Somebody’s Gotta Score: A Predictive Analysis of NFL Touchdowns

Justin Hooker
CSCI 4360
University of Georgia
Athens, GA 30602
Jwh40896@uga.edu

Abstract

This paper attempts to dispel the notion that NFL touchdowns are highly volatile and unpredictable events. Motivated by a fantasy football perspective, opportunity and efficiency are used as the defining criteria to develop a model that seeks to project touchdowns per player on a game-by-game basis. Weekly receiving data from the 2016-2017 NFL season was collected for all skill positions (WR, TE, RB). Exploring the dataset presented several challenges, such as scoring events appearing in less than twenty percent of the sample. This skewness had a clear effect on model performance, as accuracy of over eighty percent was achieved for several classification algorithms due to the high rate of true negatives. However, F1 scores for the associated models were much worse, generating results just above 0.30. Consequently, the results of this analysis do not support the proposed theory that opportunity and efficiency can be modeled to project touchdowns.

Background

1.1 Daily Fantasy Sports

The Fantasy Sports Trade Association estimates the player base for fantasy sports at nearly 60 million in North America for 2017, representing an increase of ~3.5% from the previous year. Furthermore, the fantasy sports industry has an estimated value of over $7 billion dollars with significant growth prospects expected for the future. The two market leaders DraftKings and Fanduel were some of the first to capitalize on the current trend towards a new format known as daily fantasy sports. Gone are the days of season-long commitments to leagues playing the same players drafted from months before.

Daily fantasy provides the convenience of daily commitments to contests while also allowing for new game types with massive prize pools. Draftkings hosts their weekly Millionaire Maker contest each Sunday during the NFL season with a prize pool of over $5 million dollars and the winner taking home $1 million dollars. Fanduel has claimed to have payed out over $500 million dollars in contest winnings annually since 2015. The opportunity to win potentially life changing money adds a new, interesting layer to the sports fan’s perspective while watching the games; not only do they have a rooting interest in their team, they now have a personal, rooting interest in player performance.

1.2 Motivation

The daily fantasy industry has seen tremendous growth in recent years, but two concerns need to be addressed to maintain its upward trajectory: is it a game of skill, and then how does this affect the legal status of the industry? Currently, the industry sits in a grey area in terms of legal status, as it's not entirely clear to those with influence whether daily fantasy sports should be considered gambling or not. The goal of this analysis is to gain a better understanding of some of the underlying factors that contribute to the skill component of daily fantasy sports – projecting player performance.

There undoubtedly exists a random element that affects all sport outcomes, but this isn't to say the result of the games is entirely uncertain. For example, when evaluating the win probability for two competing teams, one might consider several factors such as: overall record, head-to-head record, injuries, location, and matchup criteria. This information can have an enormous impact on projecting a team's chances of winning a particular game. Daily fantasy makes similar projections but on a micro level. These projections seek to maximize player performance subject to a salary budget and positional requirements for roster construction. Additionally, there is a scoring system to consider which awards points to fantasy rosters for various outcomes on a per-play basis.

Table 1 – NFL Skill Position Fantasy Contribution

Opportunity	Allocated Points
Rushing Yard	0.1 pts
Rushing Touchdown	6 pts
Reception	0.5 pts
Receiving Yard	0.1 pts
Receiving Touchdown	6 pts

Table 1 provides the Fanduel per-play, fantasy contribution for skill position players in the NFL. Scrimmage yards accrue 0.1 fantasy points per yard, pass receptions reward 0.5 fantasy points each, and touchdowns garner 6 fantasy points for skill players. The scoring rubric indicates that fantasy performance for NFL players should be highly dependent on touchdowns, given the disproportionate weight assigned to these events. For further context, consider that the average fantasy points accrued through the air by skill players for last season was 7.51. However, this average more than doubles to 16.50 when filtered for a scoring event and reduces to 5.31 for non-scoring events. As a result, there should be a significant advantage available to developing a model which accounts for this variation and can project scoring probabilities with a reasonable accuracy.

1.3 Relevant Work

Current wisdom from within the fantasy sports community suggests touchdowns are the most volatile event in sports, followed by predicting a homerun. This would seem to be a reasonable viewpoint with so many attributable factors to consider such as: matchup, opportunity, efficiency, and scheme. Furthermore, NFL touchdown opportunities through the air are surprisingly lower than expected for skill players. It turns out that only the best players receive more than twenty opportunities per season to score inside the red zone – inside the opponents twenty-yard line. It is also worth noting that the touchdown probability for an average skill player last season was ~18%, all else equal.

This idea of scarce opportunity, along with other potential factors, appears to have discouraged any meaningful effort from the community at understanding this perceived variation in scoring outcomes. As a result, very few references were found during the preliminary research for this analysis. One simple method that was found for predicting NFL touchdowns was a "coin-flipping" model, where essentially each player's scoring probability was fixed as his prior touchdown rate per game. Results for the model were not published but were likely very poor given the model doesn't account for game specific criteria as previously mentioned. Another method analyzed the effects of opportunity on touchdown expectation for season-over-season comparison. The author did find a correlation between opportunity and touchdown rates, but since the data was collected on a per-season basis, results were found exploratorily rather than a model driven approach. The final method found was another exploratory analysis into past season touchdown rates and some potential factors that might influence results.

A look into the relevant work of NFL touchdowns suggests a relatively undeveloped area with a lot of room for growth. This analysis will attempt to project weekly receiving touchdowns for skill players using game data from the 2016-2017 NFL season. The model will emphasize opportunity and efficiency as the basis for its projections. Rushing production will not be accounted for in the model due to the lack of opportunities for wide receiver and tight end. This allows for a simpler process but does discount the value of the running back position.

Data Collection

1.1 Sources

The source for the receiving production and efficiency data is *http://www.profootballfocus.com*. A utility feature allowed for the filtering of weekly production and associated metrics which was then scraped into a csv file. The source for the team data is *http://www.pro-football-reference.com*. The team data is comprised of matchup criteria such as offensive/defensive rankings and pace of play. This data was exportable from the website as a csv file. Any other relevant features were engineered from the dataset as part of the exploratory process.

1.2 Feature Set

Table 2 provides a list of the features used as part of the modeling process. The label for the dataset is the binary variable "Score", where a zero represents a non-scoring outcome and a one represents a scoring outcome.

Table 2

Feature	Description
Score	Binary variable indicating an observed scoring event.
Team	Ordinal value representing the offensive ranking of a player's team.
Opp	Ordinal value representing the defensive ranking of the opposing team.
Pos	The designated position of the offensive, skill player (WR, TE, RB).
Tgt	Target. A reception opportunity.
Rec	Reception. A converted reception opportunity.
Yds	Receiving yards. Scrimmage yards accrued during a game.
C%	Catch Percentage. The ratio of Receptions/Targets.
aDOT	Average depth of target. Measured distance of reception opportunities.
YPR	Yards per reception. Yards gained relative to converted opportunities.
YPT	Yards per target. Yards gained relative to total opportunities.
YAC	Yards after catch. Yards gained after the reception.
QBR	Quarterback rating. Efficiency metric used to evaluate qb performance.
SK	Sacks. Strong indicator of defensive strength.

Data Processing

1.1 Cleaning

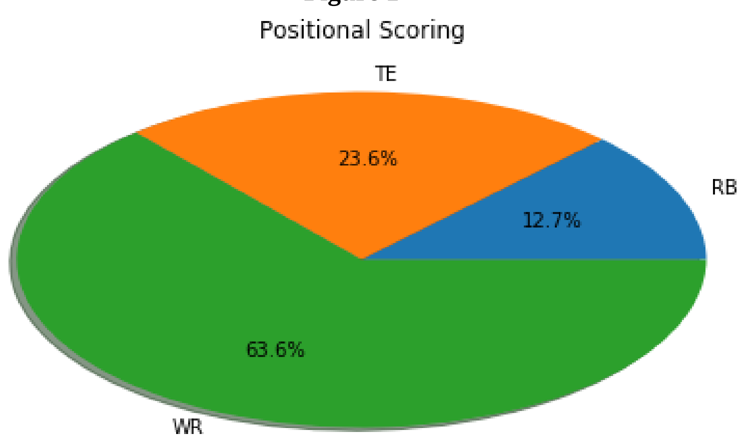
The cleaning process for the NFL dataset that was compiled for this analysis was tedious, as expected, but not terribly difficult. First, the production data was scraped by player position and needed to be conglomerated. Some of the associated metrics for production contained hyphens instead of zeros, which needed to be replaced. These columns also needed to be converted from string type to float as well. Finally, the matchup criteria needed to be incorporated into the dataset. This was done using panda's merge function.

1.2 Exploration

The NFL 2016-2017 dataset was filtered for skill position players who received at least one target during a game, which was necessary as a player with zero opportunities to catch a pass would not be able to score a touchdown. As a result, total observations for the dataset were 3645.

It is worth noting that there appears to be a positional component to fantasy production. The average fantasy points by position from last season are: 9.74 for wide receiver, 7.21 for tight end, and 5.04 for running back, through the air. A total of 753 receiving touchdowns were scored last year. Figure 1 shows the scoring frequency by position. The wide receiver position accounted for nearly two-thirds of the receiving touchdowns last season, which follows intuitively given the fantasy production. This would seem to indicate a positional component in relation to scoring touchdowns.

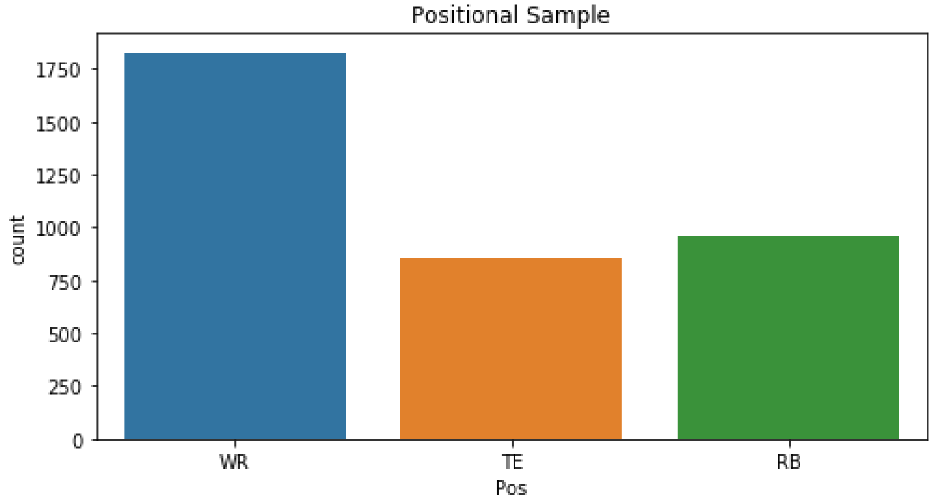
Figure 1



Formatting Source: https://www.kaggle.com/ash316/eda-to-prediction-dietanic

Figure 2 shows the positional frequency represented in the dataset. Wide receiver accounts for 1826 observations, which is about half of the dataset while scoring two-thirds of the touchdowns. Tight end and running back almost split the remaining observations with the tight end position doing well in maintaining its scoring frequency relative to its sample frequency. Could the touchdown success by the wide receiver position be explained simply by its relative frequency in the dataset? This does not appear to be true. Figure 3 provides the scoring efficiency by position.

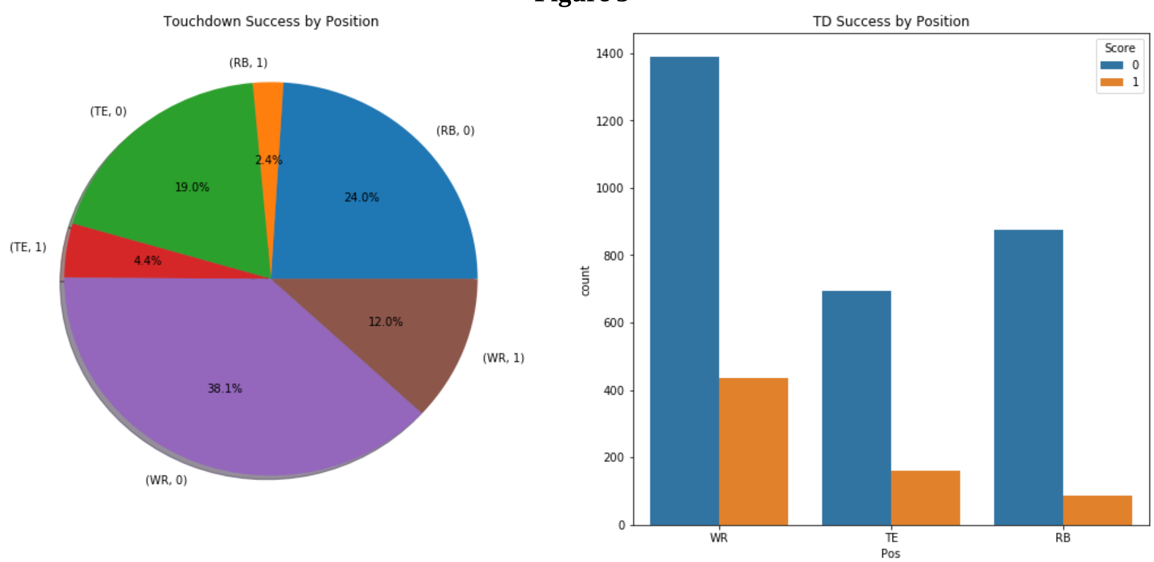
Figure 2



Formatting Source: https://www.kaggle.com/ash316/eda-to-prediction-dietanic

Figure 3 represents the positional ratio of touchdown success observed in the dataset. Wide receiver appears to also be the most efficient in terms of touchdown conversion with a rate of 485/1341 ~30%. Tight end and running back convert touchdowns at a much lower rate of about 18% and 10%, respectively.

Figure 3

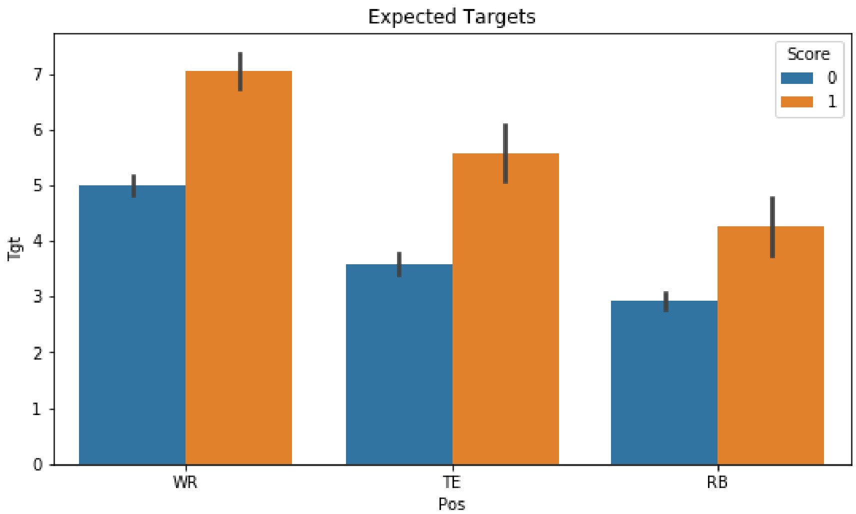


Formatting Source: https://www.kaggle.com/ash316/eda-to-prediction-dietanic

It seems that wide receiver is an important skill position in relation to fantasy production and touchdown conversion. To understand why the wide receiver position stands out as so significant, it is necessary to explore the production and efficiency features from the dataset. Is it the case that wide receivers get more opportunities? Figure 4 represents the positional breakdown of expected opportunities filtered by touchdown success.

Interestingly, there is a significant increase in expected opportunities across all positions when a player scores a touchdown. This appears to be a significant feature. The average opportunities per game for a player increases from 4.05 to 6.34 when he scores a touchdown in the game.

Figure 4

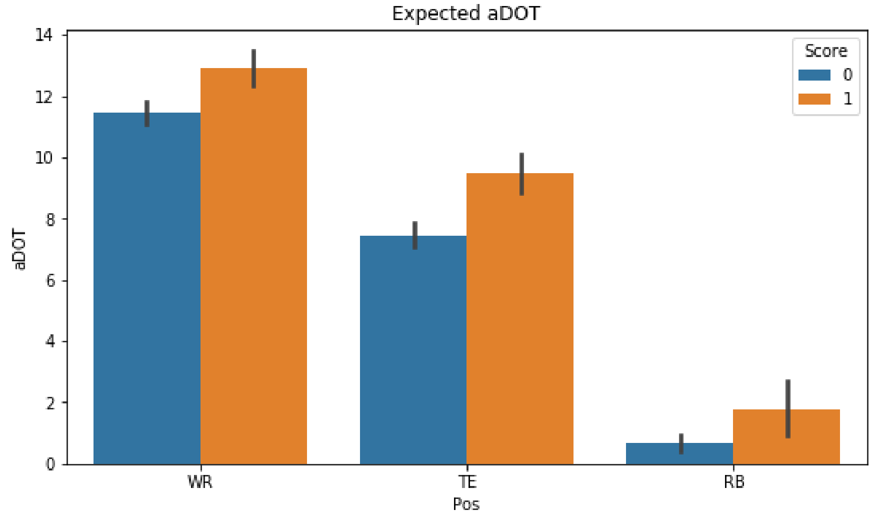


Formatting Source: https://www.kaggle.com/ash316/eda-to-prediction-dietanic

One measure of production is receiving yards gained during a game. Figure 5 shows the positional breakdown of expected receiving yards across all positions filtered by touchdown success. There is also a significant increase in production across positions when a player scores a touchdown. A player's average receiving yards per game increases from 28.76 to 60.37 when that player scores a touchdown.

Figure 6 displays the positional breakdown of the aDOT feature, which incorporates a distance component to each skill player's opportunities. It should be the case that if a player has more opportunities down the field, then he is more likely to produce big plays and score touchdowns. Wide receiver averages the most distance per opportunity with a significant increase when the player scores a touchdown. Tight end also sees an increase in distance when the player scores a touchdown. Running back averages a miniscule 0.65 yards per opportunity without a significant increase in distance for an observed touchdown. The distance component would seem to provide context for the scoring efficiency of each skill position.

Figure 6

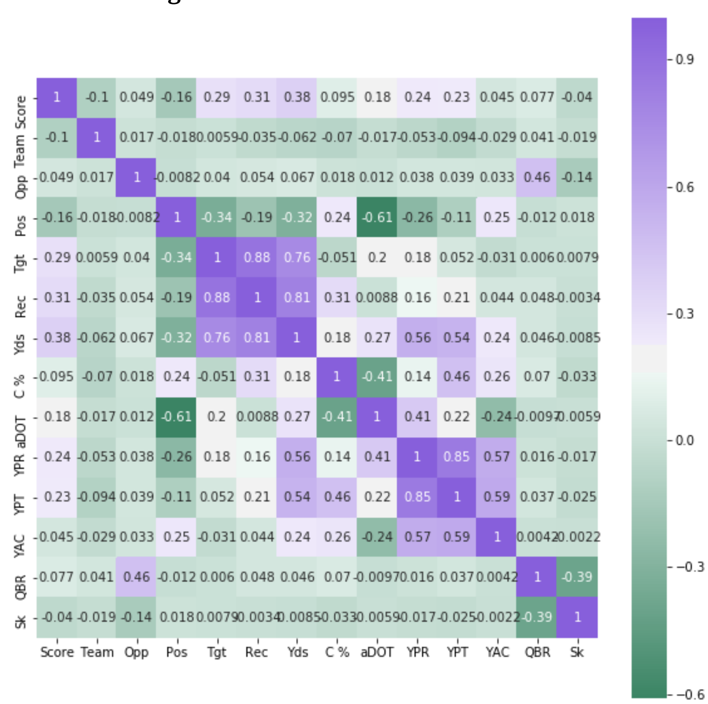


Formatting Source: https://www.kaggle.com/ash316/eda-to-prediction-dietanic

1.3 Dimensionality Reduction

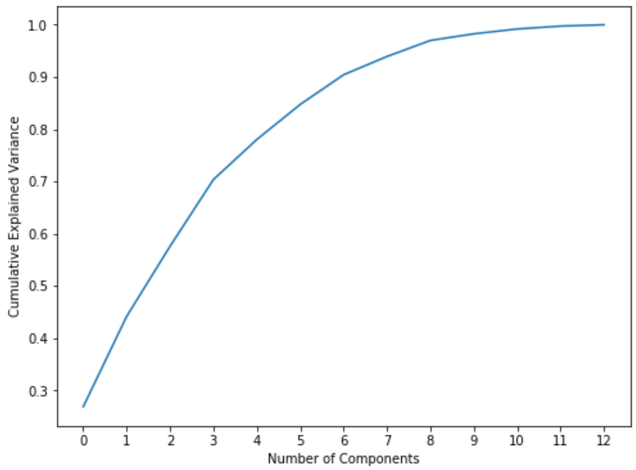
Another challenge that needed to be addressed was the high degree of collinearity between features in the dataset. It was expected that touchdowns and fantasy points would strongly correlate with the scoring label and these features were already going to be dropped before the modeling process. However, it was not expected to observe the high degrees of multi-collinearity between the production features. Figure 7 provides the correlation matrix for the feature set.

Figure 7 – Correlation Matrix



Production measures such as: targets, yards, and receptions all show a large amount of correlation with each other. Other efficiency metrics were highly correlated as well such as yards per target and yards per reception. Too many relevant features were correlated with each other to considering dropping them, so dimensionality reduction was employed to address this issue. Classic Principle Components Analysis was used to shrink the dimensionality of the dataset while maintaining a linear combination of the features. Figure 8 shows the cumulative variance obtained for each additional component. 97% of the variance was retained using nine components.

Figure 8 – Dimensionality Reduction



Methods

1.1 Evaluation Criteria

Accuracy and the F1 score were used to evaluate results during the model selection process. Accuracy produced deceivingly strong results given the skewed distribution of scoring outcomes where an observed touchdown accounted for ~18% of samples. This led to a high rate of negative outcomes which produced high accuracy but low recall for several algorithms. As a result, the F1 score was used as the primary criteria for evaluating performance. Additionally, incorporating a baseline for comparison of results was more difficult than anticipated. Several options were considered such as a fixed negative prediction where zero is always predicted and then a positional threshold where a positive prediction was enforced for wide receiver. However, these options weren't practical in application to projections of player performance. It was decided that a threshold based on opportunity would be used. The average targets for a player who scored a touchdown was about six, so anything greater than or equal to six produced a positive prediction and less than six produced a negative prediction.

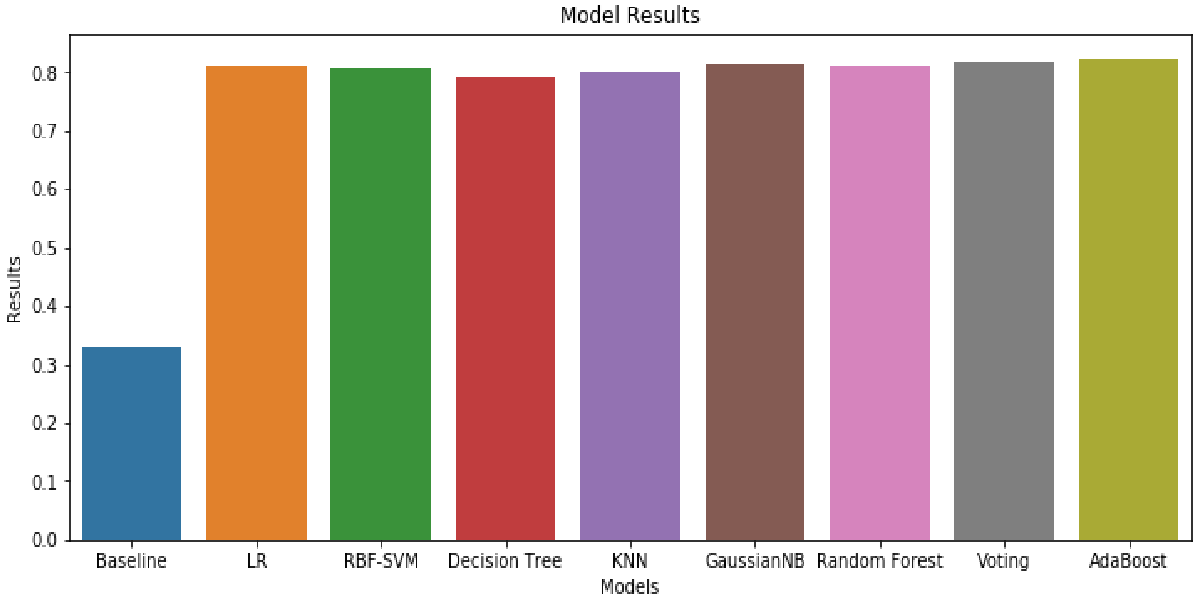
1.2 Model Selection

The classification algorithms tested in this analysis include: SVM, Logistic Regression, KNN, Decision Trees, Naïve Bayes, Random Forest, and Stochastic Gradient Descent. SVM was tuned using the 'rbf' and linear kernels, as well as the gamma and C parameters. The Decision Tree algorithm was tuned to reduce the depth parameter to account for potential overfitting to the training set. The KNN algorithm produced the best cross validated accuracy with the "n_neighbors" parameter equal to ten. Two ensemble algorithms were also tested. The first ensemble was a voting method which effectively takes an average of the independent results and the other ensemble was a boosting method known as Adaboost.

Results

1.1 Accuracy

Figure 9

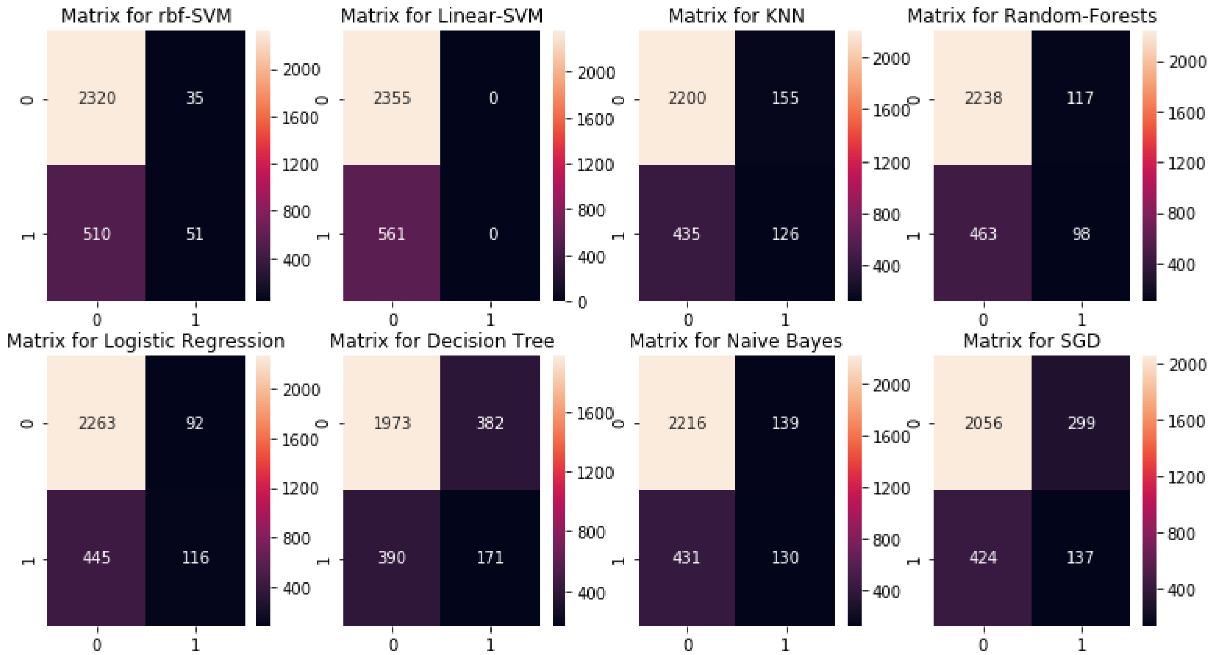


Formatting Source: https://www.kaggle.com/ash316/eda-to-prediction-dietanic

Figure 9 shows the accuracy results for each classification method that was tested. The results show very little deviation with each algorithm correctly classifying scoring outcomes with an accuracy of ~80%. The ensemble boosting method Adaboost slightly outperformed the others with a classification accuracy of 82%. This reflects poorly, however, on the skewness of the dataset. Due to the high rate of negative outcomes, these algorithms could be producing high accuracy while misclassifying touchdown events.

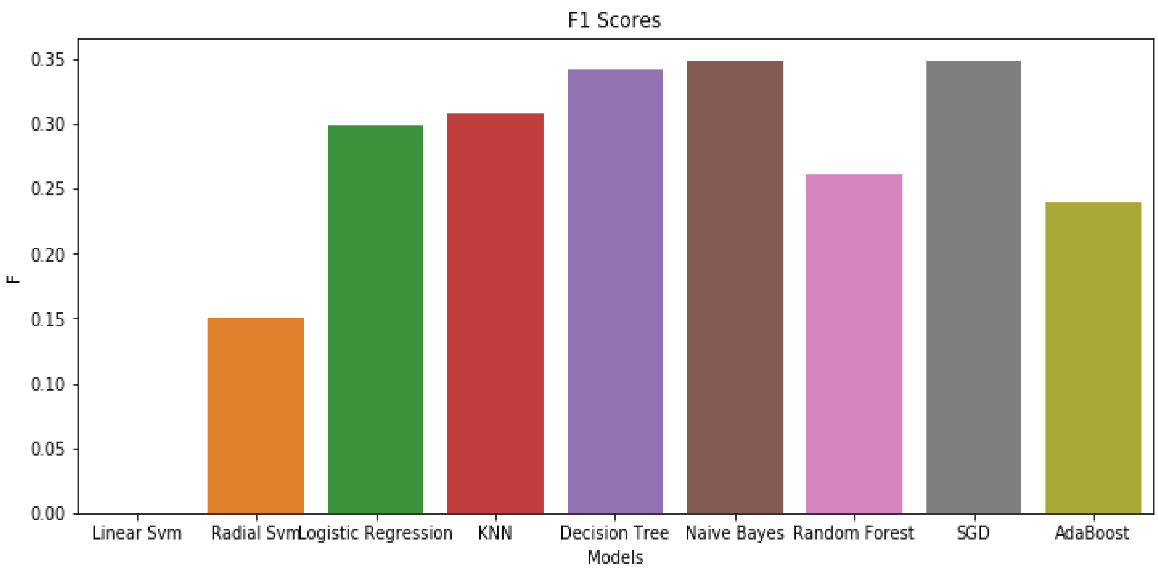
Figure 10 produces a confusion matrix for each tested algorithm. The algorithm which had the best recall and the most correctly classified touchdown events was the Decision Tree. However, the Decision Tree algorithm also had the highest rate of false positives. Interestingly, the Naïve Bayes classifier had the second highest correctly classified touchdown events with a much lower false positive rate. Linear-SVM predicted zero positive outcomes which was an interesting result as well. It is interesting to consider if this due to the PCA transformation.

Figure 10 – Confusion Matrix



Formatting Source: https://www.kaggle.com/ash316/eda-to-prediction-dietanic

Figure 11



Formatting Source: https://www.kaggle.com/ash316/eda-to-prediction-dietanic

Figure 11 shows the results of the F-scores for the tested models. Three algorithms seem to have outperformed the rest which are Decision Trees, Naïve Bayes, and Stochastic Gradient Descent. Each produced an F-score of ~0.34. This result is surprisingly close to the comparative baseline of 0.33 but clearly not a significant improvement. Linear SVM produced an F-score of zero due to its 100% negative prediction rate and was the clear outlier of the results.

Conclusion

It is interesting to note that there was considerably more variation in results from the F-scores than from the accuracy metric. This could indicate that additional tuning of parameters might improve recall performance. Current results were not particularly impressive with the best f-score of 0.34 representing a slight improvement over the baseline. Exploration of the data did reveal a positive correlation between opportunity and efficiency with touchdown outcomes, but it appears that the modeling techniques used in this analysis fell short of capturing this relationship.

Looking forward, the next step will be to train and implement a neural network for a comparison of results since there appears to be additional information that is not being captured by the traditional classification algorithms. It is crucial that model predictions correctly identify touchdown outcomes and current results do not support this notion. Therefore, the proposed theory that opportunity and efficiency can be modeled to project touchdowns cannot be validated at this time.