**Predicting NFL Rookie Contract Extensions**

Sarah Cernugel, Drew Duffy, and Josh Garzaniti

**Introduction**

The National Football League is one of the most popular sports leagues in the world, containing 32 esteemed teams. In the 2024 NFL draft, there were roughly 257 rookies acquired by teams, among a population of around 2,200 eligible players (*BBC Sport*, "NFL's Fantasy Future"). This small sub-sample then, based on their respective rookie contract player-performances, is either offered a contract extension with the team that drafted them, offered an extension with a different team, or is not offered an extension at all. Every year, contracts become a hot talking point in the NFL due to each team’s limited “cap space,” or the fiscal space teams allocate to the budgeting of their respective rosters. Teams use a mixture of this cap space to: sign rookies to their initial contracts, re-sign notable players, and sign free agents, or individual players that are currently un-signed as a part of the NFL player base. The installation of a regulated league-wide cap space helps to limit what teams can pay players on their active roster and thus spread out competition with more balanced rosters throughout the league (*Schultz*). This places a unique emphasis on which rookie players will get extended and which will not.

The problem we are trying to solve is two fold: how can NFL general managers accurately predict player extensions to their draft team in the NFL based on rookie player positions, and what factors impact rookie NFL player’s ability to get re-signed. The first part can be turned into a binary classification prediction where “1” represents a player who has been re-signed by their draft team after their rookie contract, and “0” represents a player who has not been re-signed by their draft team after said rookie contract. The second, will involve more general logistic modeling and factor analysis. It is important to note that the scope of this analysis and future scalability may be limited, because it is possible in the NFL for a rookie to not get signed to an extension by their original draft team, and still be signed to a contract by one of the other 31 teams in the league.

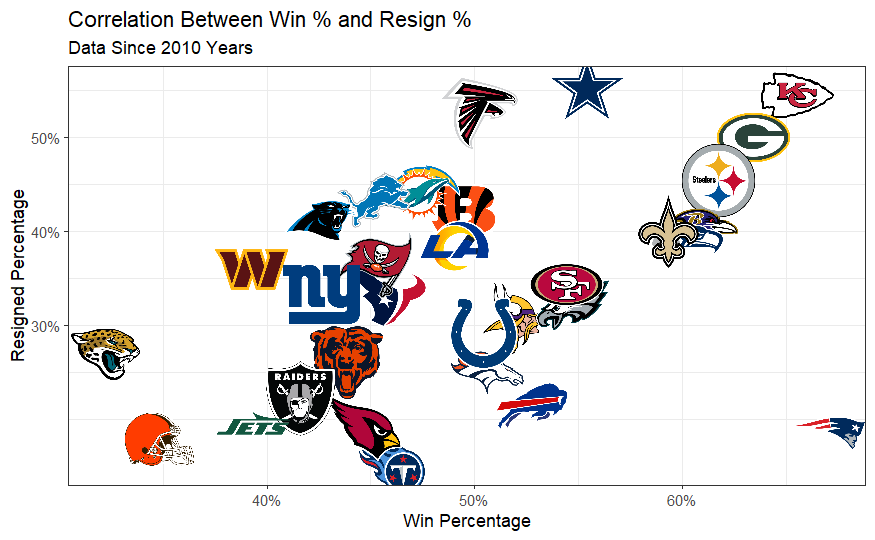
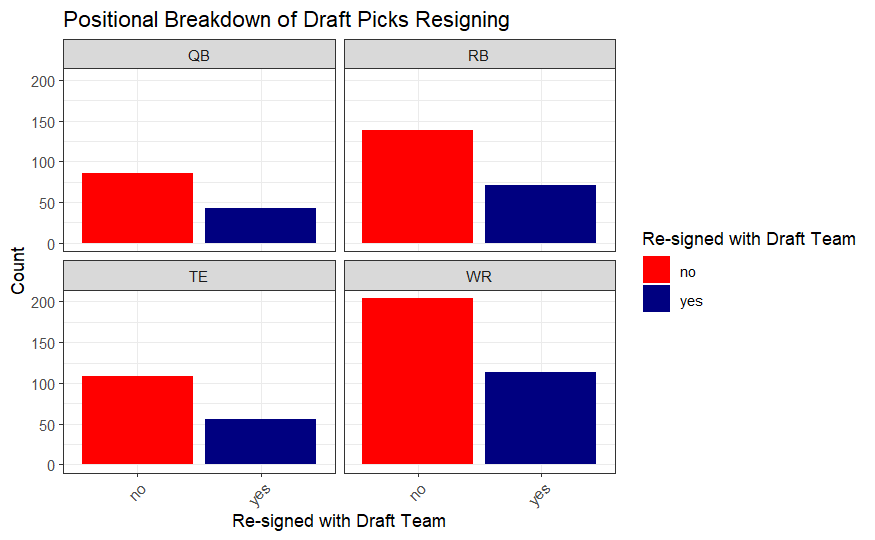
**Explanatory Data Analysis**

The data used for this project comes from the *nflverse* package in R. From that package, there were two main dataframes pulled: nfl contracts and offensive player specific stats. The NFL contracts initial data frame contained 18,520 observations and 25 unique variables, notably including features such as: player, position, team, year signed, contract fiscal details, draft round and draft overall. The offensive player specific data frame consisted of 125,076 observations and 53 columns, which contained features related to position-specific metrics such as: yards, first downs, EPA, fumbles, two-point conversions, target share, yard after the catch, etc… One of the challenges of using these datasets is that it contains duplicate observations for players, because statistics are collected on a weekly basis, and players earn multiple contracts over the span of their respective careers. What was done to address these issues is that data from the statistical side was aggregated into yearly sums, and grouped by for the first 4 years after their draft date. This became essentially “Historical Rookie Statistics.” On the contract side, since players had multiple contracts, a filter was created to limit contracts to five or less years after the player was initially drafted. This shifted the focus from career spanning contracts to only the first contract extension a player received after their rookie contract.

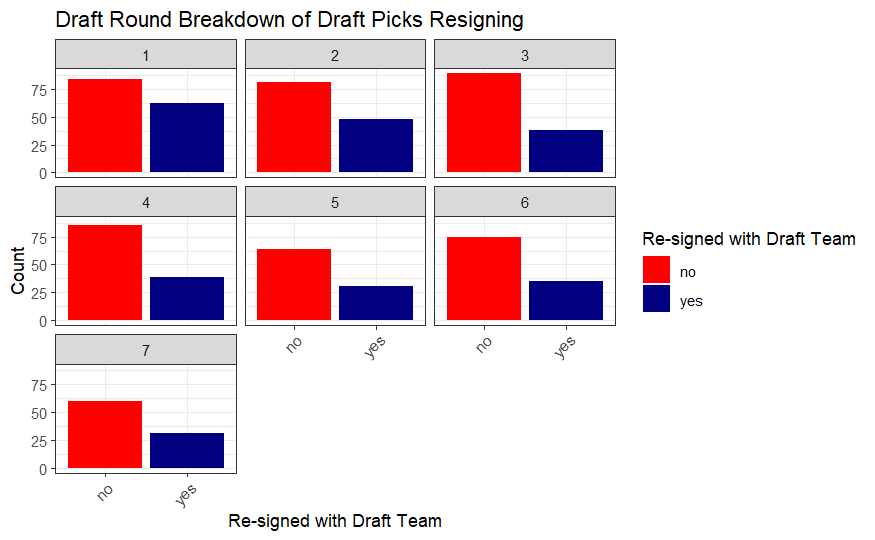
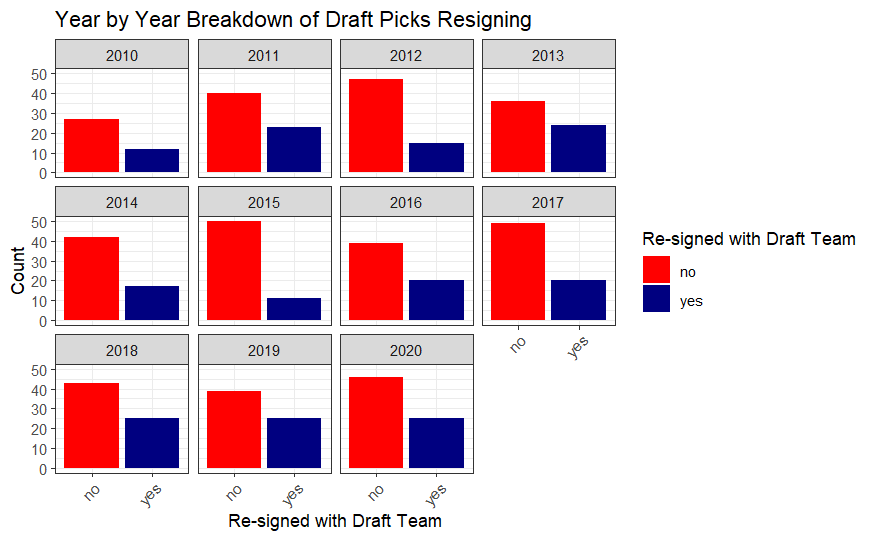
There were a few NA and null values that had to be dealt with in the process of cleaning up both data frames for analysis. Firstly, player statistics that got loaded into *nflverse,* mainly from the 49ers, assigned a NA value to their “draft team” column. This impacted the ability to analyze said players because “re-sign team” could not be compared to an NA value to create the outcome variable “re-signed\_with\_draft\_team”. Those NA values had to be extracted out of the data set and manually loaded back in with the correct “draft team” name. Outside of the 49ers there were also players who got drafted by teams but who had NA values for all of their statistics because they never played. Those players had to be removed completely from the dataset alongside an additional two players, Terelle Pryor and Josh Gordon who also had invalid data. Once all of those issues were taken care of, “contracts” and “stats” were joined on unique player IDs. The final contract and stats dataset had 822 observations and 44 columns.

For an additional question that was explored, the offensive stats dataset was adjusted to only show stat totals from players’ rookie season. That was then combined with the contract dataset and filtered to include the same players as before. After the dataset was cleaned, four separate datasets were created for quarterbacks, running backs, tight ends, and wide receivers.

The five individual attributes used for exploratory analysis were team win percentage, re-sign percentage, year, position, and draft round. Shown in Figure 1 below, team win percentage and re-sign percentage were plotted in a scatter plot to show the relationship between the two. Teams in the top right corner are the ones who have been successful on the field and in re-signing their rookies from 2010-2020. This implies that they have a great scouting department that is finding players to draft that fit their team and contribute to the teams success enough to be re-signed following the expiration of their rookie deal. The teams in the bottom left corner, like the Browns, are the teams that have not been playing well and are commonly not re-signing their drafted players. Teams like this are likely in a constant cycle of failing to develop the talent they draft and not finding players who can help their team win.

**Figure 1:** Team win percentage vs. re-sign percentage. **Figure 2:** Re-sign count by player position.

**Figure 3:** Re-sign count by draft round. **Figure 4:** Re-sign count by draft year.

The analysis for position, draft round, and draft year were all done using bar charts. Figure 2 above shows how position can impact re-signing with a draft team. There were no significant differences between the four positions, and it appears that around two times as many players do not re-sign with their draft team than do. Figure 3 breaks down the re-signing by draft round. First round players have the most re-signed players compared to not re-signed players. This is likely due to the fact that the most talented players are drafted in the first round and are more likely to make a significant enough impact on their drafted team to be re-signed. Rounds two through seven have less players resigning with their original draft team. The last visualization, Figure 4, shows the number of players re-signing based on the year they were drafted. The two years that stick out the most are 2012 and 2019. In 2012, a high proportion of players re-signed with their draft team following their rookie deal. In 2019, the opposite happened and there was a low proportion of players who re-signed with their draft team. This shows that the year you were drafted can have an impact on your likelihood of getting re-signed. There are a variety of reasons why this could be such as the talent level in that year's draft, coaching and general manager changes, etc.

**Learning Algorithm Training and Testing**

*Random Forest*

For the Random Forest Model, after the “contract\_and\_stats” data frame had been formatted correctly by applying the right data types to specific variables, and cleared of any missing or NA values, it was partitioned into two indices of training and validation data on a 60% and 40% split respectively. The “clean training data” was then filtered down to features that would not contain noise when running the binary classification model. Such features removed were player name, and player ID. For an initial Random Forest Model, the target variable “Resigned\_with\_draft\_team” was taken as a function of every other variable in the clean data set. For the first model, the parameters were left as standard considerations: a tree size of 500, node size of 1, and then using all 40 of the variables in the dataset to predict. When predicting off of the validation data index, it produced a confusion matrix with the following results: Accuracy .6494, sensitivity .3982, a specificity value of .7814. This meant effectively our first model was more effective at predicting true negatives (players who didn’t get re-signed) vs true positives (players who did). Unfortunately it also came with a p-value of .0621, which meant the results of that first model weren’t significant. That said, it was only just this first model that came with such a high p-value.

In between the first and second Random Forest models, OOB error (Out of Bag Error) estimate was plotted against the number of trees used (Figure 6). This graph displayed a downward trend where OOB bottomed out around the 300-325 tree size range. During this period between modeling, the node size was charted against the OOB accuracy (Table 1), or 1-the OOB error rate. When the resulting table was produced, it displayed that the best node size for our model was 50. With that the next Random Forest Model was created; this time with slightly better results. Accuracy increased to .6585, sensitivity slightly fell to .3894, and specificity bumped up to .8000. More importantly, the p-value for this latest model dropped all the way down to .0181, which with a benchmark of .05 now meant a significant Random Forest Model had been created.

The final Random Forest model was created using a cross validation index. 5 random groups were created out of the clean training data, and each row then randomly assigned to one of those 5 groups. This created data that was randomly arranged to use for our model, which would help it perform better on data it hadn’t already seen, and thus cause it to both be even more accurate and at the same time overfit less. The results of this final model included inverted specificity and sensitivity results from the previous models (Figure 7). This time, sensitivity dramatically increased to .9104 and specificity fell to .3415. Better yet, the total accuracy was .6944 and the p-value shrunk to .0005. All in all, this tuning helped to create a fairly solid model in terms of predicting which rookies would be re-signed and which wouldn’t.

*Linear Regression*

Binary generalized linear models were used to understand how player stats impact whether or not a player will re-sign with the team that drafted them. To ensure that the model was as accurate as possible, the data was split based on player position. This was done to avoid the value of certain stats being weakened by other positions. For example quarterbacks are the only ones with passing yards, so keeping running backs in the dataset would not give as accurate a representation of how well passing yards can predict whether or not a player will be re-signed.

Once the data was split, models were created for each position. Variable selection was done by plugging in variables one by one and selecting the model with the lowest AIC. The p-values of the variables were also monitored in the process. For running backs, the best model included receiving EPA (expected points added), rushing fumbles, rushing yards, and receiving air yards. This shows that teams value great runners, who can be a threat in the passing game and take care of the football. The best wide receiver model has receiving EPA, receiving air yards, and receiving first down. The model shows the importance of moving the chains and putting points on the board as a receiver. The quarterback model includes carries, rushing fumbles, rushing first downs, rushing EPA, and rushing two-point conversions. It was interesting that this model did not contain any passing variables, and that some of these rushing variables had negative coefficients. This model shows that versatile dual-threat quarterbacks are valuable and the ability to gain first downs and convert two-point conversions is crucial. The last position model, the tight end model, values receiving yards, receiving touchdowns, and receiving air yards. The only positive coefficient was receiving yards which signifies that tight ends who can be a passing threat and have great yards after the catch have a better chance of re-signing with their draft team. Overall, the models highlighted what stats are telling of who will be resigned and showed that versatility and production at each position matters.

Models were also created using team, position, and draft round to predict whether a player will re-sign or not. The team model did not have significant results, however the position and draft round model did. The position model showed the QB position was significant and had a coefficient of 0.69. This result signifies the importance of the quarterback position in the NFL. If players are not able to produce for a team while on their rookie contract, they will go out and replace them, likely with another rookie who will have a lower contract value than a veteran. The draft round model had all negative coefficients, and the third round of the draft proved to be the largest and most significant. Players in the third round of the draft are not as likely to re-sign with their draft team.

**Discussion and Conclusion**

Our highest random forest model yields a 69% accuracy, which is a fairly accurate estimation predicting players that will re-sign based purely off of their individual/team statistics. This appeared to be a decent indication of if a player would resign with the team that originally drafted them. Since the primary model only accounted for a few very narrow variables (draft team, position, and offensive statistics), there is still a lot of room left for interpretation. A considerable amount of interest was placed in the faceted bar plots constructed, showing that there is a very similar relationship among those who are resigned and not when broken down into levels.

The positional chart (Figure 2) shows that there isn’t a clear advantage in terms of specific skilled positions when it comes to who is more likely to be re-signed with their original team. When breaking it down by the round that the players were drafted (Figure 3), there were no telltale signs either. This prompted some questions into external factors, not modeled, that impact whether or not a player will be re-signed.

Going forward, if we were expanding the reach of the model, we would want to first use the random forest model to predict on future draft classes that are entering the end of their respective rookie contracts, as well as expand the models used to take into account a number of other factors. The main ones revolve around salary cap management and draft strategy. We discussed how the salary cap and teams being able to or unable to spend will have a huge impact on whether they re-sign their players or not. There are simply too many unpredictable factors in the draft, such as a player rising or falling on a draft board, a team selecting a position of need or a position of value, and many others. However, these are surely factors that top executives weigh when deciding whether they want to make the financial investment on these players.

In all, when taking into account the limited parameters and variables accounted for, we deemed this to be a fairly successful model. Tailoring a model that could potentially use historical data from a certain team or a certain general manager’s propensity to resign his players would then be the next step. Then that future model could assign predicted values to each player based on their individual likelihood of being re-signed with their respective team.

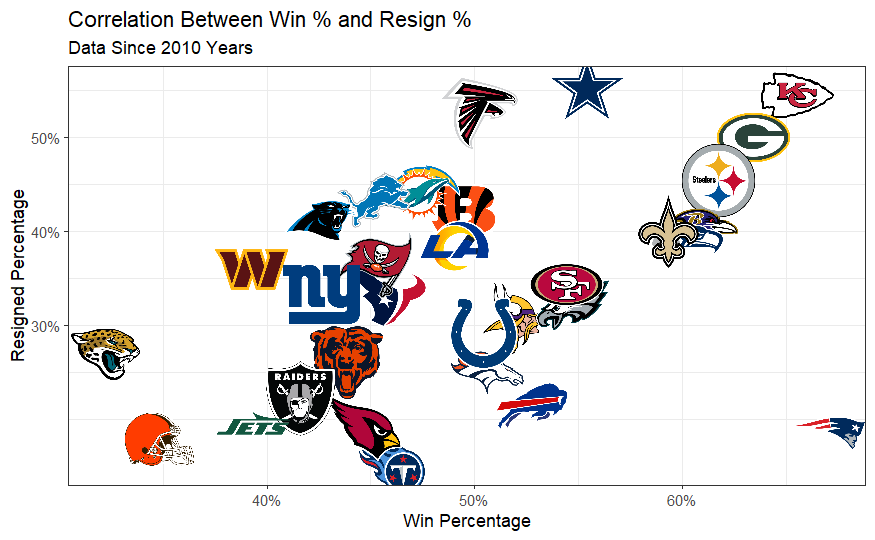
**Works Cited**

BBC Sport. "NFL's Fantasy Future: The Impact of Data and Analytics on the Game." *BBC Sport*, 29 Sept. 2024,<https://www.bbc.com/sport/american-football/articles/cxx8pjdn32zo>. Accessed 29 Sept. 2024.

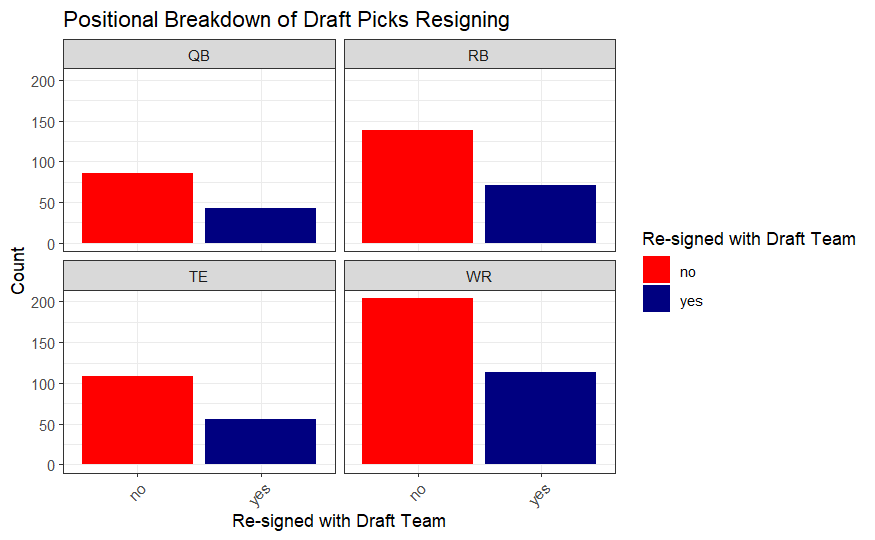
*nflverse*. Version 1.0.3, nflverse, 2023,<https://github.com/nflverse/nflverse>. Accessed 4 Oct. 2024.

Schultz, Jason. "How Does the NFL Salary Cap Work?" *Pro Football Network*, 27 Sept. 2023,<https://www.profootballnetwork.com/how-does-nfl-salary-cap-work/>. Accessed 29 Sept. 2024.

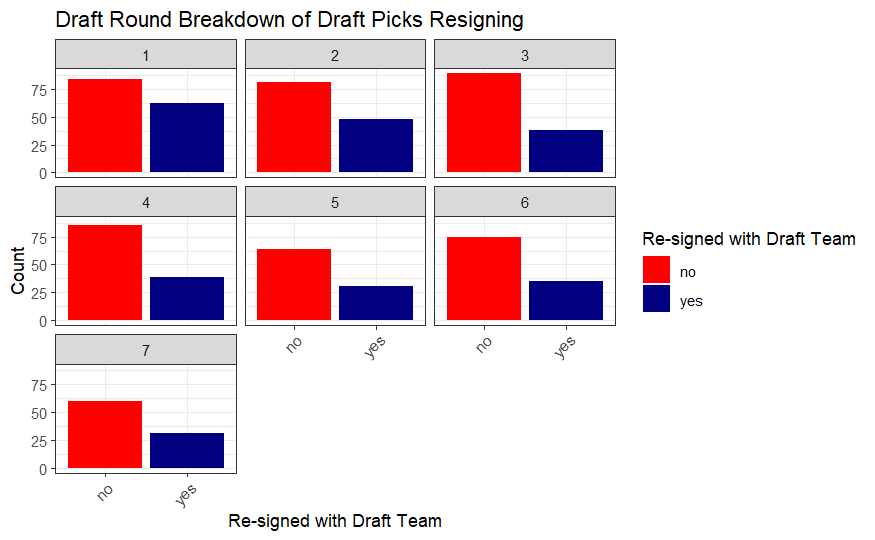
**Appendix**



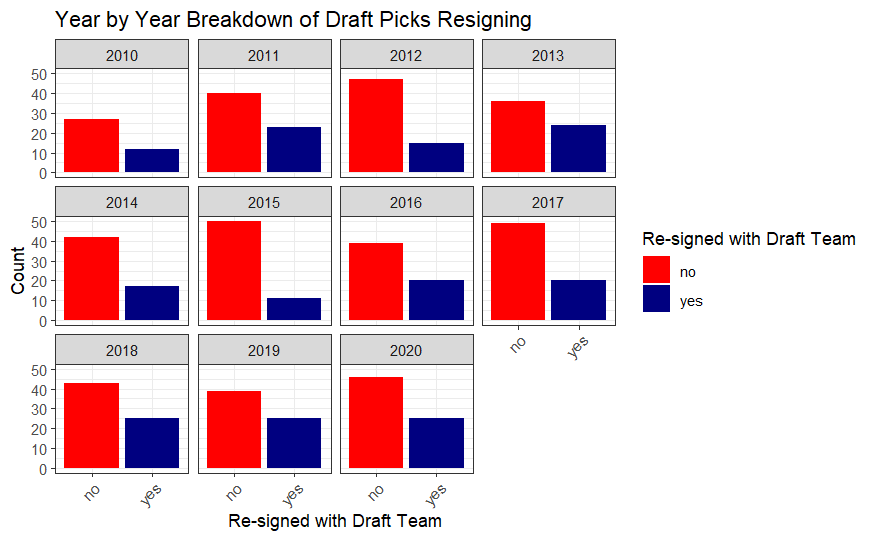
**Figure 1:** A scatter plot showing team win percentage vs. rookie re-sign percentage since the 2010 season.



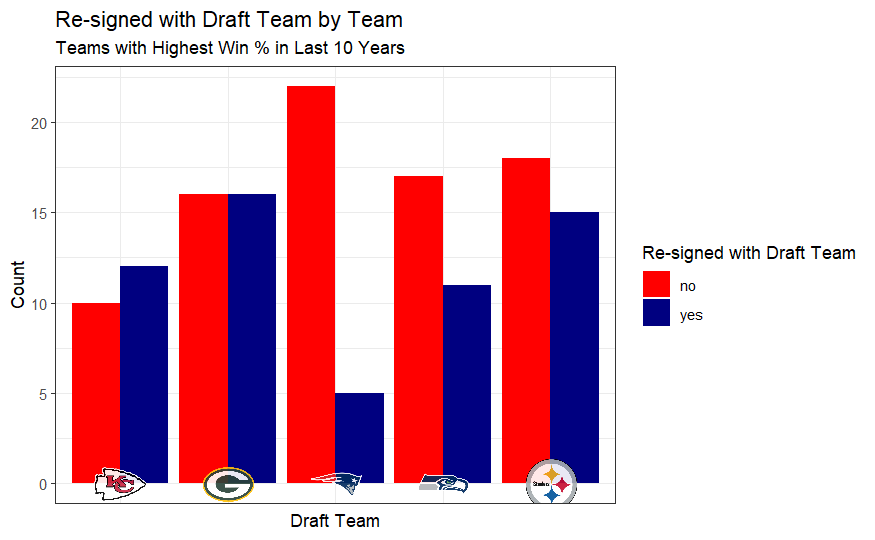
**Figure 2:** Bar charts showing the number of rookies who did or did not re-sign with their draft team, split up by position.



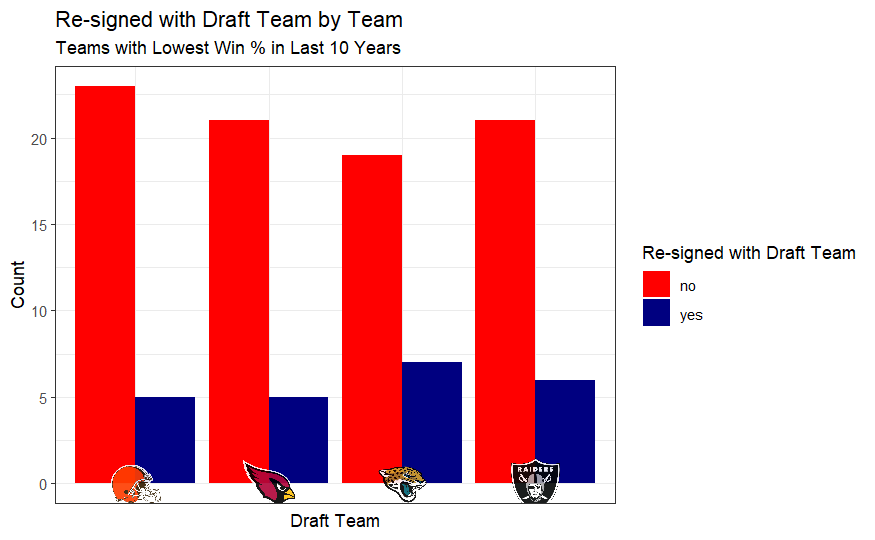
**Figure 3:** The number of rookies who did and did not re-sign with their draft team, split up by draft round.



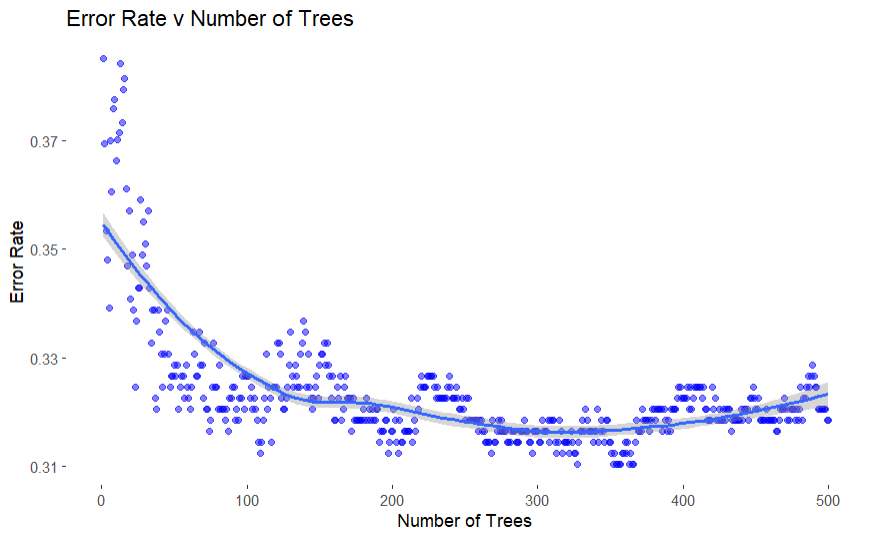
**Figure 4:** The number of rookies who did and did not re-sign with their draft team, split up by draft year, since 2010.



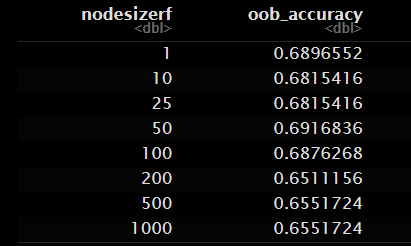
**Figure 5:** Bar chart showing the five teams with the highest win percentage over the last ten years and the total number of rookies who re-signed and did not re-sign with the team.



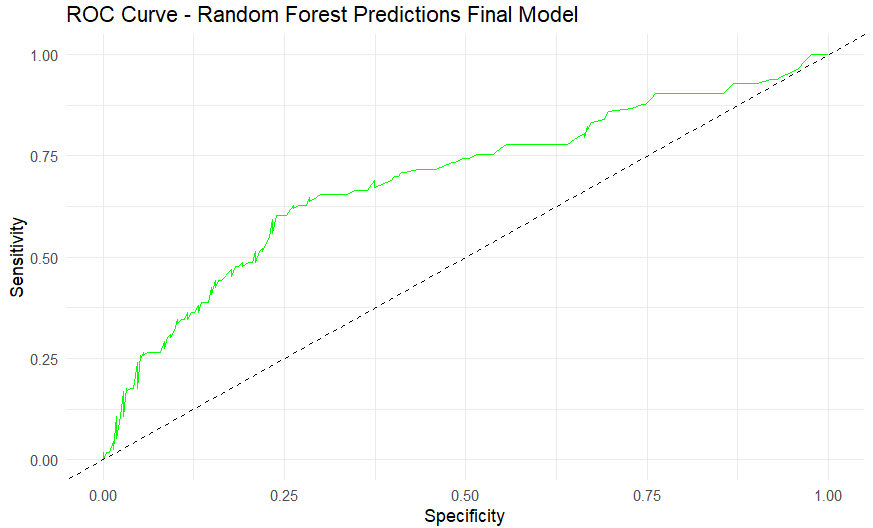
**Figure 5:** Bar chart showing the five teams with the lowest win percentage over the last ten years and the total number of rookies who re-signed and did not re-sign with the team.



**Figure 6**: The OOB Error Estimate by N Trees Curve showing the best range at ~300.



**Table 1**: The Node Size OOB Accuracy Table displaying the best size of 50.



**Figure 7:** The ROC curve for the Random Forest model, containing an accuracy of 0.6944.