Data-Driven Offensive Insights: Using Analytics to Improve Offensive Success Rates

Clayton Fogler

2025-12-17

Table of contents

Abstract

There are a lot of factors that go into winning football games. One of the biggest factors is an efficient offense, and at the core of an efficient offense is an offense with high success rates on their plays. This paper will investigate how a certain Division 3 football team can maximize their offensive success rates. Using data collected from their past two seasons, this study aims to examine how variables like Distance and Down can impact the chances of a successful play. After cleaning and organizing over 1,000 football plays, I created many models, using multiple methods, to determine the best way to predict success rates in the team's offense. Five models were built throughout the paper that each predict success rate based on a combination of the following variables: RP (run pass), DN (Down), DIST (Distance to go), and PERSONNEL (Offensive Personnel). To compare the models, I calculate a classification rate for each to see which model best predicts the team's success rates on offense. By identifying the most effective model, I will be able to highlight the most important factors that give the best chances for a successful play. This will provide further insights that can be used in live game-play calling to create a more efficient offense.

Introduction

In this decade, we have seen major advancements in the use of analytics in sports. Whether used from a business perspective, player development prospective, or game play prospective, there are so many cool ways to use analytics to improve your team. In football, one of the greatest uses of analytics is finding tendencies in a team's playing schemes, measuring their efficiency, and finding a way to get the upper edge on the field. As a kicker on a small Division 3 football school, the use of analytics is small as stats are kept on a very minimal level. However, just because very little data is recorded, doesn't mean we can't find ways to use the data to improve our success.

The goal for my Senior Year Experience is to evaluate my teams' offensive success rates using both exploratory statistics, logistic regression analysis, classification trees, bagging, and random forest to improve our offensive efficiency. By studying the play by play data from both past and current seasons, I will be able to find patterns in both our play calling tendencies, our efficiency in specific situations, and potentially develop a program that simulates our drives.

Throughout this paper, I will use the term SUCCESS as the response variable. The variable SUCCESS refers to whether or not the offensive football play was successful or not. To figure out how to measure success, I went online and found multiple sources that defined success similarly. One website, titled Football Study Hall, defined success rate as "at least 40 - 50% of the yards to go on 1st down, at least 50-70% of yards to go on second down, and first down achievement on third and fourth down" (C, 2012). To get a first down achievement on third and fourth down, an offense must gain 100% of the yards to go. For the purpose of my paper,

I defined success as needing 40% of the yards to go on 1st Down, 60% of the yards to go on second down, and 100% of the yards to go on third and fourth down.

This research will directly help the team as identifying our tendencies and efficiency levels will tell us our strengths and weaknesses, improving our play designs, playbook, and player development. By using our different model methods, along with our exploratory research, we will be able to build future models that can be updated as the season moves on to see if our changes are working. This will help our coaching staff handle uncertainties when trying to game plan for the game.

My project will begin with basic exploratory data analysis based off the last two years of games to get an understanding for the offense and find early trends. Things we will look at include success rates based on run/pass, distance, down, and personnel. We will then use these findings to build some logistic regression models before diving into classification trees, bagging, and random forests. In the end, we will compare our models to determine which one is the best at predicting successful football plays.

Data Gathering and Tidying

Before discussing how we tidied our data, we must begin to explain how it was gathered. Like most teams at the division 3 football level, we used a software called Hudl to keep track of game film and statistics. It is a subscription-based website in which teams can upload film from games and practice, along with data such as down, distance (distance to go), personnel, and more. After recording plays and uploading them to hudl, the coaching staff will then go in and enter all of the information related to each play. Thankfully, Hudl has an option for coaches to download the statistics as an Excel file, that can then be transformed into a csv file.

After uploading each game, I combined all the data into one data set called `all_seasons`. This data set includes a lot of variables, however, the main variables that are used throughout the paper include:

- DN (stands for what down the play is. 0 stands for the first down to begin a drive, 1 for 1st down, 2 for 2nd down, 3 for 3rd down, and 4 for 4th down.)
- DIST (stands for the distance to go)
- RP (whether the play was a run or a pass. R for run and P for pass.)
- PERSONNEL (what players were on the field. With numbered personnel, the two numbers are used to tell the offense how many tight ends and running backs will be on the field. After factoring 5 offensive lineman and 1 quarterback, this means there are 5 remaining players allowed on the field. The first digit refers to the number of Tight Ends on the field. The second digit refers to the number of running backs on the field.

Finally, if those two digits do not add to 5, the remaining number of players are wide receivers. For example, 12 personnel means their is 1 tight end, 2 running backs, and 2 wide receivers on the field. 11 personnel means their is 1 tight end, 1 running back, and 3 wide receivers on the field. With named personnel's, this usually refers to a specific game plan package that is unique to certain players.)

- SUCCESSFUL (whether or not the play was successful. We defined SUCCESSFUL in the introduction).
- SITUATION (Situation was used in our exploratory section to help look at the distances to go a little easier. Short means the DIST was between 1-3 yards. Medium meant 4-7 yards, Long meant 8-10 yards, and very long meant 11+ yards).

With our data collected and gathered, it was time to begin some exploratory research to see if we notice any early patterns in SUCCESS.

Exploratory Research

To begin my exploratory research, I wanted to create an assortment of graphs and charts to help visualize my data and see if any early trends arise. I created a chart for each one of my predictors to see how success related to each of them:

Success based on RP

```
# A tibble: 2 x 3
  RP    plays `Success Rate`
  <chr> <int> <chr>
1 P       657 41.2 %
2 R       448 37.3 %
```

Our offense appeared to have a little more success passing the ball then running, however, it is not by much. This shows potential signs of RP not being the strongest predictor, especially on its own.

Success based on DN

```
# A tibble: 5 x 3
     DN plays `Success Rate`
  <dbl> <int> <chr>
1     0   194 44.3 %
2     1   261 42.9 %
3     2   362 39.8 %
4     3   244 31.6 %
5     4    44 43.2 %
```

Success rate on Third Downs seems to be a lot less compared to the other Downs. This shows that downs has the potential to be a strong predictor.

Success based on DIST

```
# A tibble: 4 x 3
  SITUATION       plays `Success Rate`
  <chr>           <int> <chr>
1 Short (1-3)       127 59.8 %
2 Medium (4-6)      137 45.3 %
3 Long (7-10)       692 39.5 %
4 Very Long (11+)   149 18.1 %
```

Because there are so many different Distances that occurred over the past two seasons, I created the variable `SITUATION` to help visualize the effect of distance on success. Situation is broken into 4 categories: Short (distance = 1-3 yards), Medium (distance = 4-6 yards), Long (distance = 7-10 yards), and Very Long (distance = 11+ yards). We see above that distance has a huge impact on the success rate, with an over 20% decrease between Long and Very Long. Early on, DIST seems to be the strongest predictor of SUCCESS.

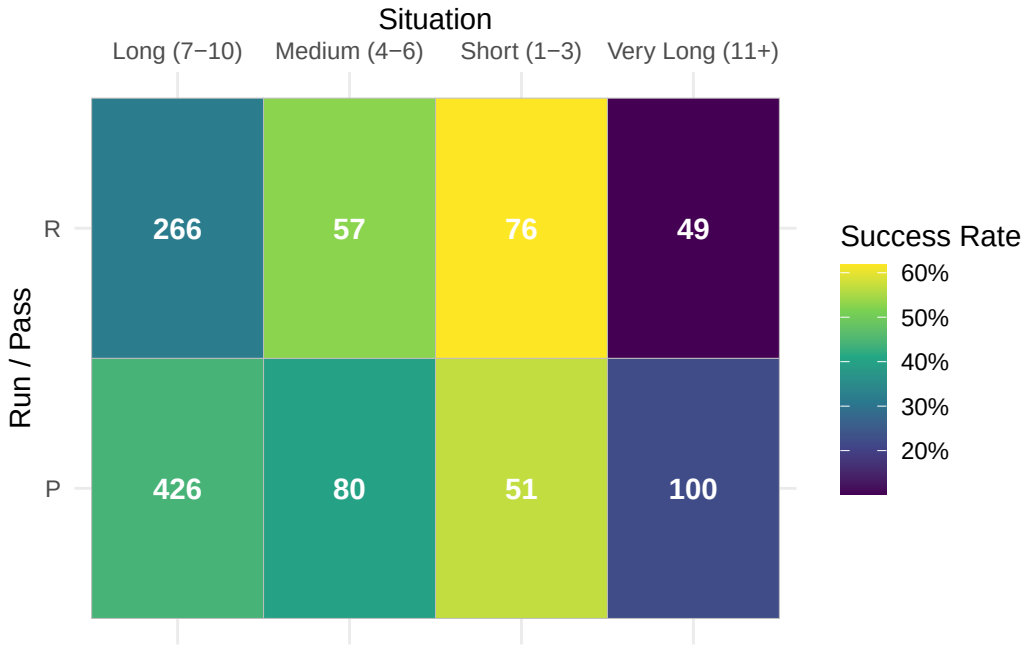
Success based on PERSONNEL

```
# A tibble: 14 x 3
   PERSONNEL plays `Success Rate`
   <chr>     <int> <chr>
 1 0             4 100 %
 2 1            44 40.9 %
 3 10           75 52 %
 4 11          747 38.4 %
 5 12          153 40.5 %
 6 2             1 0 %
 7 20            2 0 %
 8 21           30 33.3 %
 9 ALPHA         6 16.7 %
10 BEEF         17 47.1 %
11 JET           3 0 %
12 PIG          13 46.2 %
13 PORK          9 33.3 %
14 VICTORY       1 0 %
```

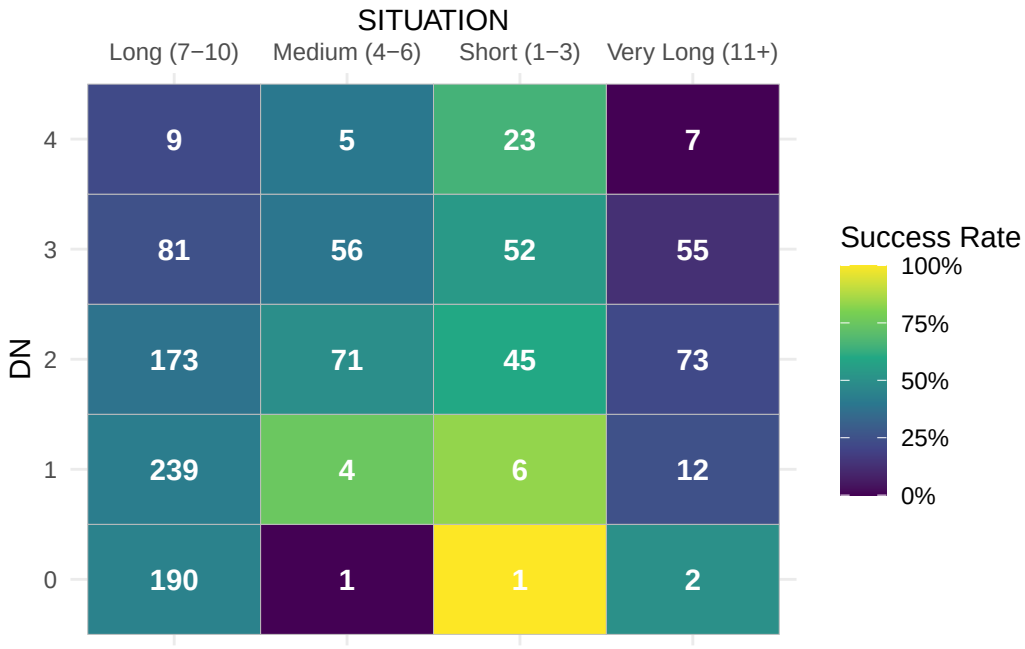
We notice that there are some large jumps in success rate between personnel groupings, however, there is also a large difference in play count for each. 11 personnel (1 running back, 1 tight end, 3 wide receivers) was used 747 times, with the next closest being 12 personnel (1 running back, 2 tight ends, 2 wide receivers) with 153 plays. The lack of even plays will probably limit the effect personnel has on predicting success.

We can also visualize multiple variables at once, using various heat maps to try to find some correlation between variables. Because we have early evidence that DIST will be our strongest predictor, lets compare the other 3 variables with Situation (our modified distance variable).

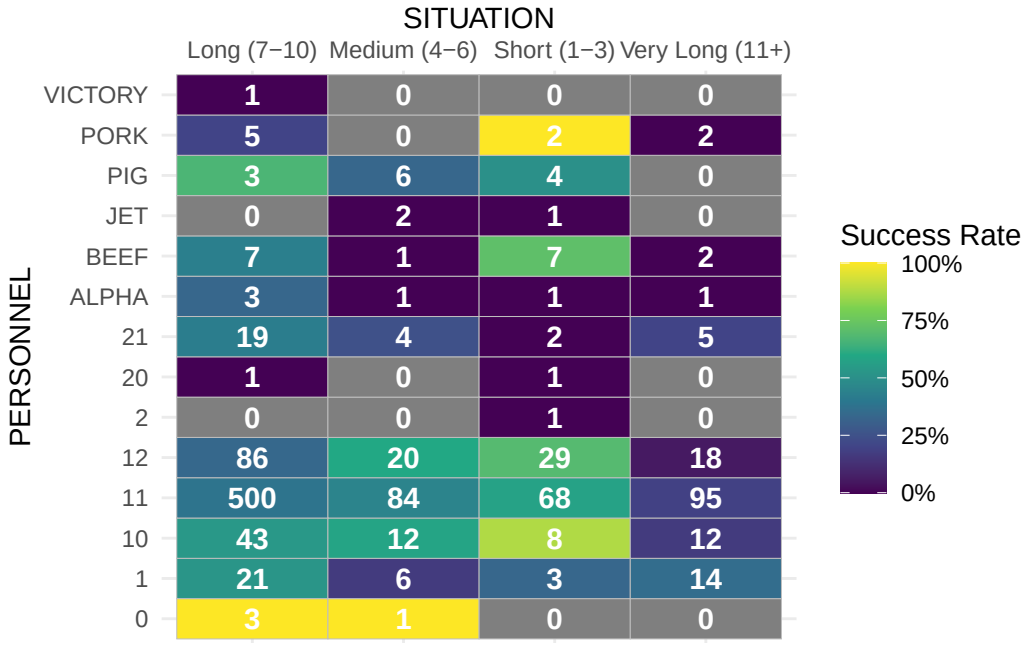
Situation and RP



Situation and DN



Situation and Personnel



The least effective graph above was our graph displaying the relationship between personnel and situation. Again, this isn't a surprise as there were 14 different personnel groupings recorded over the last two seasons, with 11 personnel being used over 700 times. However,

this doesn't mean it will be a bad or ineffective variable. We will learn more later on in our model building if personnel is worth having. While a RP and DN definitely have relationships with Success Rate, it is clear to see with these graphs that the strongest relationship is the relationship between SITUATION and SUCCESS. After doing some exploratory research, it is time to dive into our model building.

Modeling Methods

Method 1: Logistic Regression

To start off our model building, I wanted to use a logistic regression model. Logistic regression models are nice because they allow us to look at the relationship between a response variable and however many predictor variables we want to use. Because we are modeling SUCCESS, a categorical variable, we use Logistic Regression instead of Linear. My first step was to build 4 models, one for each variable that I am interested in using: DN, DIST, PERSONNEL, and RP:

```
logRP <- glm(SUCCESSFUL ~ RP, data = all_seasons_logit, family = binomial)
AIC(logRP)
```

```
[1] 1486.291
```

```
logDIST <- glm(SUCCESSFUL ~ DIST, data = all_seasons_logit, family = binomial)
AIC(logDIST)
```

```
[1] 1442.719
```

```
logDN <- glm(SUCCESSFUL ~ DN, data = all_seasons_logit, family = binomial)
AIC(logDN)
```

```
[1] 1484.043
```

```
logPERSONNEL <- glm(SUCCESSFUL ~ PERSONNEL, data = all_seasons_logit, family = binomial)
AIC(logPERSONNEL)
```

```
[1] 1489.597
```

With these starter models, the model with the lowest AIC was logDIST (AIC = 1442.7), which uses the DIST variable to predict SUCCESSFUL. We saw a strong relationship with DIST and SUCCESS in our exploratory section, so it makes sense as to why that model had the best AIC. However, we know as we begin to add more variables, we will probably start to see better models. The best way to find the best combination of variables for a logistic regression model would be to create a model with all of our variables, and then use the `step()` function to find which model is the best.

```
Start:  AIC=1414.76
SUCCESSFUL ~ RP + DIST + DN + PERSONNEL

            Df Deviance    AIC
<none>           1374.8 1414.8
- PERSONNEL 13   1400.8 1414.8
- RP         1   1381.8 1419.8
- DN         4   1411.0 1443.0
- DIST       1   1447.9 1485.9

[1] 1414.765
```

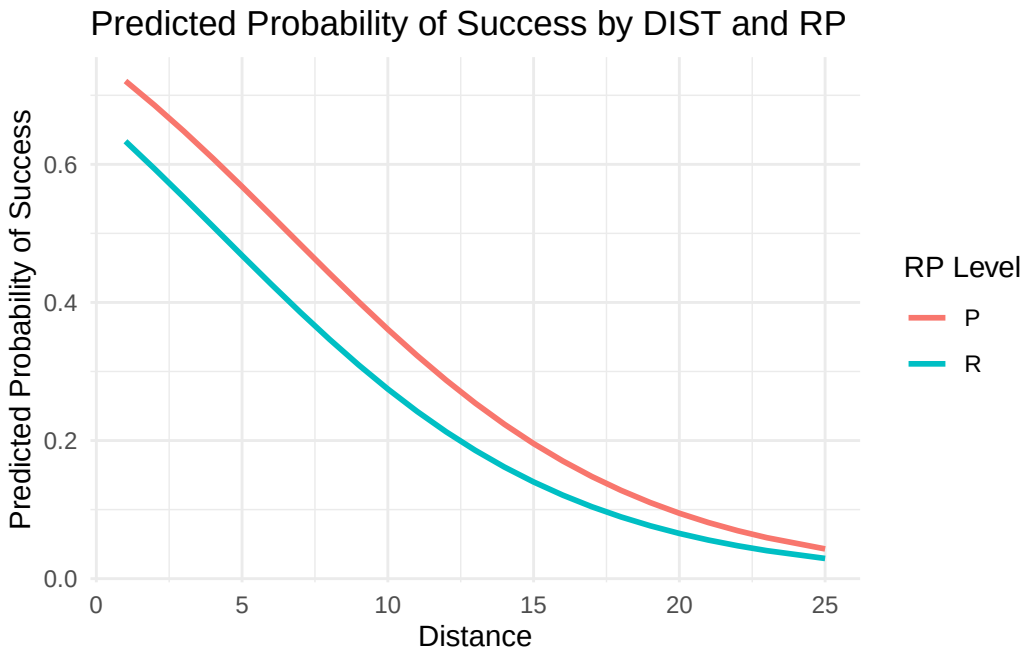
LogBEST returned back a model with all four variables with an AIC of 1414.8, which was better than our logDIST model. However, I noticed in the coding output that a model with DN, DIST, and RP (no PERSONNEL) got the same AIC of 1414.8. To decide whether or not to keep the PERSONNEL term, I compared the two models BIC and found that the model without PERSONNEL resulted in a lower BIC. Because of this, my final predictor variables are DN, DIST, and RP.

Next, I used the `train()` function to build the model again, but added 5 fold cross validation. The biggest reason for doing this was to be able to put every model I build through the same validation process to make them easy to compare.

```r
# Extract cross-validated predictions
cv_preds <- logFinal$pred

# Convert probability of "Yes" to predicted class
cv_preds$pred_class <- ifelse(cv_preds$Yes > 0.5, "Yes", "No")

# Actual outcomes
actual <- cv_preds$obs

# Compute classification rate
classification_rate <- mean(cv_preds$pred_class == actual)
classification_rate
```

```
[1] 0.60181
```

I then calculated the classification rate for this model. Using DN (Down), DIST (Distance to go), and RP (whether the play was a run or pass) to predict SUCCESS resulted in a classification rate of 0.602, meaning our model correctly predicted 60.2% of the actual plays. I wanted to also visualize this model so I built the following graph.



In the graph, we notice that run and pass are fairly parallel, as they begin to get closer as we approach a distance greater than 15 yards. It is important to note that there is no overlap in the two slopes (this will be important to remember for later).

To be correct 60.2% of the time is fairly good, however, there are many methods that can be used to predict outcomes. Lets now try Logistic Regression Models that include interaction terms.

Method 2: Logistic Regression with Interaction

To begin our logistic regression model with interaction, we again build a model with not only all the predictors, but all of the possible interactions as well.

```
logALL_inter <- glm(SUCCESSFUL ~ (RP  + DIST + DN + PERSONNEL)^2, data = all_seasons_logit, f
```

We then use the step function to find the best possible model.

```
Call:
glm(formula = SUCCESSFUL ~ RP + DIST + DN + PERSONNEL + RP:DIST,
    family = binomial, data = all_seasons_logit)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       17.05173  725.80130   0.023  0.98126
RPR                0.69986    0.35927   1.948  0.05142 .
DIST              -0.12171    0.02681  -4.540 5.63e-06 ***
DN1               -0.03441    0.19637  -0.175  0.86090
DN2               -0.60149    0.19323  -3.113  0.00185 **
DN3               -1.15386    0.22428  -5.145 2.68e-07 ***
DN4               -1.08286    0.38026  -2.848  0.00440 **
PERSONNEL1       -15.46133  725.80132  -0.021  0.98300
PERSONNEL10      -15.28114  725.80129  -0.021  0.98320
PERSONNEL11      -15.86130  725.80126  -0.022  0.98256
PERSONNEL12      -15.78246  725.80128  -0.022  0.98265
PERSONNEL2       -31.65865 1626.33622  -0.019  0.98447
PERSONNEL20      -31.45731 1226.91043  -0.026  0.97954
PERSONNEL21      -15.85799  725.80137  -0.022  0.98257
PERSONNELALPHA   -17.15094  725.80210  -0.024  0.98115
PERSONNELBEEF    -15.62492  725.80145  -0.022  0.98282
PERSONNELJET     -31.95057 1090.40555  -0.029  0.97662
PERSONNELPIG     -16.15056  725.80150  -0.022  0.98225
PERSONNELPORK    -15.84392  725.80165  -0.022  0.98258
PERSONNELVICTORY -30.75752 1626.33621  -0.019  0.98491
RPR:DIST          -0.13086    0.04015  -3.259  0.00112 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1484.1  on 1104  degrees of freedom
Residual deviance: 1364.0  on 1084  degrees of freedom
AIC: 1406

Number of Fisher Scoring iterations: 14
```

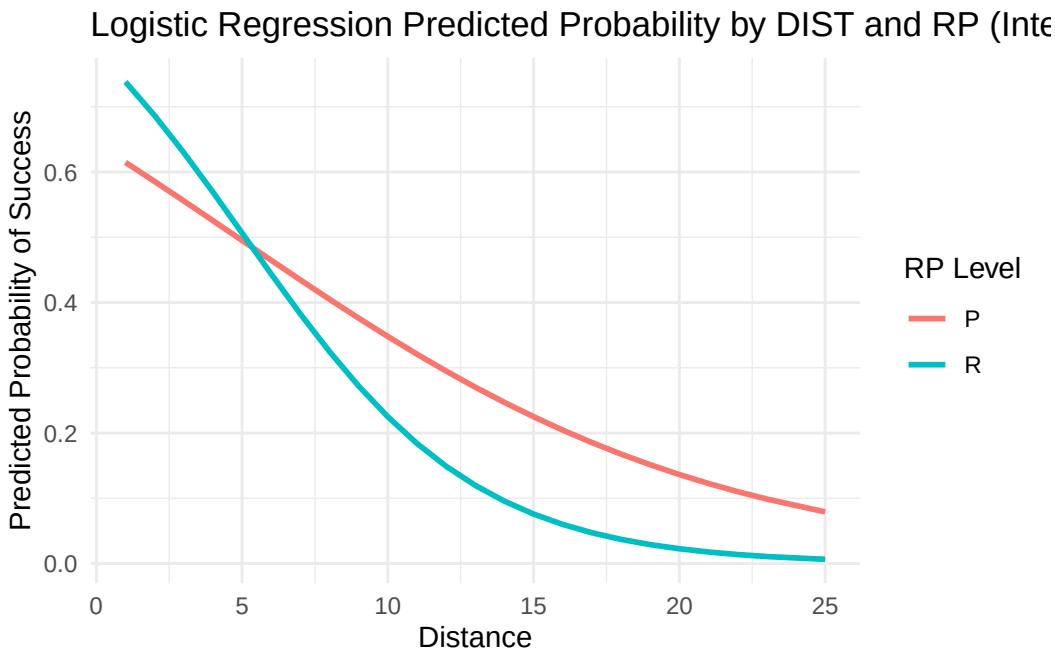
It seems that the best possible model using our 4 variables to predict success, based on AIC and using interaction terms, is a model with RP, DIST,DN, PERSONNEL, and RP:DIST interaction(AIC = 1406).

Next, I used the `train()` function again to build the model using 5 fold cross validation.

```r
# Extract cross-validated predictions
cv_preds_int <- LOGINTERACTION$pred

# Convert probability of "Yes" to predicted class
cv_preds_int$pred_class <- ifelse(cv_preds_int$Yes > 0.5, "Yes", "No")

# Actual outcomes
actual_int <- cv_preds_int$obs

# Compute classification rate
classification_rate2 <- mean(cv_preds_int$pred_class == actual_int)
classification_rate2
```

```
[1] 0.6162896
```

After the model goes through 5 fold cross-validation, I can now calculate the classification rate for this model. A model using DN, DIST, RP, PERSONNEL, and an interaction with RP and DIST to predict SUCCESS resulted in a classification rate of 0.616, meaning our model correctly predicted 61.6% of the actual plays. I wanted to also visualize this model so I built the following graph.



The biggest different between our first two methods so far is our graph of the predicted probabilities of success. In our graph with interaction, we see that the fitted curves of run and pass cross when distance is at about 5 yards. This means our interaction model is seeing a

shift in which pass plays are more successful in longer distances. We will do more comparisons at the end, however, we do begin to see that a model with interaction is a little better at predicting success than our model without interaction.

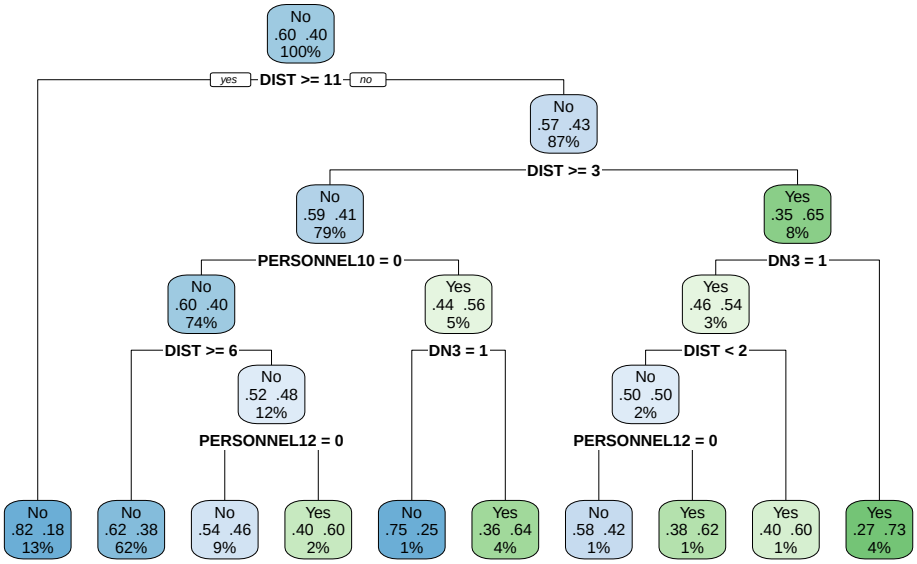
Method 3: Classification Trees

Logistic regression is not the only method that can be used in this scenario. Another effective method that can be used are classification trees. Classification trees is a kind of model that starts with one yes or no question that then breaks off into branches of yes or no questions that best predict what your model is trying to predict. For this paper, the classification tree is trying to find what pathways the data best follows to predict whether a play was successful or not. Our model is as follows:

```
ctrl <- trainControl(
  method = "cv",
  number = 5,
  classProbs = TRUE,
  summaryFunction = twoClassSummary
)
```

```
set.seed(123)

tree <- train(
  SUCCESSFUL ~ RP + DIST + DN + PERSONNEL,
  data = all_seasons_trees,
  method = "rpart",
  trControl = ctrl,
  metric = "ROC",
  tuneLength = 10
)
```

To help further understand what a classification tree looks like, here is the classification tree for my model. Note that this isn't my exact classification tree as my tree has too many branches to understand what is going on. I created a second version of my model that caps the tree depth to make the chart easier to see and understand.



In our partial tree, you see that the first partition is distance, showing that distance was the strongest predictor out of our variables. It is also interesting to note that personnel ranks higher than some of the other variables, which is not something I would've predicted based on our exploratory research.

```
pred_tree <- predict(tree, newdata = all_seasons_trees)

actual_tree <- all_seasons_trees$SUCCESSFUL

classification_rate3 <- mean(pred_tree == actual_tree)

classification_rate3
```

```
[1] 0.6651584
```

Since we already built our model with 5 fold cross-validation, we can jump right into finding our classification rate. Classification Rate for our Classification Tree is 0.665, meaning our model correctly predicted 66.5% of the actual plays. This is better than both our base logistic regression model, and our logistic regression model with interaction. Before we do a full comparison of all our models, we will build our final two models using bagging and random forest.

Method 4: Bagging and Random Forest

Classification trees can be really effective, however, they can sometimes be very sensitive to changes in the training data, as well as overfitting. That is where bootstrapping comes in. Bootstrapping repeatedly re-samples with replacement your original data, and creates new training data. Then, it runs a new classification tree based on the training data. It will do this as many times as you would like, helping reduce the variance in your model. Our bagging model is as follows:

```
set.seed(123)

football_bagging <- train(
  SUCCESSFUL ~ `RP` + DIST + DN + PERSONNEL,
  data = all_seasons_bagging,
  method = "treebag",
  trControl = trainControl(method = "cv", number = 5), # cv = cross validation, number is am
  nbagg = 200,  #the number of trees being built
  control = rpart.control(minsplit = 2, cp = 0) #
)
```

```
# predicted class
pred_class_bagging <- predict(football_bagging, type = "raw")

# actual outcomes
actual_bagging <- all_seasons_bagging$SUCCESSFUL

# classification rate
classification_rate_bagging <- mean(pred_class_bagging == actual_bagging)
classification_rate_bagging
```

```
[1] 0.7230769
```

We then can use this model and calculate our classification rate, which we get to be 0.724. This means our model correctly predicted 72.4% of the actual plays. This is a big jump from our classification tree method. Lets see if random forest makes a similar jump.

Random forest

```
ctrl2 <- trainControl(
  method = "cv",
  number = 5,
  classProbs = TRUE,
  summaryFunction = twoClassSummary,
  allowParallel = FALSE
```

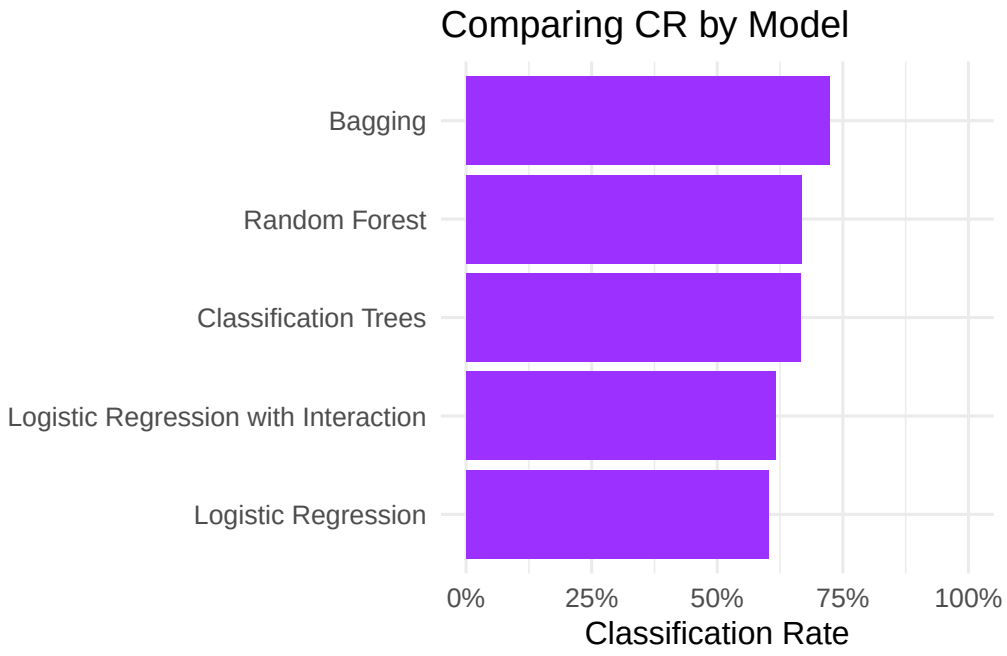
```
)


set.seed(123)

football_rf <- train(
  SUCCESSFUL ~ RP + DIST + DN + PERSONNEL,
  data = all_seasons_bagging,
  method = "ranger",
  trControl = ctrl2,
  metric = "ROC",
  tuneLength = 10
)
```

```
[1] 0.6669683
```

For our random forest model, we found a classification rate of 0.667. This means our model correctly predicted 66.7% of the actual plays.

Comparing the Four Methods

To easily compare our 5 classification rates we have found, I created the following graph:



As you can see, the far away best model we built was our bagging model. We notice that logistic regression, with or without interaction, were the worst two models. This isn't because GLM models are not as good as bagging and random forest, but instead means that the data is not best fit for a linear model. Because we have a strong interaction between RP and DIST, as well as non-linear patterns in the log-odds for success, a linear model like our logistic regression model does not best fit our data. Instead, re-sampling methods like classification trees, bagging, and random forest will have more success capturing the complexities of our data.

Conclusion

Throughout this model, we see time and again Distance be the strongest predictor of success. This makes a ton of sense as a shorter distance to go is much more obtainable than a longer distance to go. While this project is done based on a single team's data, Dist should be the strongest or one of the strongest predictors of success with whatever school you applied this research to. We also saw downs, specifically 3rd down, be one of the stronger predictors of success. 3rd downs are a pivotal part of a game, as third downs can either extend your drives and bring more chances for successful drives, or end your drives and kill the momentum. A model built using bagging had the greatest success of predicting success. With its re-sampling and simulation ability, the bagging model better fit the data strongly outperformed the others. The model correctly predicted 72.4% of the plays, which is a very successful rate. Using this model and my research, I hope to communicate with the coaches of the team about a plan for next season to incorporate our findings. As a staff, the team needs to find a way to focus on creating short down situations, especially on 3rd downs.

References

C, B. (2012, February 16). In Defense Of Success Rates. Football Study Hall. https://www.footballstudyhall.com/2012/2/16/2798555/in-defense-of-success-ra tes?utm_source=chatgpt.com