Fantasy Football Predictor for Wide Receiver's Yards (December 2019)

Manish Goud and Ram Bala

***Abstract*—The National Football League is a $15 billion industry making it the most popular sport in the United States of America. Fantasy football has played a major role in making football the most popular sport in America with more than 50 million people playing fantasy football. This demonstrates the popularity and reach of fantasy football and why people are looking to gain the competitive edge in order to perform well in fantasy football. Therefore, we are predicting how well a wide receiver will do in fantasy football season by using machine learning models such as linear regression, support vector regression, and k-nearest neighbors regression in order to predict a wide receiver's reception yardage in a season. Wide receiver performances from 2016 to the 2018 were used in order to train the models by looking at statistics such as games played, receptions, and touchdowns. This was then tested against the 2019 season in order to see how well the models did in predicting each wide receiver's yardage for the 2019 season. All three models performed well, with the linear regression model performing the best on average across different thresholds. Within an error of 50 yards, the linear regression model was accurate 66% of and within 100 yards, it was accurate 86% of the time.**

I. Introduction

Fantasy football has helped make the National Football League (NFL) the most popular sport in the United States of America by engaging over 50 million users yearly and is one of the biggest reasons that there is such a huge interest in the NFL. Every fall, these millions of people will make fantasy teams and draft their teams in hopes of winning their leagues, with some of the leagues having cash prizes and pride at stake. These millions of people take their fantasy football very seriously, and thus try to exploit any advantage they can get in order to win their leagues.

Our goal for this report was to predict the best performing wide receiver in the 2019 season by predicting each wide receiver's projected yardage by using data from the 2016 to the 2018 season. Eventually, we will expand to predict the other positions but we decided to first focus on wide receivers as there are generally less variables in determining their performance and the fact that there was a large enough dataset with relevant performers that would allow the results to be more accurate when compared to other positions.

We used Pro-Football-Reference to get the statistics for the 2016 to 2018 season at the wide receiver position and then cleaned the data by only focusing on the top 200 performing wide receivers each season, and by filling in null values with 0. We then ran a correlation matrix to see which statistics highly correlated in determining a wide receiver's yardage. As we see in Figure 1, "Games", "Games Started", "Receptions", and "Targets" play a strong part in determining yardage.

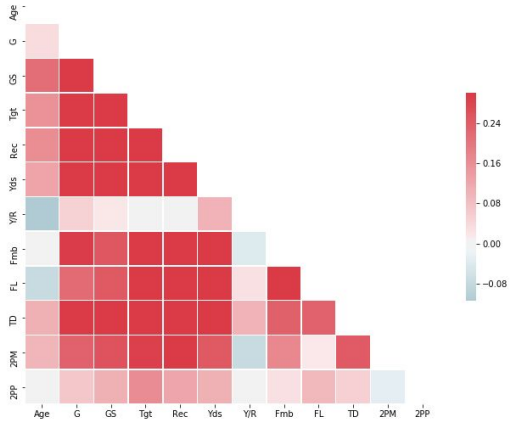

Figure 1: Correlation Matrix

After seeing which variables had high correlations in determining a wide receiver's yardage, we ran three different machine learning models in order to predict the yardage. We looked at linear regression, support vector regression, and k-nearest neighbors regression and had the training data be the statistics from the 2016 to 2018 seasons testing on the 2019 season.

II. Linear Regression

The linear regression model was 66% accurate within 50 yards and as seen in Figure 11, most of the predictions were considered to be very accurate.

The *train_test_split* function in Python was used to split the data into training and testing subsets. The linear regression function then runs the analysis for different subsets to train the model. This feature was used for KNN and SVR as well.

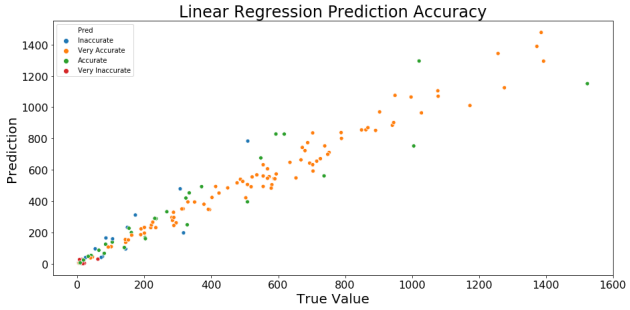

Figure 2:Linear Prediction Accuracy

The accuracy for each of the predictions of the Linear Regression model can be seen in Figure 2 above. The majority of the points were very accurate or accurate, and few of the points were inaccurate or very inaccurate. This tells us that the linear regression model supplies a good fit for the data, which is proven by the high R^2 value 0.952.

The method for calculating the inaccuracies was based on the relationship between the actual value and the prediction. If the ratio between the two was greater than 2, the prediction was labeled very inaccurate. If the ratio between the two was less than 1.2, then the prediction was labeled as very accurate. Values in between were labeled accurate and inaccurate based on which side they were closer too.

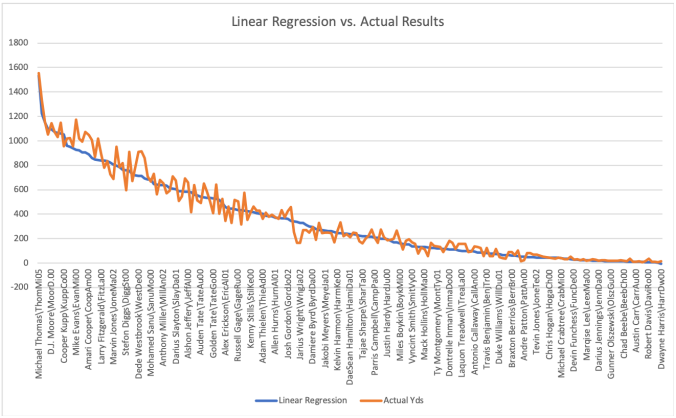

Figure 3: Linear Regression vs. Actual

The trend of the linear regression model compared to the actual results is shown in Figure 3. The linear regression model predictions are all close to the actual yard values, but the curve does not have as much variation as the actual. This was something we assumed would happen since it is difficult to predict anomalies in the data.

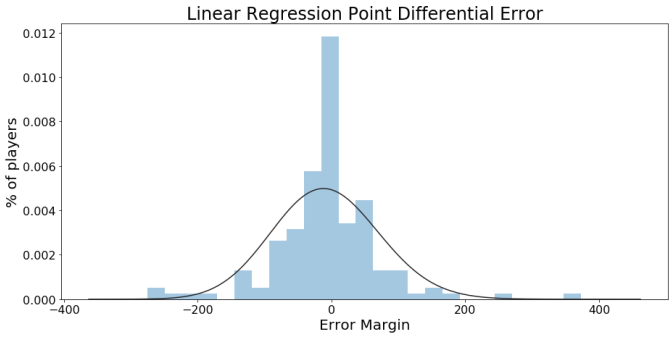

Figure 4: Point Differential Error for LR

To further explore the accuracy of the Linear Regression model, we checked if each prediction was within 200 yards of the actual value. Figure 4 shows that most of the predictions fit this criteria and there were few points that were out outside of this range.

III. Support Vector Regression

The support vector regression model was 62% accurate within 50 yards and as seen in Figure x, most of the predictions were considered to be very accurate but there are more predictions considered to be accurate and inaccurate.

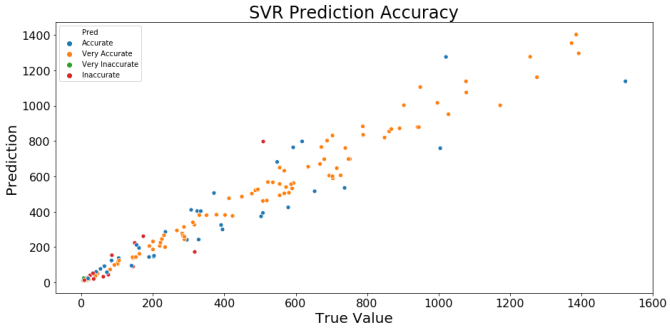

Figure 5: SVR Prediction Accuracy

The SVR model also proves to be an accurate model for the data with a R^2 value of 0.951. As shown in Figure 5, there are very few points that are labeled as inaccurate or very inaccurate.

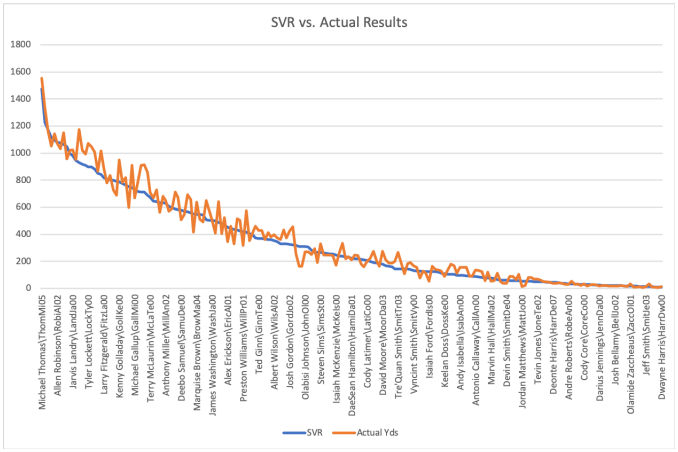

Figure 6: SVR vs. Actual

The trend of the support vector regression model compared to the actual results is shown in Figure 6. Similar to the linear regression trend, the SVR model predictions are all close to the actual yard values as well. There are subtle differences between the SVR and LR curve, but it is evident that both models fit the data well. The SVR model is also more accurate towards the middle whereas the linear regression model tends to overestimate.

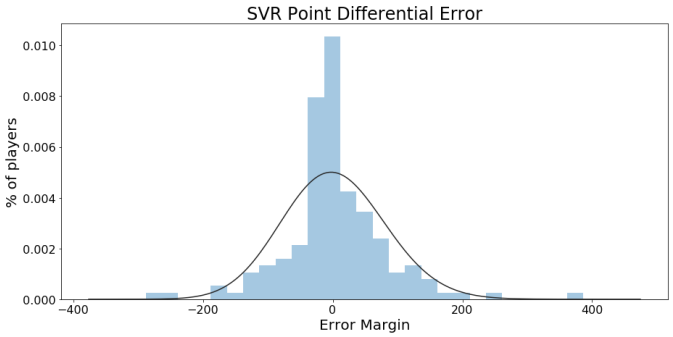

Figure 7: Point Differential Error for SVR

The SVR model also had very rare occurrences of predictions that were more than 200 yards apart from the actual value. As shown in Figure 7 above, you can see that most of the predictions for the SVR model are close to the 0 error margin and point differential also resembles a normal distribution curve.

IV. K-Nearest Neighbors Regression

The k-nearest neighbors regression model was 61% accurate within 50 yards and as seen in Figure 8, most of the predictions were considered to be very accurate but there are more predictions considered to be inaccurate.

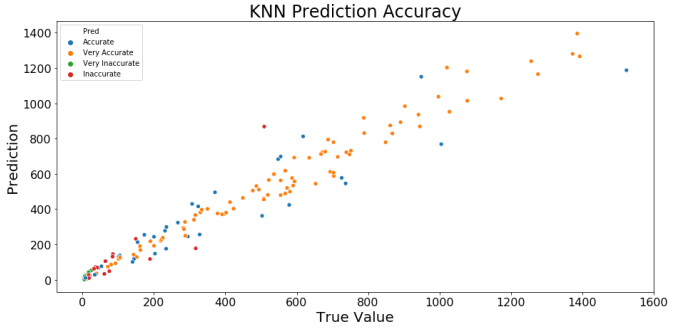

Figure 8: KNN Prediction Accuracy

The reason that there is likely more inaccuracies in the k-nearest neighbors regression model compared to the other models is because we placed a higher emphasis on the distance between the neighbors instead of a uniform distance. This means that closer neighbors will have a stronger influence than compared to neighbors which are further away thus strongly pulling the points one way or the other.

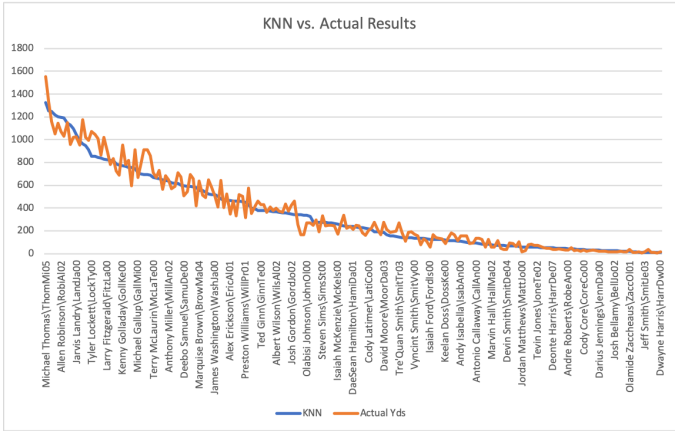

Figure 9: Point KNN vs. Actual

Figure 9 shows the actual results compared to the predicted results by the k-nearest neighbor model. The k-nearest neighbor model performs similarly when compared to the other models but tends to have a more jerky fit to the actual result.

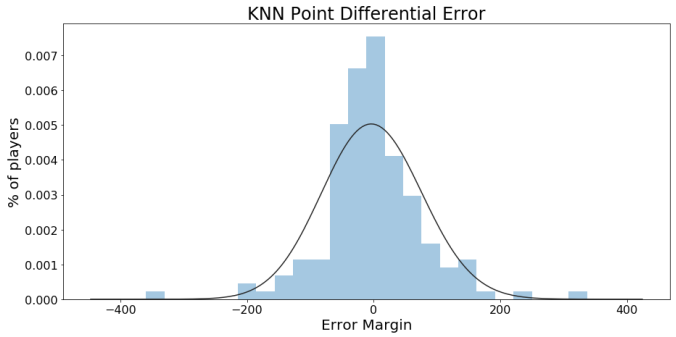

Figure 10: Point Differential Error for KNN

Similar to the other two models, the KNN model also had very rare occurrences of predictions that were more than 200 yards apart from the actual value. You can also see from the Point Differential Error graphs from all the models, that the KNN model has the most values concentrated near 0 error margin and so has the tallest normal distribution curve (Figure 10). Linear regression has the highest mean number of values concentrated at 0 error margin (Figure 4), as you can tell by the highest bar at 0 error margin.

V. Conclusion

All three models predicted fairly well as there is an extremely high correlation between the variables chosen to predict the yardage for each wide receiver. The linear regression model performed the best on average, however, when compared to the other two models. As seen in the figures throughout the report, the linear regression model had a higher total of "Very Accurate" predictions, a better fit compared to the actual results, and a smaller differential error.

	LR	SVM	KNN
Accuracy within 25 yards	38.78%	41.72%	37.09%
Accuracy within 50 yards	65.99%	61.59%	60.93%
Accuracy within 100 yards	86.39%	84.77%	84.77%
Accuracy within 200 yards	95.92%	96.69%	97.35%

Figure 11: Accuracy by Threshold

Actual		LR Predicted		SVM Predicted		KNN Predicted	
Player Names	**Yds**	Player Name	Yards	Player Name	Yards	Player Name	Yards
Michael Thomas	1552	Michael Thomas	1533.6	Michael Thomas	1473.1	Michael Thomas	1323.7
Chris Godwin	1333	DeAndre Hopkins	1223.4	DeAndre Hopkins	1223.5	Julio Jones	1251.9
D.J. Moore	1174	Julian Edelman	1151	Julian Edelman	1169.1	Allen Robinson	1246.7
Mike Evans	1157	Chris Godwin	1099.5	Keenan Allen	1115.2	DeAndre Hopkins	1213.5
Julio Jones	1150	Keenan Allen	1095	D.J. Moore	1097	Julian Edelman	1198.6
DeAndre Hopkins	1142	Allen Robinson	1074.3	Allen Robinson	1079.8	D.J. Moore	1196.8
Amari Cooper	1073	D.J. Moore	1062.4	Chris Godwin	1075.5	Keenan Allen	1189.1
Stefon Diggs	1073	Cooper Kupp	1059	Julio Jones	1065	Chris Godwin	1146
Kenny Golladay	1052	Julio Jones	1051.3	Cooper Kupp	1048.8	Cooper Kupp	1127.6
Keenan Allen	1046	Jarvis Landry	961.91	Tyler Boyd	997.38	Tyler Boyd	1095.8
Cooper Kupp	1031	Stefon Diggs	953.41	Jarvis Landry	982.55	Jarvis Landry	1043.2

Figure 12: Top 11 Receiving Yards for 2019

The linear regression model also was more accurate on average across the various accuracy thresholds performing the complete best for two thresholds, 50 yards and 100 yards as seen in Figure 11.

When seeing the predictions for the top 11 wide receivers with the most yardage, the linear regression model had the most success as it was able to predict eight of the top 11 receivers, compared to the seven for support vector and k-nearest neighbor as seen in Figure 12.

There are anomalies in the predictions in all the models due to various factors such as injuries, suspensions, new coaches, etc. One factor that would help in more accurately predicting the yardage for a wide receiver is how well each wide reciever's quarterback performs. This is due to the inherent dependence that a wide receiver has on his quarterback.

References

[1] B. McCormick, "Rise of fantasy football played big part in league's growth," *Sports Business Daily*, 02-Sep-2019. [Online]. Available: https://www.sportsbusinessdaily.com/Journal/Issues/2019/09/02/Media/Fantasy.aspx.

[2] N. Kapania, "Predicting Fantasy Football Performance with Machine Learning Techniques," *Stanford University*, [Online]. Available: http://cs229.stanford.edu/proj2012/Kapania-FantasyFootballAndMachineLearning.pdf

[3] Pro Football Statistics and History, "Pro-Football-Reference. [Online]. Available: https://www.pro-football-reference.com/