AMERICAN FOOTBALL: AN ANALYTICAL APPROACH

BY SOURCE, FAIR USE, HTTPS://EN.WIKIPEDIA.ORG/W/INDEX.PHP?CURID=17153791

5/2/2019

The Best Team: Lavada Blanton, Aaron Carmichael, Levi Taylor, and Robert Underwood

Can football teams' performance be improved by analyzing data regarding various play aspects and game conditions? Our goal in this analysis is to uncover potential insights into strategies that will result in more touchdowns and, therefore, more wins.

American Football: An Analytical Approach

THE BEST TEAM: LAVADA BLANTON, AARON CARMICHAEL, LEVI TAYLOR, AND ROBERT UNDERWOOD

TABLE OF CONTENTS

TABLE OF FIGURES

INTRODUCTION

Football, America's most watched sport (Davis), is beloved by millions of people across North America and is home to one of the most watched event every year, the Super Bowl. Thirty-two teams compete for the coveted title of Super Bowl Champions every year in order to earn fame, fortune, and glory, as they make history for their city. With 103.3 million viewers during the 2017 Super Bowl, there is opportunity for a large revenue stream with $14.1 billion spent on just watching the Super Bowl in 2017. (Reilly)

Fans attribute their favorite team's success to a wide variety of factors, from more reasonable concepts such as their star player's abilities, or the financial status of their favorite team, to more superstitious ideals, like luck, fate, or divine intervention. But what if we actually look at the numbers? What if we were to take into account every last bit of information recorded by analysts over all the details during a game? By analyzing this data and creating predictive models, we are able to see how teams can predict the most effective type of plays to call, and when to call them. This will allow a more strategic approach to coaching and helping teams win based off of data.

By determining the dependent variables for our model, touchdown and yards gained, we can segment our search down to certain values that have a high correlation to scoring allowing teams to use these to their advantage. By understanding which plays under certain conditions have the highest likelihood of scoring then this will help both offense and defense in calling the best plays in game to increase their chances of winning.

MOTIVATION

Our goal for this data analysis is to help NFL teams perform better throughout the season by implementing a strategic approach to coaching and play calling. If a team finds out that their highest chance of scoring is in the 1st quarter, on 3rd down with a pass, then the coaches and players can be reasonable and assume that calling a pass play is the best option for them to take, as it yields the highest percentage of a touchdown. This will allow coaches to have a more thorough understanding of the game better by having accurate data that is relevant and useful and allow players to make more educated judgments on the field. By having play data that shows predictive stats and implementing this data, a team would ideally have a higher chance of scoring and winning more games. By having a more analytical approach to football this can help coaches, players, and play coordinators make the best decisions for the outcome of their franchise.

Finally, by extension, more wins for a team will yield various positive results. If a team starts performing better, local fans will be more inclined to purchase tickets and buy merchandise, increasing sales revenue for the whole program. In addition to this, increasing wins can show benefits beyond that football franchise, as according to CNBC, a win for a football team can yield increases in stock returns for the company that sponsors their home stadium (Wells) as well as increased commercial airtime profits, due to the increased number of games.

DATA

Our Data consists of different variables from every recorded play from every NFL game from the years 2009-2018. This involves 4737 observations of data. Our target variables from this data set are touchdowns and yards gained. By using these as our dependent variables, we are able to see how all the other variables contribute to either increasing or decreasing the chance of a touchdown. Since the

primary goal of the game is to score, then setting this as our primary outcome can help determine if our data can predict the outcome of each play, which can change the direction of the game and overall team strategy. Our variables include the following:

Table 1: Description of Variables

Variable Name	Description	IV or DV	Variable Type
Down	Down of play. (1,2,3,4)	IV	Nominal
Posteam_score	Score of team with posession of the ball	IV	Continuous
Score_differential	Difference between the team scores	IV	Continuous
Yardline_100	Yard line starting with offensive team goal.	IV	Continuous
game_half	Half of the game. (1,2,0)	IV	Nominal
yards_gained_binary	Nominal variable of yards lost (-1), yards gained (1), or no yards (0)	IV	Nominal
Touchdown	Binary variable of either touchdown(1) or no touchdown (0).	DV	Binary
DefTeam_score	Score of defensive team	IV	Continuous
qtr	Quarter of the game (1,2,3,4)	IV	Nominal

These variables were decided first through a logical analysis of the data, by interpreting what we already knew about the data. As an example, due to the format of the data, we will never have a penalty and a touchdown on the same play. Next, we filtered out the variables that essentially meant the same things, such as yardline_100 and net yards. Next, we eliminated variables we would not need for this particular analysis, such as data about players. Finally, with the variables that were left, we initialized a linear regression to determine the significance of each variable and disregarded all variables that had a significance less than 0.05. By utilizing these variables for our prediction, we can see the major different factors in a game that can cause a higher or lower chance of scoring.

Data Preprocessing

The first step in our data preprocessing steps was cleaning the data and finding useful variables that were relevant in our search for predicting if a team will score. Our initial dataset taken from Kaggle was too large to clean manually, and had too many irrelevant variables that were unfit for modeling. We determined which variables would be most useful in predicting future plays, and selected only a few significant variables to use from the whole data set. This process took some trial and error, as we ran various models to find out which variables gave us the most accurate results. After that was complete,

we had to use file impute in SAS to get rid of all the N/A values that we had, as many columns of data had skewed values and missing data. The results of this impute are as follows:

Table 2: Summary of Class Variables

Class Variable Summary Statistics				
Role Name	Role	Missing	Mode	Mode %
TRAIN IMP_down	INPUT	0	1	48.91
TRAIN qtr	INPUT	0	2	27.6
TRAIN touchdown	TARGET	3065	0	94.12

Table 3: Descriptive Statistics

Distribution of Class Target and Segment Variables								
Variable	Mean	Standard Deviation	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
IMP_posteam_score	10.33006	9.355487	0	0	9	59	0.947577	0.733757
IMP_score_differential	-0.98635	10.83127	0	-59	0	59	0.031546	1.744129
IMP_yardline_100	48.95476	24.90329	0	1	50	99	-0.20114	-0.9276
defteam_score	11.3012	9.856854	3643	0	10	59	0.860309	0.394879
qb_hit	0.051292	0.220594	3065	0	0	1	4.068262	14.55106
yards_gained_binary	0.40579	0.61117	0	-1	0	1	-0.50704	-0.63536

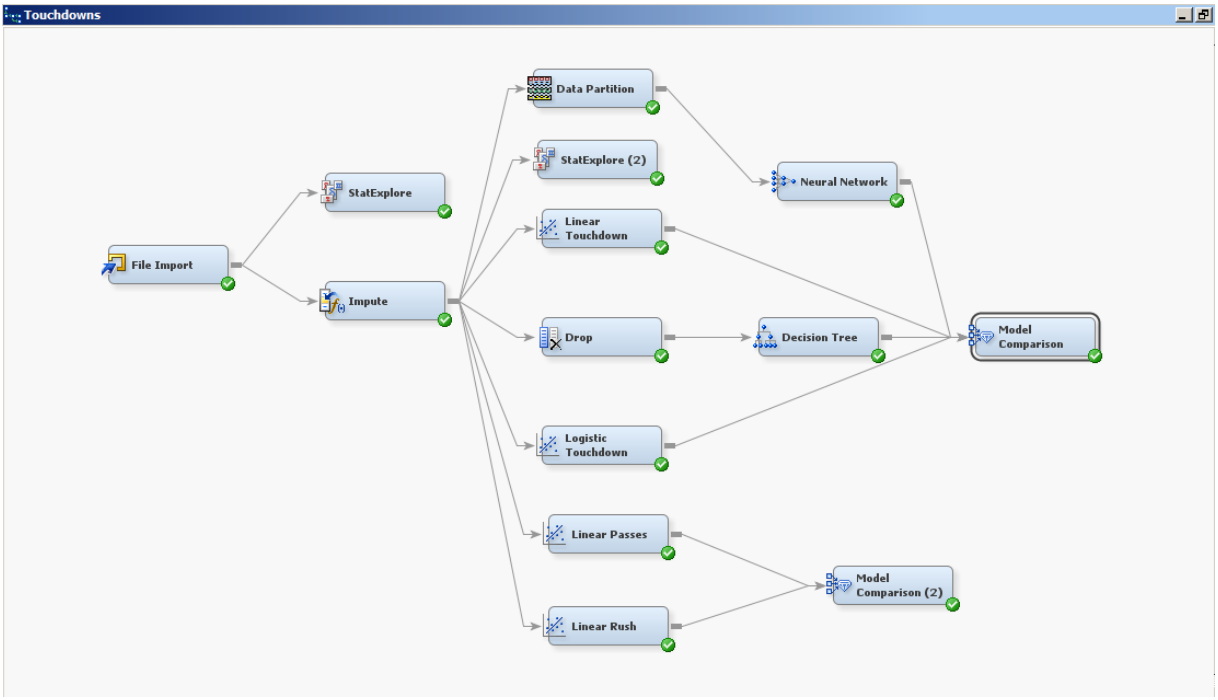
According to our sample statistics above, 94% of the data did not have a touchdown, which emphasizes the rarity of the touchdown play.

After fixing these variables, we started to notice cleaner more logical results. All these variables are significant with high standard errors that show the relevance they have to our target variable, touchdown. This allowed us to narrow our scope on specific items that we believe can help predict more accurate play calling in games. By utilizing the variables significance to the model, we were able to form an insight as to which variables had a high correlation to achieving a touchdown and which variables had little to zero value in our analysis.

MODELS

To analyze our data, we used a few different models in order to gain various perspectives into the data, in addition to having multiple sources for claims that we make. These models include linear and logistic regressions, a neural network and a decision tree.

Figure 1: Full SAS Enterprise Miner Diagram



Logistic Regression

We ran a logistic regression with the same variables as the prior regression with the binary variable touchdown being the dependent variable. As seen below, in a logistic regression, possessing team's score was not significant while the rest of our independent variables were. Also, the significance of the model was <.0001 which was very significant as compared to the average model. In our maximum likelihood estimate, we analyzed a 0.5635 ChiSq in the 3rd quarter. This would infer that the third quarter is less significant than the other quarters. According to the maximum likelihood estimate, downs 1 and 2, score differential, and yard line all have a negative correlation to touchdown. Another, inference to point out is the expected estimate of 17.35 for yards gained.

Table 4: Logistic Regression Likelihood Ratio Test

Likelihood Ratio Test for Global Null Hypothesis: BETA=0				
-2 Log Intercept Only	Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
119095.913	83284.61	35811.3029	11	<.0001

Table 5: Logistic Regression Variable Significance

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Standard Estimate	Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-3.2876	0.0689	2279.46	<.0001		0.037
IMP_down	1	1	-0.3482	0.0176	392.46	<.0001		0.706
IMP_down	2	1	-0.1611	0.0189	72.84	<.0001		0.851
IMP_down	3	1	0.2442	0.02	149.59	<.0001		1.277
IMP_posteam_score		1	0.00179	0.00173	1.07	0.3016	0.00946	1.002
IMP_score_differential		1	-0.00847	0.0012	49.48	<.0001	-0.0507	0.992
IMP_yardline_100		1	-0.0697	0.000554	15800.57	<.0001	-0.9626	0.933
game_half_	1	1	0.2275	0.0591	14.83	0.0001		1.255
game_half_	2	1	0.193	0.0566	11.63	0.0006		1.213
yards_gained_binary		1	2.8536	0.0324	7734.3	<.0001	0.9744	17.35

Decision Tree 1

Our data set contains hundreds of thousands of plays, most of which are not touchdowns. Therefore, it is important to note that in the decision tree, many of the pure nodes indicate that plays that follow that path are not touchdowns. Many initial observations follow common sense regarding football. Most obvious is that more touchdowns occur when the team is closer to their opposition's end zone. However, the difference is not as drastic as most fans would make it sound. Approximately 14.3% of touchdowns occur within 1.5 yards of the end zone, 43.7% occurring between 1.5 and 13.5 yards of the end zone, and the remaining 42% occur from beyond 13.5 yards away. Given the fact that the vast majority of plays are not touchdowns, most pure nodes result in non-touchdowns.

Table 6: Decision Tree 1 Confusion Matrix

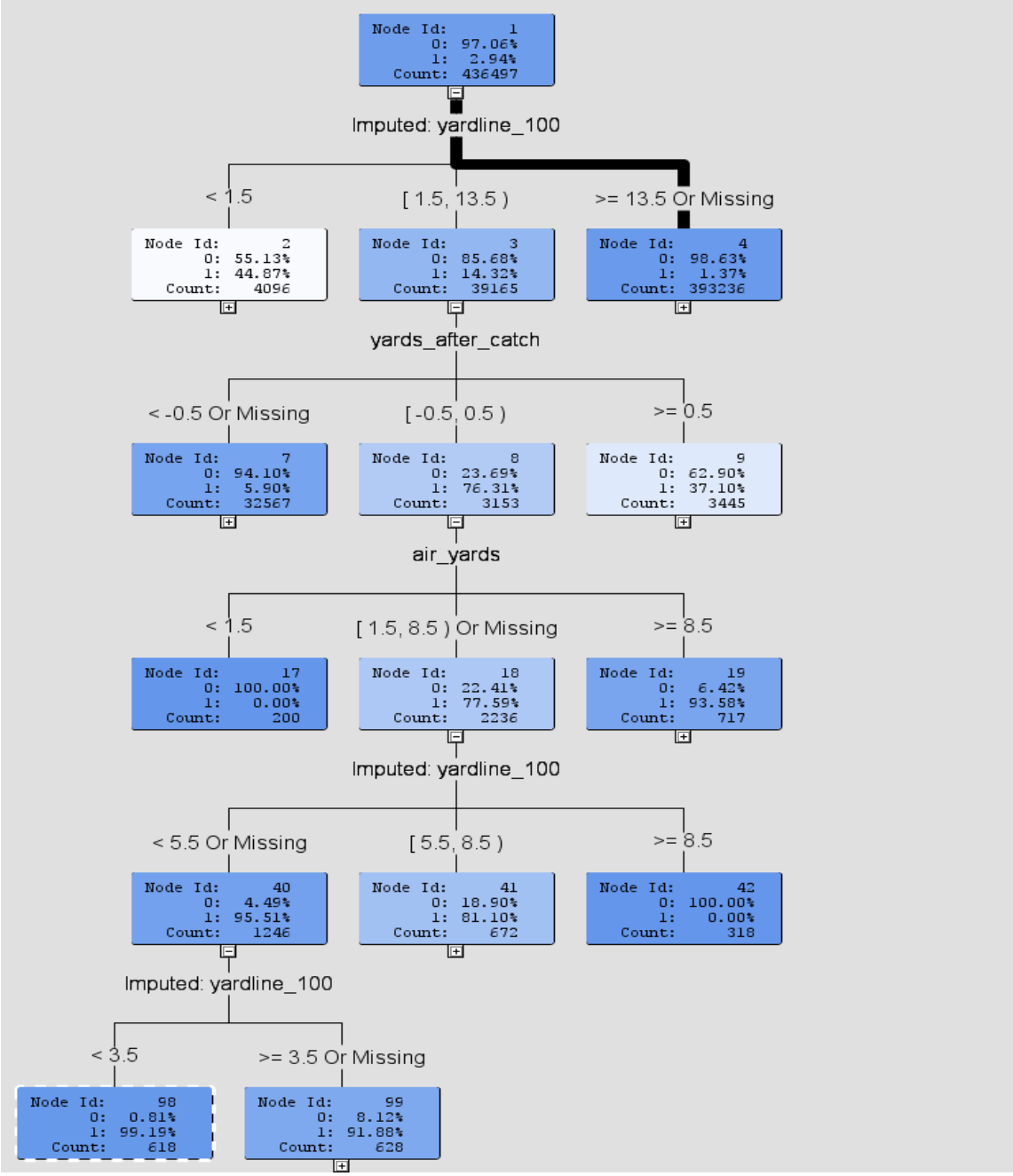
FALSE Negative	TRUE Negative	FALSE Positive	TRUE Positive
4775	435332	1206	8058

Table 7: Decision Tree 1 Confusion Matrix Analysis

Accuracy	98.67%
Sensitivity (True Positive Rate)	62.79%
Specificity (True Negative Rate)	99.72%
False Positive Rate	0.28%
False Negative Rate	37.21%
Precision	86.98%

As demonstrated in the screenshots above, the model performs fairly well. Given the astronomical amount of non-touchdowns in our data, the model is great at identifying negatives. Also given the small amount of touchdowns relative to the total number of observations, the model has a small false positive rate. In addition to this, this decision tree had a misclassification rate of 1.37%

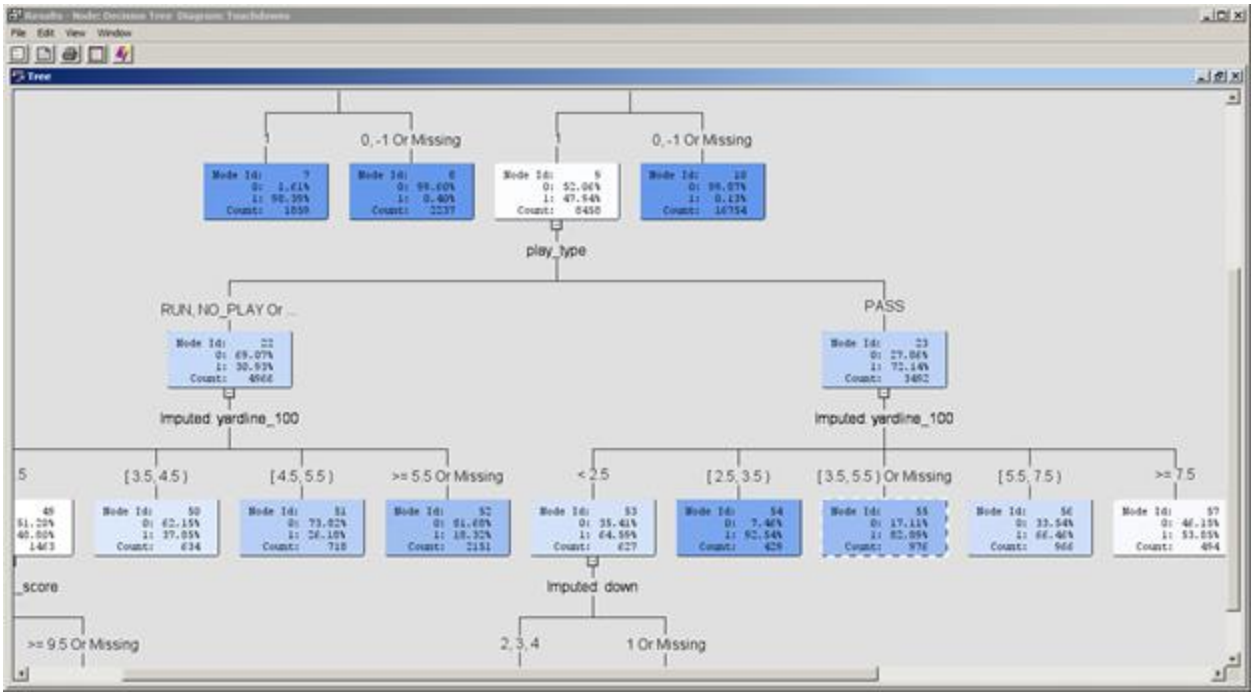
Figure 2: Decision Tree 1[1]



[1] Figure 2 only shows the path to the pure node containing the most touchdown

Decision Tree 2

Figure 3: Decision Tree 2: Standard Variables



This Decision tree gives us a model to help predict the best type of play to call depending on the yardage away from a touchdown. This shows that pass plays between 2.5-3.5 yards away have 92% of being a touchdown pass. Pass plays less than 2.5 away yards have a 64% of being a touchdown, and pass plays between 5.5-7.5 yards away have 66% of being a touchdown pass. Passes between greater than 7.5 yards away have 56% of being a touchdown pass. This shows that as the further we get from the touchdown, our probability of trying to score by pass decreases exponentially even by slight increments of yardage. Compared to run plays, these offer a higher percentage of touchdowns. Run plays do offer a high yield for scoring, given the rarity of a touchdown, but when put against pass plays do not offer the same effectiveness. By our model we can see that calling a pass play is a more effective play type than running. However, based off the various factors we cannot measure this only gives us a general approach and does not highlight special talents that can offset these results in trying to predict the most effective offensive play to call.

Passing and Rushing Linear Regressions

One question that was not answered by the previous decision trees: if the best way to score touchdowns is via a pass from less than a few yards from the end zone, what is the best way to satisfy those conditions? In other words, how are teams supposed to get that close to the end zone? These two models aim to shed some light on the answer.

Passing

Table 8: Passing Linear Regression Type 3 Analysis

Type 3 Analysis of Effects				
Effect	DF	Sum of Squares	F Value	Pr > F
IMP_down	3	0.9394	2.23	0.0826
IMP_posteam_score	1	0.0267	0.19	0.6626
IMP_score_differential	1	0.0083	0.06	0.8084
IMP_yardline_100	1	0.0607	0.43	0.5108
air_yards	1	3160141	2.25E+07	<.0001
defteam_timeouts_remaining	1	0.0184	0.13	0.7174
game_half_	2	0.041	0.15	0.8642
game_seconds_remaining	1	0.08	0.57	0.4503
pass_length	2	1.8365	6.54	0.0014
pass_location	2	0.0643	0.23	0.7954
posteam_timeouts_remaining	1	0.286	2.04	0.1536
qb_hit	1	0.006	0.04	0.8369
yards_after_catch	1	5361145	3.82E+07	<.0001
ydstogo	1	0.3645	2.6	0.1072

Table 9: Passing Linear Regression Model Fit Statistics

Model Fit Statistics			
R-Square	0.9987	Adj R-Sq	0.9987
AIC	-213725	BIC	-213723
SBC	-213533	C(p)	20

Table 10: Passing Linear Regression Analysis of Variance

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	11568475	608867	4334348	<.0001
Error	108882	15295	0.140475		
Corrected Total	108901	11583771			

Rushing

Table 11: Rushing Linear Regression Type 3 Analysis

Type 3 Analysis of Effects				
Effect	DF	Sum of Squares	F Value	Pr > F
IMP_down	3	654555.905	3644.56	<.0001
IMP_posteam_score	1	0.0436	0	0.9785
IMP_score_differential	1	2092.3326	34.95	<.0001
IMP_yardline_100	1	172807.476	2886.57	<.0001
defteam_timeouts_remaining	1	17138.0571	286.27	<.0001
game_half_	2	1.5461	0.01	0.9872
game_seconds_remaining	1	21.7619	0.36	0.5466
posteam_timeouts_remaining	1	1787.4977	29.86	<.0001
qtr	2	749.333	6.26	0.0019
ydstogo	1	408933.843	6830.81	<.0001

Table 12: Rushing Linear Regression Model Fit Statistics

Model Fit Statistics			
R-Square	0.0577	Adj R-Sq	0.0577
AIC	1785222	BIC	1785224
SBC	1785387	C(p)	15

Table 13: Rushing Linear Regression Analysis of Variance

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	1599968	114283	1908.98	<.0001
Error	436241	26116028	59.866055		
Corrected Total	436255	27715995			

Interpretation

The first thing to note is that the models are significant, with p-values of <.0001. However, the differing R-Squared and Adjusted R-Squared values indicate a strong difference between the two. The

high R-Squared value in the passing model implies a strong correlation between the target variable (yards gained) and the independent variables. On the other hand, the low R-Squared value in the rushing model implies a very weak correlation between the target and independent variables. Based on p-values, we can determine that the following variables are significant in regards to passing:

- Air_yards
- Pass_length
- Yards_after_catch

To football fans, there is nothing particularly shocking about these variables being significant. However, on the other hand, it rules out the significance and correlation for other variables. There is no strong correlation between gaining yards via passing and game conditions such as score, time left, or down. In addition, the location of the pass does not have a strong effect on yards gained.

The following variables are significant in the rushing model:

- IMP_down
- IMP_score_differential
- IMP_yardline_100
- Defteam_timeouts_remaining
- Posteam_timeouts_remaining
- Qtr
- Ydstogo

As compared to the passing model, it seems that various game conditions have a correlation with yards gained via rushing. However, with the low R-Squared value, all of these variables only account for a very small portion of yards gained.

Neural Networks

After all this analysis, we initialized a neural network of the specific variables that were already classified as significant in the earlier models and omitted, possessing team score. As shown below, the lift did not differ from the baseline much at all. This gives the inference that the results of this dataset are uncertain and more instances of touchdown are needed. Despite our picking of variables and a misclassification rate of 0.02, the data still does not fit a comprehensible pattern that differs from a guess.

Figure 4: Neural Network Lift Analysis

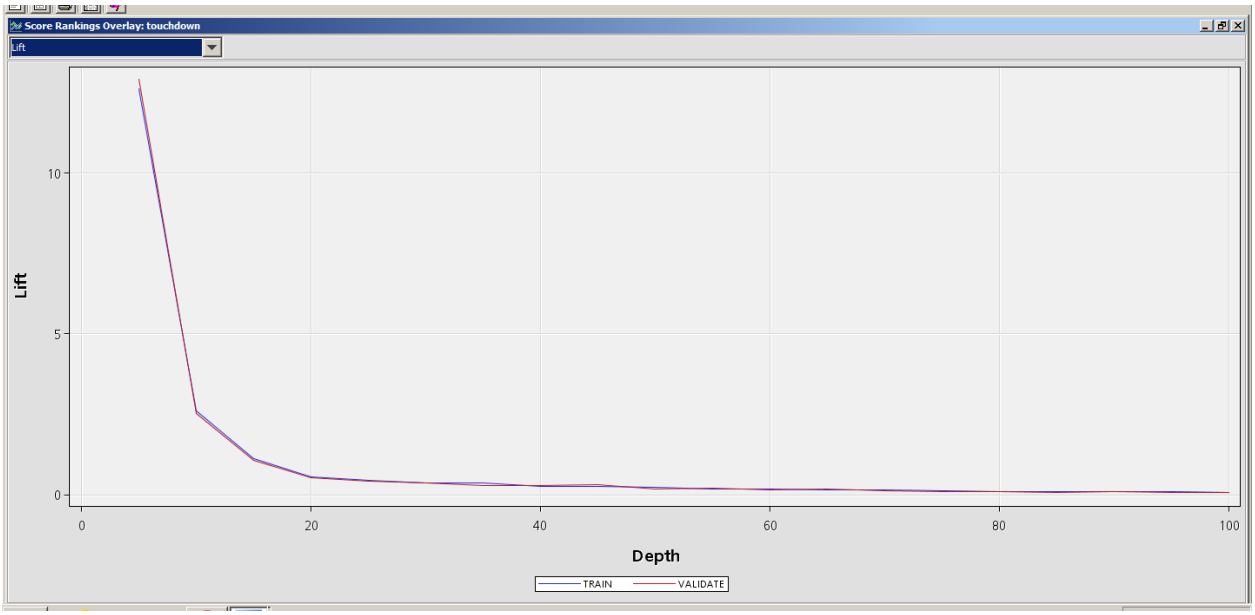


Table 14: Neural Network Classification Tables

Classification Table					
Data Role=Train Target Variable= touchdown Targe Label=touchdown					
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
.	0	2.9076	100	5149	2.8646
0	0	95.2278	99.512	168638	93.8196
1	0	1.8646	64.329	3302	1.837
0	1	31.1136	0.488	827	0.4601
1	1	68.8864	35.671	1831	1.0187
Classification Table					
Data Role=Validate Target Variable= touchdown Targe Label=touchdown					
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
.	0	2.908	100	3862	2.8648
0	0	95.2322	99.509	126475	93.8165
1	0	1.8598	64.156	2470	1.8322
0	1	31.1377	0.491	624	0.4629
1	1	68.8623	35.844	1380	1.0237

Table 15: Neural Network Event Classification Tables

Event Classification Table			
Data Role= Train Target=touchdown Target Label=touchdown			
FALSE Negative	TRUE Negative	FALSE Negative	TRUE Positive
3302	173787	827	1831
Event Classification Table			
Data Role= Validate Target=touchdown Target Label=touchdown			
FALSE Negative	TRUE Negative	FALSE Negative	TRUE Positive
2470	130337	624	1380
	Train	Validation	Test
Misclassification Rate	0.02	0.02	0.02

FINDINGS

Our initial motive for analyzing this data was to try and predict the best plays to call that have the highest chance of being a touchdown and examine all the different variables that affect the outcome. From the data shown by our various models, we attempted to determine predictive results for where scoring a touchdown is more likely, and the best plays to call to achieve a touchdown. We did find correlations and results, however, touchdowns are not a simple statistical measure that can be intricately

probed, as they are unexplainable and based off too many chance based measures. Our models reflected that there were some statistical proof that some plays are more likely to result in a touchdown. However, the touchdowns were still dispersed across an array of different conditions. In an effort to find predictable patterns within the data, we came to the conclusion that touchdowns are too rare to make any significant predictive measure. With a vast amount of irrational variables taken into account, producing useful and accurate predictions of numerical data to show past history is difficult in determining predictive analytics. This shows us that even with a large dataset from years of plays and a varied range of variables, we cannot accurately predict when a team will score. Despite the significance of our models and variables, our models' ability to predict the outcome still faltered. This leaves many issues up for debate, as NFL owners need to determine where to invest their money since trying to spend more on analyzing and predicting plays is not a feasible investment. Instead of investing in play by play predictive analytics, NFL teams need to spend more money into drafting better players and staff to help their organization move forward. With so much of the game residing on the players themselves, it is simply not enough to try and predict the best play if the opposing team has more skill and talent.

CONCLUSION

With the percentages of touchdowns being only a miniscule amount of our data, it is best that NFL teams do not treat football as purely numerical strategy, as play results are highly unpredictable. From our data, we can analyze the best types of plays calls that historically have had the highest percentage of touchdowns, but football has too many human variables affecting the outcome. If football was entirely formulaic and predictable, nobody would watch it; the human factor is what makes it exciting and appealing. Instead of spending money on trying to uncover analytics about games, despite the vast resources of statistical data, revenue should be spent elsewhere to improve the analysis of players, team morale, and ticket sales.

REFERENCES

Davis, Scott. “TV Viewing Habits Shows How the NFL Is Still as Dominant as Ever.” *Business Insider*, Business Insider, 5 Jan. 2019, www.businessinsider.com/nfl-ratings-most-watched-sports-events-2018-2019-1

Reilly, Lucas. “By the Numbers: How Americans Spend (More of) Their Money.” *MentalFloss.com* 11 September 2017. http://mentalfloss.com/article/94623/numbers-how-americans-spend-more-their-money.

Wells, Nick. “How an NFL team's record can affect its sponsor's stocks.” *CNBC* 10 September 2015. https://www.cnbc.com/2015/09/10/how-an-nfl-teams-record-can-affect-its-sponsors-stocks.html.