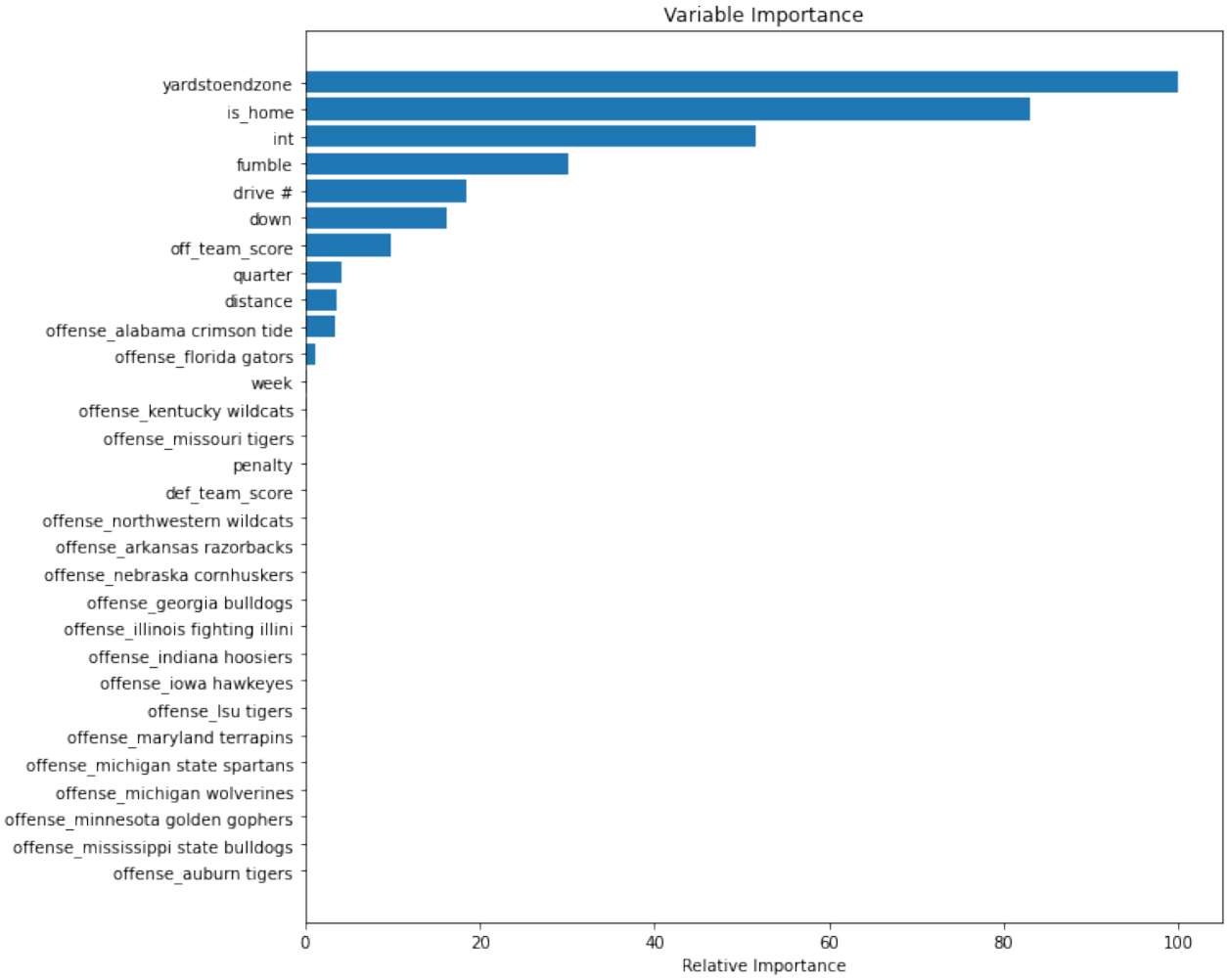
The model metrics are as follows:

Metric	**Model**
Accuracy	0.9864139741979678
Precision	0.8745454545454545
Recall	0.9058380414312618

There is not a large difference in the model's prediction of scoring plays and non-scoring plays.



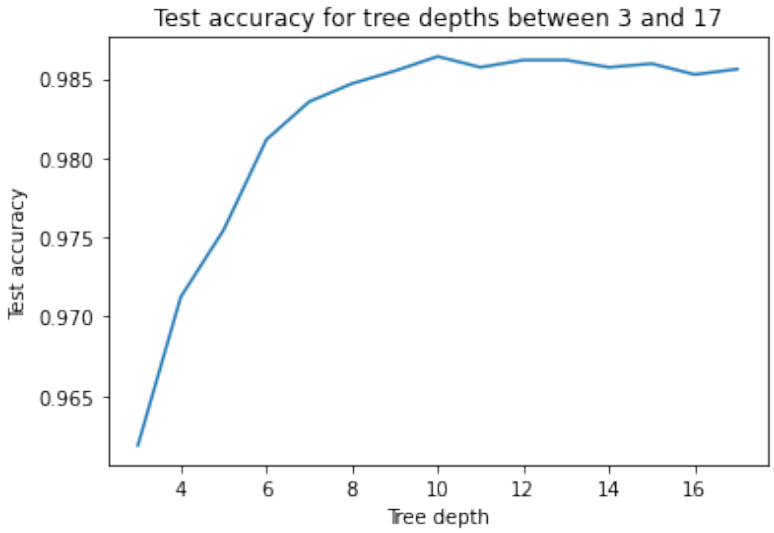
The most important features for predicting whether or not a play result in a score were yards to end zone, whether or not the game is a home game, and the presence of big mistakes – fumbles and interceptions.

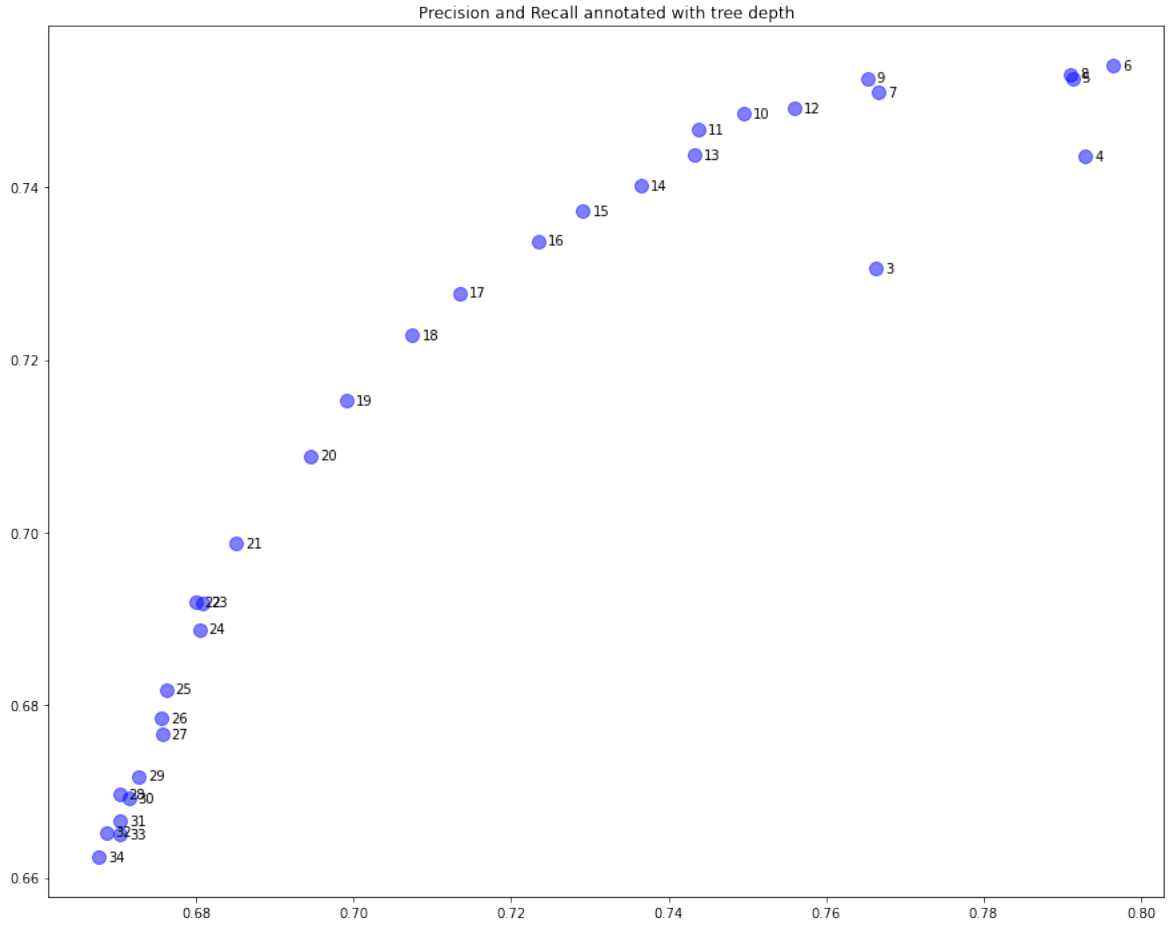


There is a chance of overfitting with a tree depth of 10. This potential overfitting could be checked with data from other conferences in the 2019 season.

Predicting play outcome – Gain, loss, no yards ('gain_loss')

As with predicting scoring, I ran a decision tree, a random forest model, and logistic regression. The best model was again the decision tree model, this time with a depth of 6. A tree with the depth of 6 had the highest accuracy, precision, and recall.

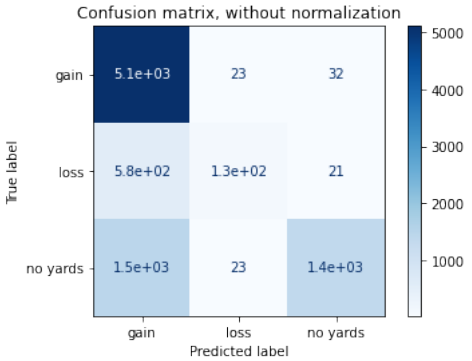


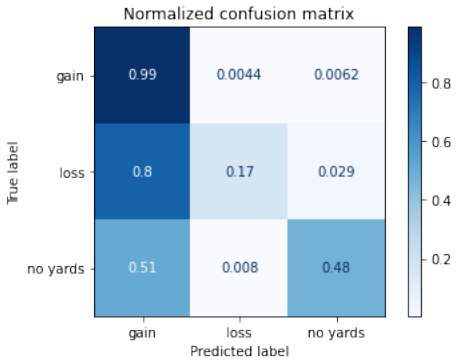


For this model, a depth of 6 was the obvious choice – it had the highest accuracy, precision, and recall.

The model's metrics are as follows:

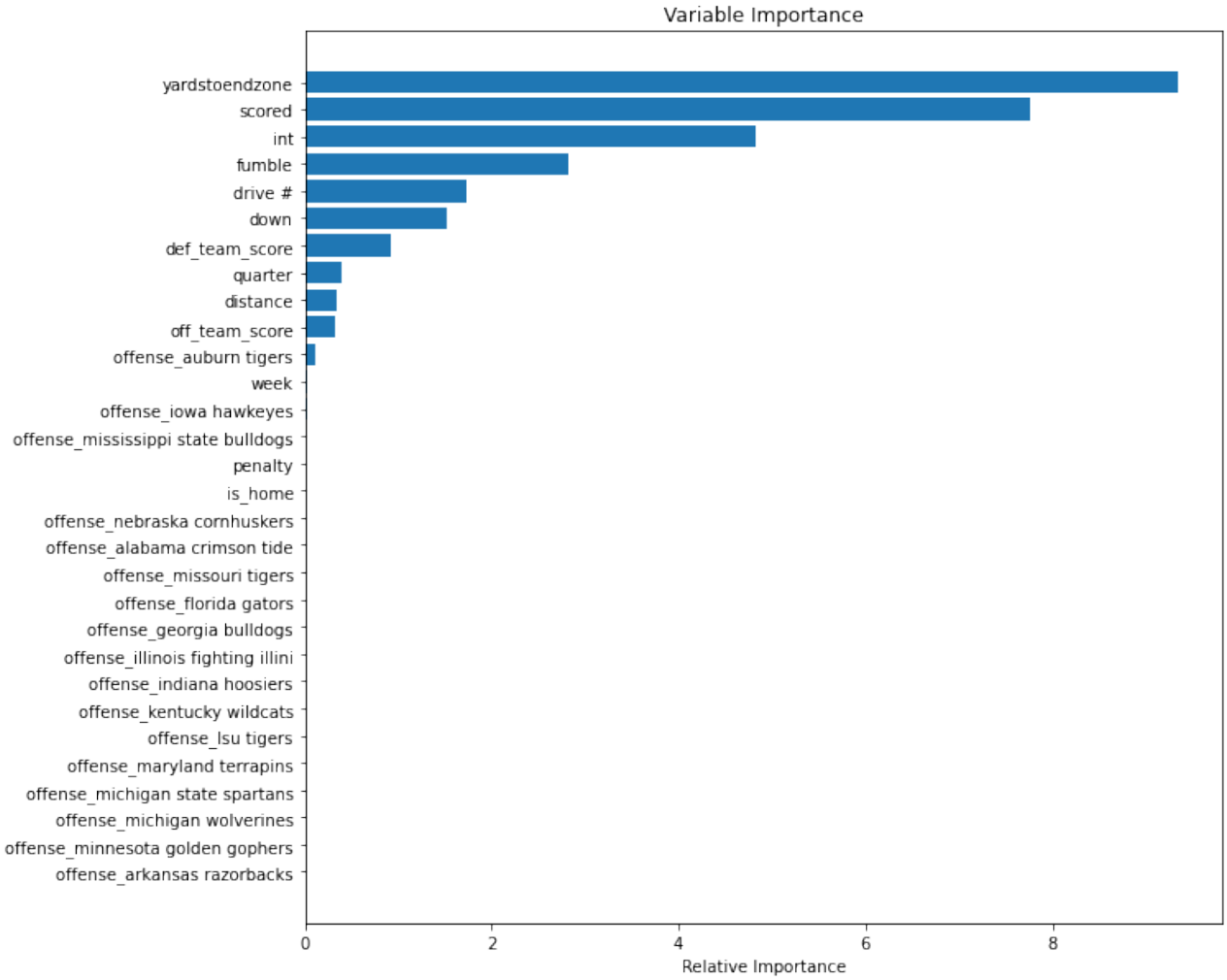
Metric	**Model**
Accuracy	0.7539673478707615
Precision	0.7961585724506773
Recall	0.7539673478707615





This model generally predicts plays with yards gained or lost correctly but does not predict plays with no yards gained or lost as well, incorrectly classifying these plays as a gain 51% of the time while classifying it correctly 48% of the time. However, this is still higher than the 33% of all plays that no yardage plays constitute.

The most important features for predicting play outcomes were yards to endzone, whether or not the team scored, the presence of interceptions and fumbles, drive number, and down.



Yards

To predict the yards gained on a play, I ran a linear regression model and a random forest regressor.

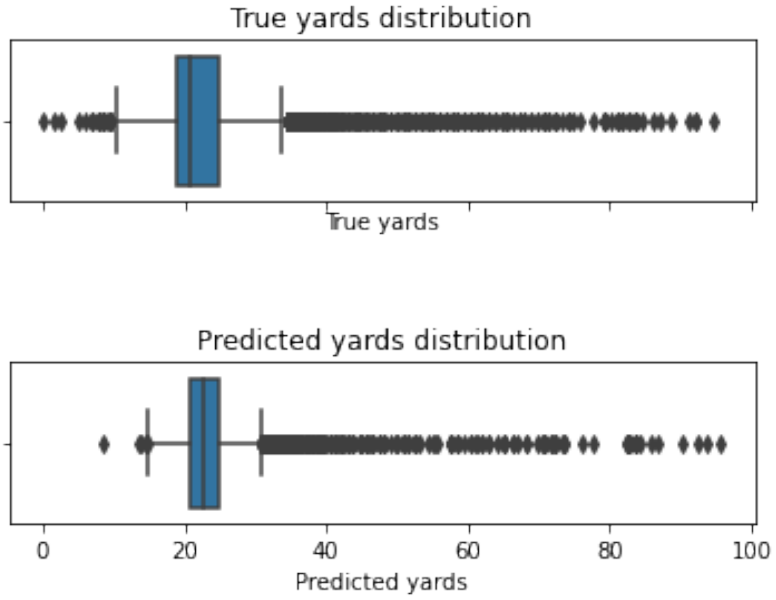
The random forest model was the best choice, although neither it nor the linear regression model were as accurate as the predictions for scoring and play outcome.

The model metrics are as follows:

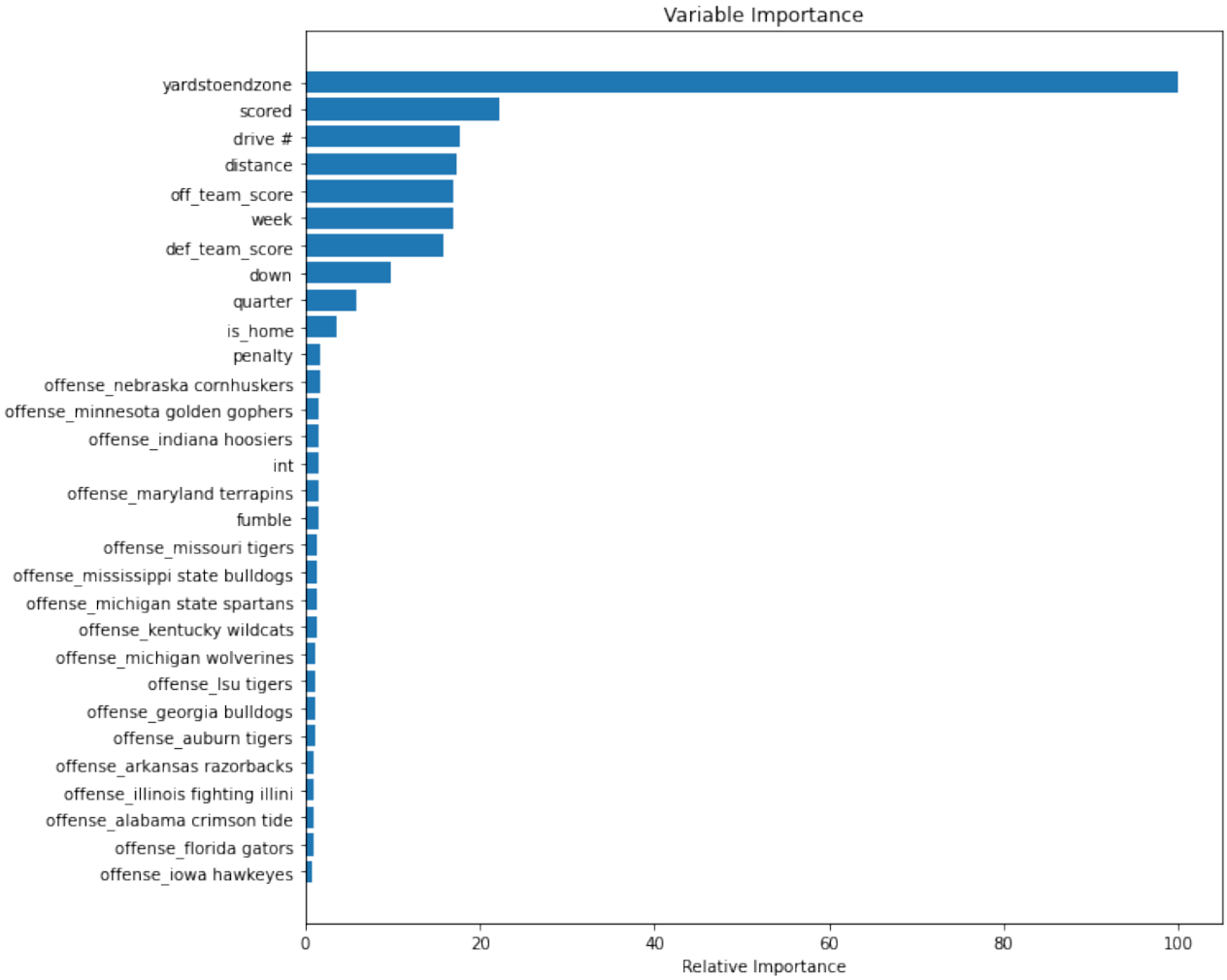
Metric	Model
RMSE	6.4724018987167735
R2	0.38100796297820283

The model's R2 is not high, but it is higher than the linear regression model by .14.

The random forest regressor's predicts a narrower yards IQR than the true distribution and also does not predict the lower yardage plays well. *(Note: the yards target variable is scaled from 0 to 100. There is negative yardage in the original data.)*



Yards to end zone is again the most important feature in predicting yards gained on a play. Whether or not the team scored is second in importance, but relatively much less important than yards to end zone. Yards to end zone limits the maximum number of yards that can be gained on a play, so it is not unexpected that this feature is important to predicting the yards gained on a play.



Future Work / Considerations

These models are limited to offensive plays during home games for Big 10 and SEC teams. There is an abundance of data from 2019 for away games and teams outside these conferences and utilizing this data would be a logical next step.

Access to more detailed play-by-play data, such as offensive formation, would be an interesting addition to these predictions, as well as an avenue for exploring differences between conferences.