# *PREDICTING FANTASY FOOTBALL OUTCOMES*

*Daniel Scholnick, David Brewer, Peter Mathews, Craig Beach*

## Introduction

Americans love professional sports, and of the available professional sports, the National Football League (NFL) is king. The National Football League is an organization that manages professional American football operations. American football can be best described as a homegrown mixture of rugby and soccer. The sport is uniquely American. Competition, regional identity, aggression, creative thinking – the sport not only embraces American values, but it has woven these values into its very fabric. While American football was invented on college grounds in 1919, Americans have found new ways over the years to further fall in love with the sport.

Enter: fantasy football. Fantasy football has contributed to a dramatic rise in popularity of the American football over the previous two decades. The rise of fantasy football is directly attributable to the increasing widespread use of technology. Football data became available to the consumer, consumers had increasingly better means access to that data, and the past time skyrocketed in popularity. The massive popularity of the sport means there is massive economic opportunity and bragging rights on the line.

According to Bleacher Report, approximately 60 million people participate in fantasy football each year and spend an average of $595 on fantasy football per season. This equates to one out of every five Americans participating in the phenomena. Data science and fantasy football have clear synergies – both are data driven domains that intend to understand or improve outcomes. Data science has many robust and highly interpretable machine learning algorithms that are freely available to the public. Given these factors, the team's goals are to build models that predict fantasy football outcomes and identify the associated drivers of these outcomes.

## Analysis

### About the Data

The lack of publicly available National Football League (NFL) data sources has been a major obstacle in the creation of modern, reproducible research in football analytics. While clean play-by-play data is available via open-source software packages in other sports (e.g. nhlscrapr for hockey; PitchF/x data in baseball; the Basketball Reference for basketball), the equivalent datasets are not freely available for researchers interested in the statistical analysis of the NFL. This analysis uses the ArmchairAnalytics.com dataset, which is considered one of the premier resources for American Football statistical information.

The Armchair Analytics dataset consists of 30 tables and over 7000 data points, spanning from 2000-2018. Of the thirty tables, data from five of the tables were extracted, transformed and combined to create a dataset for further analysis. The five selected tables centered around the 2018 Play-by-Play data and the player's availability. In addition, tiers for coaching and defense were imputed.

| uid | gid | wk | day | v | h | stad | temp | humd | wspd | cond | surf | games_played | team |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 99867 | 4791 | 1 | THU | ATL | PHI | Lincoln Financial Field | 81 | 71 | 8 | Cloudy | DD GrassMaster | 1 | ATL |
| 99868 | 4791 | 1 | THU | ATL | PHI | Lincoln Financial Field | 81 | 71 | 8 | Cloudy | DD GrassMaster | 1 | ATL |
| 99869 | 4791 | 1 | THU | ATL | PHI | Lincoln Financial Field | 81 | 71 | 8 | Cloudy | DD GrassMaster | 1 | ATL |
| 99870 | 4791 | 1 | THU | ATL | PHI | Lincoln Financial Field | 81 | 71 | 8 | Cloudy | DD GrassMaster | 1 | ATL |
| 99871 | 4791 | 1 | THU | ATL | PHI | Lincoln Financial | 81 | 71 | 8 | Cloudy | DD GrassMaster | 1 | ATL |

*Fig 1. Raw Data Analysis*

**Preprocessing**

As highlighted below, very few fields had NULL values. Still, there were a number of steps required to take the initial dataset and condense it down to the key variables required to predict a player's points.

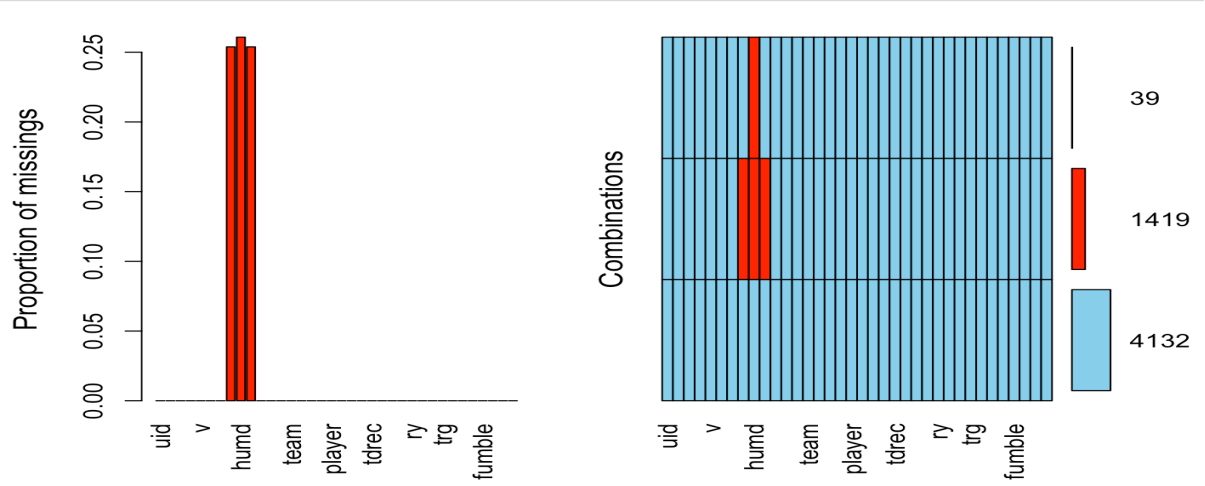


*Fig 2. Key Variables*

**Key Steps:**

1. Combine the key fields from Armchair Analysis csv dataset tables.
2. Perform EDA to determine if there are any issues with the data

3. Calculate the player's total fantasy points and compare with Armchair Analysis's fantasy points from NFL.com
4. Create user-defined variable metrics
5. Run Associate Rules Mining and R's Boruta package and compare results for final feature selection

| Variable | Definition |
|---|---|
| Scoring Tier | bin/ranking of scoring |
| Player Market Share | Percentage of Touches per game |
| Player Availability | Based upon Injury Data |
| Coaching Rank | bin/ranking of Coach |
| Depth Chart | A player on 1st or 2nd team |
| Pass Attempts per game | Average Pass Attempts |
| Passes Completions per game | Average Pass Completions |
| Average PTDs | Average PTD (passing) Touchdowns |
| Average RTDs | Average TD (rushing + receiving) Touchdowns |
| Average rush/receiving yards per game | Average yards (rush + receiving) |
| Average Turnovers | Average Fumbles + Interceptions |
| Average Team First Downs per game | Average Snap Count - average |

*Fig 3: User Defined Variables*

**Feature Selection: Association Rules Mining vs Boruta:**

To determine which features were statistically important, Association rules mining was run and compared to the results of the Boruta package.

**Association Rules Mining:**

Association Rule Mining (also called Association Rule Learning) is a common technique used to find associations between many variables. It is often used by grocery stores, e-commerce websites, and anyone with large transactional databases. The team utilized association rules mining package to explore the data and to inform which variables are potentially significant. The team ran the item frequency plots and the Apriori algorithm. The team sorted by high support and lift, and the team filter on RHS => Fantasy Tier 1, in order to determine which variables were associated with favorable fantasy outcomes.

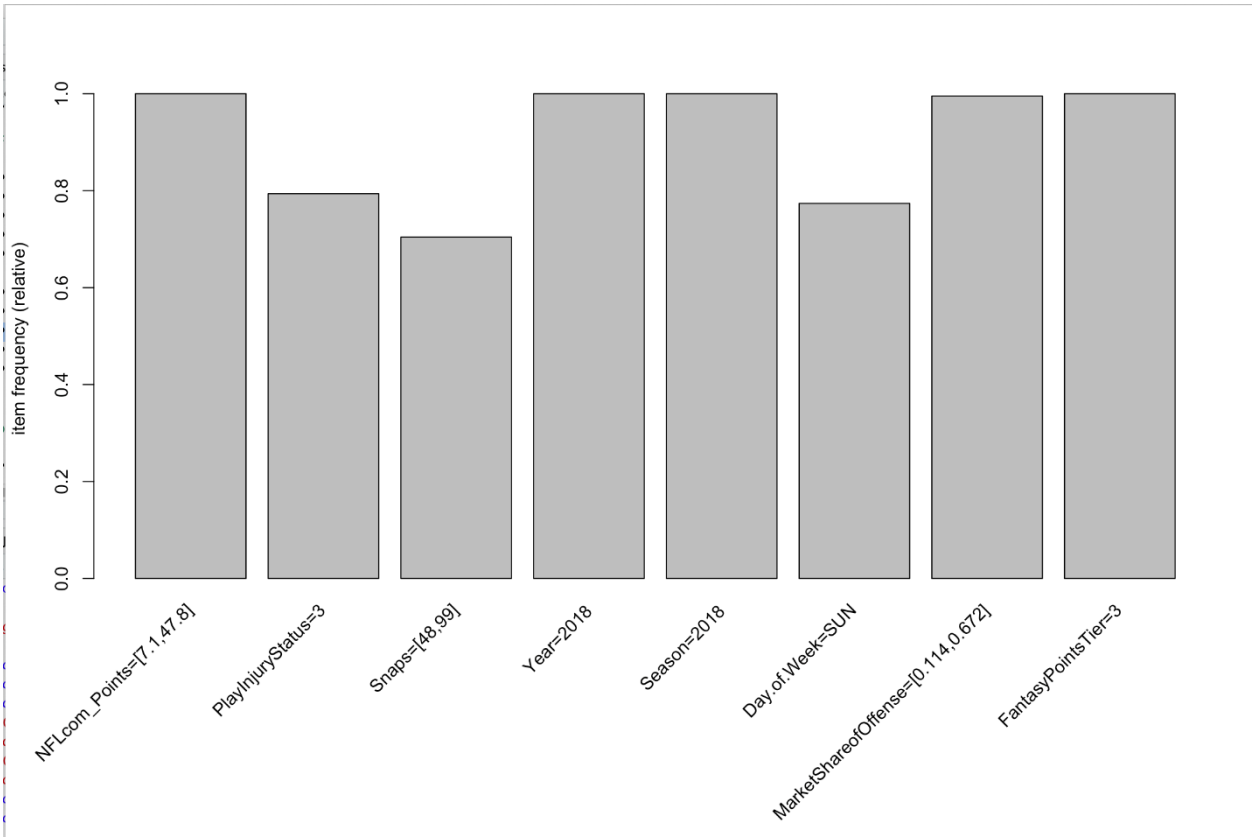


*Fig 4: Item Frequency Plot from Association Rules Package*

**Boruta Algorithm:**

The Boruta algorithm is a wrapper built around the random forest classification algorithm. It tries to capture all the important, interesting features you might have in your dataset with respect to an outcome variable.

As highlighted below the Boruta results confirmed many of the variables identified within the association rules evaluation. With one notable exception, Boruta rejected “player_injury” status as not being significant.

| name | meanImp | medianImp | minImp | maxImp | normHits | decision |
|---|---|---|---|---|---|---|
| stadium | 2.7484692 | 2.6936016 | 0.05233952 | 5.377627 | 0.69411765 | Confirmed |
| nfl_week | 3.5149861 | 3.3222963 | 1.26219667 | 6.704698 | 0.87058824 | Confirmed |
| opposing_defense_tier | 3.5089406 | 3.7042263 | 1.09658828 | 6.104741 | 0.89411765 | Confirmed |
| game_number | 3.4532724 | 3.3768663 | 1.57176711 | 5.752079 | 0.89411765 | Confirmed |
| seasons_played | 3.6946227 | 3.6506335 | 1.48296317 | 5.592815 | 0.92941176 | Confirmed |
| nfl_season | 3.8576887 | 3.8772364 | 1.72523476 | 5.748276 | 0.94117647 | Confirmed |
| defensive_index | 4.5135771 | 4.6504856 | 2.23717628 | 7.451507 | 0.96470588 | Confirmed |
| coach_tier | 4.0332178 | 4.036105 | 0.78133709 | 5.902723 | 0.97647059 | Confirmed |
| average_rrypg | 14.0589877 | 14.0397749 | 11.31640452 | 16.674225 | 1 | Confirmed |
| average_pypg | 10.5537708 | 10.5017795 | 9.15241289 | 12.148751 | 1 | Confirmed |
| average_rt_ds | 12.2292899 | 12.2881781 | 10.56591573 | 13.719613 | 1 | Confirmed |
| average_pt_ds | 10.6379078 | 10.7407355 | 9.05284098 | 11.908196 | 1 | Confirmed |
| average_pass_attempts_per_game | 9.826896 | 9.796303 | 8.53677112 | 11.485724 | 1 | Confirmed |
| average_touches_targets_per_game | 12.9169453 | 12.9256947 | 11.04682583 | 14.573337 | 1 | Confirmed |
| depth_chart_pos | 6.391541 | 6.2946301 | 4.34461405 | 7.772948 | 1 | Confirmed |
| average_turnovers | 8.627706 | 8.5963914 | 7.32197927 | 9.806889 | 1 | Confirmed |
| team_fp | 38.4753153 | 38.7891925 | 33.63229071 | 41.563632 | 1 | Confirmed |
| average_snap_count | 11.8941195 | 11.8879362 | 9.83903003 | 13.559842 | 1 | Confirmed |
| snaps | 16.4052855 | 16.4213352 | 14.54767869 | 18.113592 | 1 | Confirmed |
| over_under | 5.5265203 | 5.5389795 | 3.67267204 | 7.401793 | 1 | Confirmed |
| market_shareof_offense | 32.5494109 | 32.5143516 | 30.54639141 | 34.907528 | 1 | Confirmed |
| fantasy_points_tier | 41.0989951 | 41.1729007 | 38.07210544 | 43.892525 | 1 | Confirmed |
| stadium_surface | 0.7384015 | 1.0474425 | -2.08305941 | 2.472506 | 0 | Rejected |
| weather | 0.1108376 | 0.3322683 | -1.44055346 | 1.196358 | 0 | Rejected |

*Fig 5: Boruta Variable importance*

As listed below, the final dataset included all of the user-defined variables, with the weather being the notable exception.

```
Observations: 5,590
Variables: 16
$ FantasyPointsTier                  <int> 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 3, 1, 2, 1, 1, 2, 1, 1…
$ Average_PassAttemptsPerGame_Tile_Num   <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 1, 2…
$ Average_PTDs_Tile_Num              <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 1, 2…
$ Average_PYPG_Tile_Num              <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 1, 2…
$ Average_RRYPG_Tile_Num             <int> 4, 4, 3, 7, 3, 6, 4, 9, 3, 9, 1, 8, 9, 3, 10, 3, 10,…
$ Average_RTDs_Tile_Num              <int> 4, 6, 1, 8, 8, 3, 3, 8, 1, 6, 2, 6, 4, 4, 8, 5, 6, 8…
$ Average_SnapCount_Tile_Num         <int> 3, 8, 4, 8, 6, 6, 7, 9, 7, 9, 3, 8, 1, 3, 1, 2, 1, 9…
$ Average_TouchesTargetsPerGame_Tile_Num <int> 4, 3, 4, 9, 2, 7, 2, 9, 1, 9, 2, 10, 8, 4, 7, 4, 9, …
$ Average_Turnovers_Tile_Num         <int> 4, 4, 3, 3, 4, 4, 4, 4, 2, 4, 2, 4, 1, 3, 4, 4, 1, 4…
$ Coach.Tier                         <int> 2, 4, 2, 4, 2, 4, 4, 2, 4, 2, 2, 2, 2, 2, 4, 4, 4, 4…
$ MarketShareofOffense_Tile_Num      <int> 6, 5, 9, 7, 4, 8, 5, 9, 1, 9, 1, 9, 2, 6, 9, 3, 4, 9…
$ DepthChartPos                      <int> 3, 4, 3, 4, 2, 3, 5, 4, 2, 3, 2, 3, 2, 2, 2, 2, 2, 3…
$ PlayInjuryStatus                   <int> 3, 3, 3, 3, 3, 3, 3, 3, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3…
$ DefensiveIndex_Tile_Num            <int> 3, 2, 3, 2, 3, 2, 2, 3, 2, 3, 3, 3, 3, 3, 2, 2, 2, 2…
$ OpposingDefenseTier                <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4…
$ nfl_week                           <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
```

*Fig 6: Finalized Dataset for Modeling*

The Final NFL data set was imported into R as a .csv file. A summary of the data set was generated. Then, the data set was transformed to numeric from int type. The label, Fantasy Point Tier, was changed to factor type. The data was then split into a training and test set. The training set contained weeks 1 to 14 of the 2018 NFL games data with 4361 observations of 15 variables. The test set consisted of weeks 15 to 18 with 1229 observations of 15 variables. The data sets had similar groupings. The nfl_week variable was removed from both data sets.

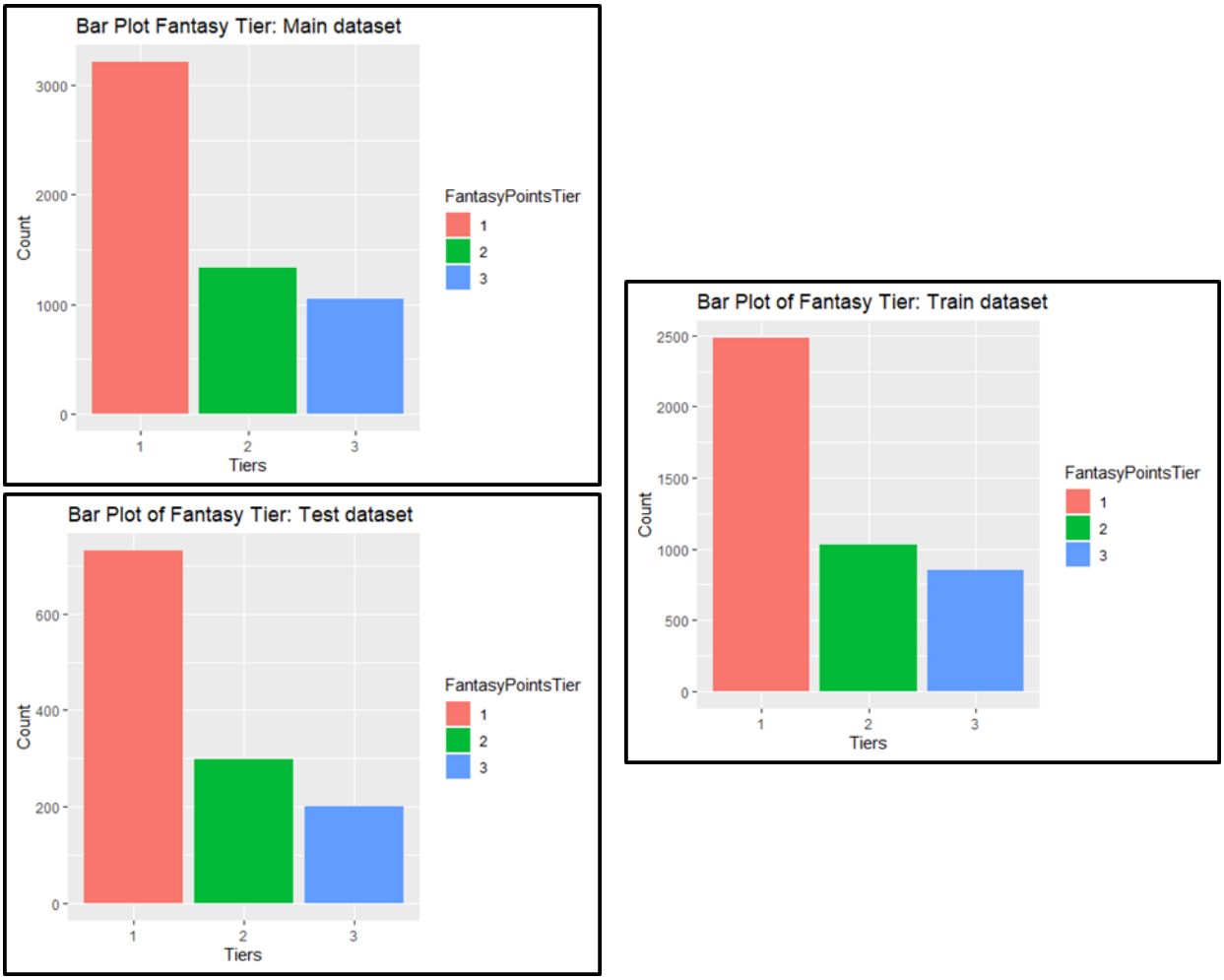


*Fig 7: Bar plots of predictor variable by data set – shows the test & train sets are balanced.*

**Clustering, Decision Tree, Random Forest, Naïve Bayes, kNN, SVM**

After this preparation, HAC and kmeans clustering analysis with a k of 3 was run on the main data set. A matrix, silhouette plot, and optimal cluster plot were created to compare the results to the actual tiers. The first machine learning model ran to analyze the training and test data was a decision tree. The caret package was utilized with the "rpart" method. A decision tree was generated as well as a confusion matrix to describe these results. Two random forest models were then generated utilizing the Caret package with the e1071 "rf" method and the randomForest packages. A variable of importance plot, histogram of tree size and confusion matrix were produced.

Three other machine learning models were run. The first was Naïve Bayes model using the naiveBayes package. From the results, a plot of the counts of observations by tier and a confusion matrix were created. The final two machine learning models were a kNN model and an SVM model utilizing the Caret package. For the kNN model, a confusion matrix and accuracy plot were generated and for the SVM model results, a confusion matrix, and a variable of importance were produced.

## Results

### Clustering

The cluster analysis did not group the fantasy tiers into the 3 groups expected (Figure 8). The optimal number of clusters was 10 (Figure 9). The silhouette cluster plot also demonstrated that the clusters were not far apart and at a width of .18 the data is not structured (Figure 10). As these results were not exculpatory, no further analysis was done using clusters. The accuracy of the clustering results varied between as range of 20 to 38% percent for the three tiers.

Clustering Results

| | Prediction | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Prediction | 2546 | 1596 | 1448 |
| Actuals | 3209 | 1329 | 1052 |
| Accuracy | 21% | 20% | 38% |

*Figure 8. Clustering accuracy.*

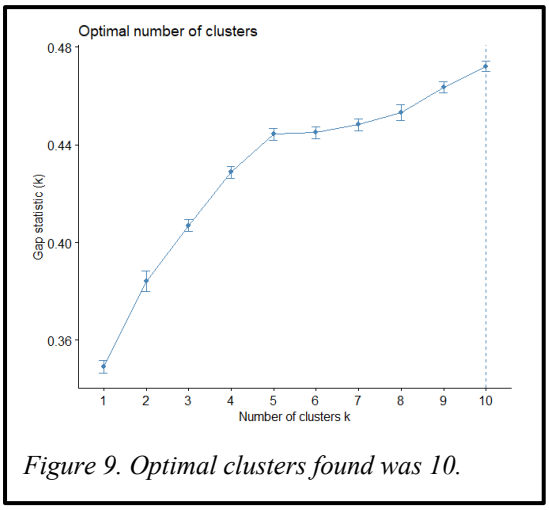


*Figure 9. Optimal clusters found was 10.*

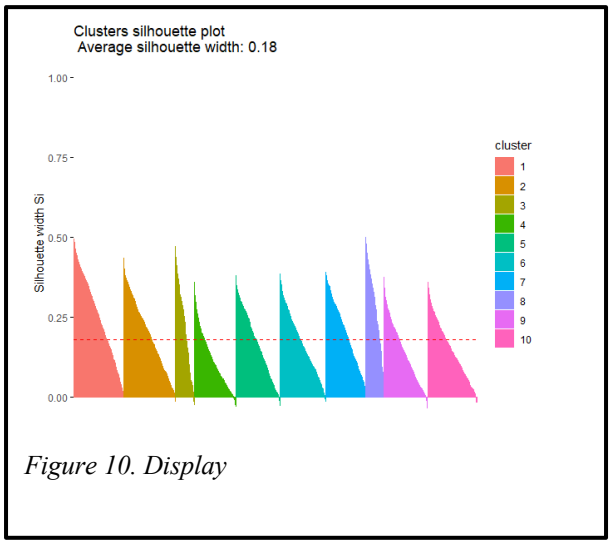


*Figure 10. Display*

After the clustering, predictive models were run to gain more insight into the data set. The predictive results ranged between 75%- 89% accuracy across all models.

### Decision Trees

The Decision tree with a tune length of 10 and a repeated cross validation of 10 produced an accuracy of 89%. This model generated the results with the highest accuracy (Figure 11). It had the highest accuracy predicting tier 1 and the lowest accuracy with tier 2. It also has a high Kappa number and minute P value lending more support for its accurate result. The decision tree had 7 splits with Market Share as the initial split as the variable of most importance (Figure 12).

Other variables with high importance were Average Snaps, Average Pass Attempts, Average Yards per game, Average Running Touchdowns and Average touches/targets per game.

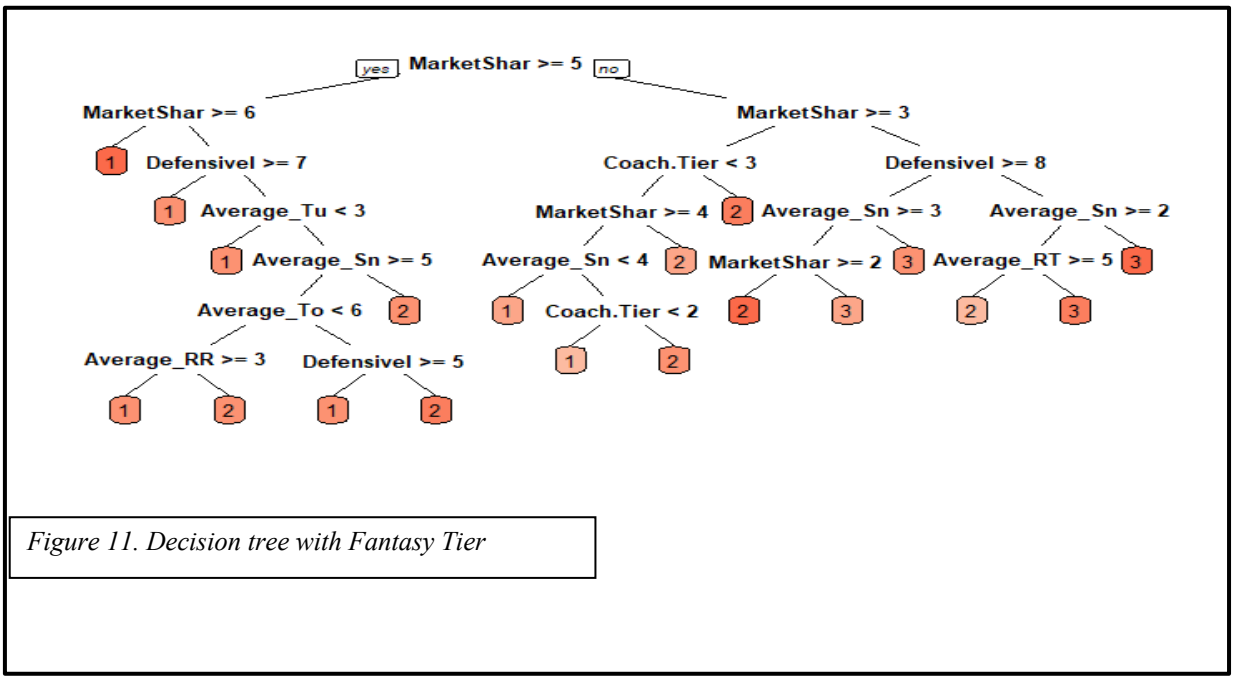


*Figure 11. Decision tree with Fantasy Tier*

Decision Tree Confusion Matrix

| Prediction | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 688 | 29 | 1 |
| 2 | 42 | 236 | 35 |
| 3 | 0 | 33 | 165 |

| Overall Statistics | |
|---|---|
| Accuracy | 0.8861 |
| 95% CI | (0.867, 0.9033) |
| Info Rate | 0.594 |
| P-Value | 2E-16 |
| Kappa | 0.7983 |
| McnemaR P-Value | 0.3287 |

| Statistics by Class | Class: 1 | Class: 2 | Class: 3 |
|---|---|---|---|
| Sensitivity | 0.9425 | 0.7919 | 0.8209 |
| Specificity | 0.9399 | 0.9173 | 0.9679 |
| Pos Pred Value | 0.9582 | 0.754 | 0.8333 |
| Neg Pred Value | 0.9178 | 0.9323 | 0.9651 |
| Prevalence | 0.594 | 0.2425 | 0.1635 |
| Detection Rate | 0.5598 | 0.192 | 0.1343 |
| Detection Prevalence | 0.5842 | 0.2547 | 0.1611 |
| Balanced Accuracy | 0.9412 | 0.8546 | 0.8944 |

*Figure 12 .DT Confusion Matrix*

## Random Forest

The random forest was then run using the caret package with e1071 and the randomForest package. The accuracy was measured at 82% (Figure 14). Similar to the decision tree, it had the most success predicting tier 2. Market share was again the variable of most significance. No other variable even came close (Figure 13).

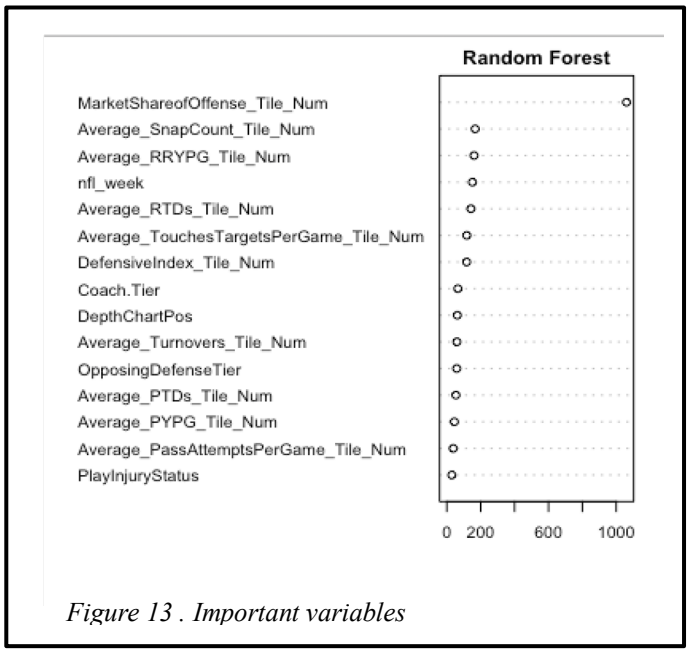


*Figure 13 . Important variables*

Random Forest Confusion Matrix

| Prediction | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 643 | 30 | 2 |
| 2 | 82 | 197 | 28 |
| 3 | 5 | 71 | 171 |

| Overall Statistics | |
|---|---|
| Accuracy | 0.8226 |
| 95% CI | (0.8001, 0.8436) |
| No Information Rate | 0.594 |
| P-Value [Acc > NIR] | 2.2E-16 |
| Kappa | 0.6943 |
| P-Value | 1.43E-09 |

| Statistics by Class | Class: 1 | Class: 2 | Class: 3 |
|---|---|---|---|
| Sensitivity | 0.8808 | 0.6611 | 0.8507 |
| Specificity | 0.9359 | 0.8818 | 0.9261 |
| Pos Pred Value | 0.9526 | 0.6417 | 0.6923 |
| Neg Pred Value | 0.843 | 0.8905 | 0.9695 |
| Prevalence | 0.594 | 0.2425 | 0.1635 |
| Detection Rate | 0.5232 | 0.1603 | 0.1391 |
| Detection Prevalence | 0.5492 | 0.2498 | 0.201 |
| Balanced Accuracy | 0.9083 | 0.7715 | 0.8884 |

*Figure 14. RF Confusion Matrix*

## kNN

The kNN model was run next. It had an accuracy of 82% (Figure 15). It had a strong sensitivity to tier 1 at .92 and the lowest sensitivity for tier 2 at .58.

k-Nearest Neighbors

| Prediction | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 674 | 63 | 2 |
| 2 | 43 | 174 | 40 |
| 3 | 13 | 61 | 159 |

| Overall Statistics | |
|---|---|
| Accuracy | 0.8194 |
| 95% CI | (0.7967, 0.8405) |
| No Information Rate | 0.594 |
| P-Value | < 2.2e-16 |
| Kappa | 0.6781 |
| Mcnemar's Test P-Value | 0.001029 |

| Statistics by Class | Class: 1 | Class: 2 | Class: 3 |
|---|---|---|---|
| Sensitivity | 0.9233 | 0.5839 | 0.7910 |
| Specificity | 0.8697 | 0.9108 | 0.9280 |
| Pos Pred Value | 0.9120 | 0.6770 | 0.6824 |
| Neg Pred Value | 0.8857 | 0.8724 | 0.9578 |
| Prevalence | 0.5940 | 0.2425 | 0.1635 |
| Detection Rate | 0.5484 | 0.1416 | 0.1294 |
| Detection Prevalence | 0.6013 | 0.2091 | 0.1896 |
| Balanced Accuracy | 0.8965 | 0.7474 | 0.8595 |

*Figure 15. KNN Confusion Matrix*

## Naïve Bayes

The Naive Bayes model had an accuracy of 75%. This was the weakest result. It struggled to identify both tier 1 and 2 at close to 50% accuracy (Figure 16).

NaiveBayes Confusion Matrix

Prediction

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 664 | 105 | 10 |
| 2 | 53 | 155 | 86 |
| 3 | 13 | 38 | 105 |

Overall Statistics

| | |
|---|---|
| Accuracy | 0.7518 |
| 95% CI | (0.7267, 0.7758) |
| No Information Rate | 0.594 |
| P-Value | 2.2E-16 |
| Kappa | 0.5444 |
| Mcnemar's Test P-Value | 7.182E-08 |

Statistics by Class

| | Class: 1 | Class: 2 | Class: 3 |
|---|---|---|---|
| Sensitivity | 0.9096 | 0.5201 | 0.52239 |
| Specificity | 0.7695 | 0.8507 | 0.95039 |
| Pos Pred Value | 0.8524 | 0.5272 | 0.67308 |
| Neg Pred Value | 0.8533 | 0.8471 | 0.91053 |
| Prevalence | 0.594 | 0.2425 | 0.16355 |
| Detection Rate | 0.5403 | 0.1261 | 0.08544 |
| Detection Prevalence | 0.6338 | 0.2392 | 0.12693 |
| Balanced Accuracy | 0.8396 | 0.6854 | 0.73639 |

*Figure 16. Naïve Bayes Confusion Matrix*

## Support Vector Machines

The last model run was an SVM model with a cross validation of 10 and a range of Cost from .01 to 2 using the Caret method. It had a predictive accuracy of 84% (Figure 17). It had similar sensitivity (86%, 89%) to both tier 1 and 3 but was only 70% percent sensitive to tier 2. Tier 2 was skewed towards tier 3 which was in contrast to all other models (Figure 18). The variables of importance table emphasized the strength of the Market Share in the model. It has a strong importance across all tiers. Interestingly, the other variables had similar importance in tier 1 and 3 but had more variation for tier 2.

SVM Confusion Matrix

Prediction

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 652 | 23 | 0 |
| 2 | 75 | 207 | 28 |
| 3 | 3 | 68 | 173 |

Overall Statistics

| | |
|---|---|
| Accuracy | 0.8397 |
| 95% CI | (0.818, 0.8598) |
| No Information Rate | 0.594 |
| P-Value | 2.20E-16 |
| Kappa | 0.7237 |
| Mcnemar's Test P-Value | 3.062E-10 |

Statistics by Class

| | Class: 1 | Class: 2 | Class: 3 |
|---|---|---|---|
| Sensitivity | 0.8932 | 0.6946 | 0.8607 |
| Specificity | 0.9539 | 0.8894 | 0.9309 |
| Pos Pred Value | 0.9659 | 0.6677 | 0.709 |
| Neg Pred Value | 0.8592 | 0.901 | 0.9716 |
| Prevalence | 0.594 | 0.2425 | 0.1635 |
| Detection Rate | 0.5305 | 0.1684 | 0.1408 |
| Detection Prevalence | 0.5492 | 0.2522 | 0.1985 |
| Balanced Accuracy | 0.9235 | 0.792 | 0.8958 |

*Figure 17. SVM Confusion Matrix*

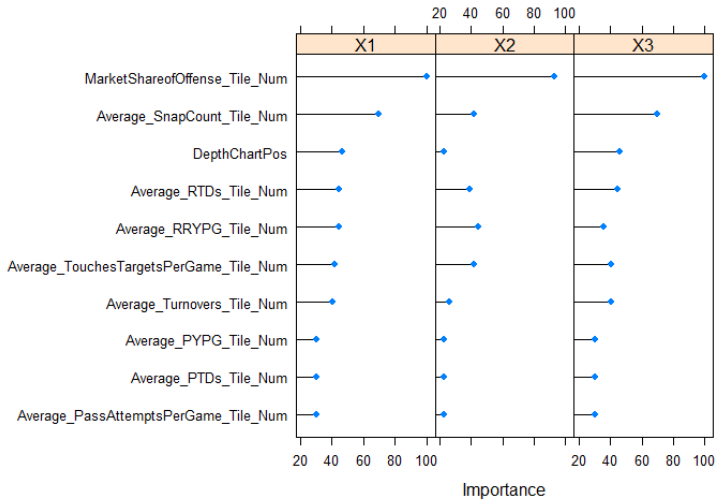


*Figure 18. SVM important variables.*

There are a number of consistent themes throughout all the models. The clearest is that the Market Share variable has the most impact on fantasy football scoring. The second interesting result is that the players who are in tier two have more variation in their outcomes. This is likely highly dependent on their place on the depth chart. If they are not a top player at their position, they potentially struggle against good defenses and thrive against poor ones. Thus, causing the models disparities when trying to predict their fantasy tier. And the final consistent theme is that the models have excellent accuracy and high confidence for predicting fantasy football scoring.

**Conclusion**

After iterating through many machine learning model configurations, the key drivers of positive fantasy football outcomes became apparent. The following are the significant variables that players need to consider in order to succeed at fantasy football:

- Market Share
- Opposing Defensive Competency
- Player Availability
- Practice Status Depth Chart Position
- Coaching
- Average TDs Scored

For market share, fantasy football players must identify who will receive a largest portion of their respective team's fantasy output. This is the most critical consideration to dominating in fantasy football. The higher the market share a player has, the more fantasy points they are going to score. Players with high market share are matchup independent – this means it doesn't matter which defense they face. Opposing defensive competency becomes a factor when a football player's market share is average. The better the defense the player faces, the lower the amount of points the football player will score.

The remaining variables are player availability, depth chart position, coaching, and average TDs scored. Player availability is critical factor to fantasy football success. If the player is out or doubtful, they are unlikely to score points. Where if a player is questionable or limited, the player is likely to score points. Depth chart position is significant in the sense that it doesn't matter if a player is a 2nd string player. Many second-string players have consistently high fantasy football outcomes. For coaching: the better the coaching, the higher the fantasy football scores. Finally, the higher number of average TDs the player scored, the more likely the player was to have positive fantasy football outcomes. Touchdowns are high value fantasy plays, so players that consistently score touchdowns are predicted to have future success.

The team successfully completed its goals of building models that predict fantasy football outcomes and identifying the associated drivers of these outcomes. The team developed a productional model in Excel for future use and created a checklist to ensure the above variables are considered when playing fantasy football in the future.