Revolutionizing Football: Using Machine Learning to Predict Future Performances for Quarterbacks

Anya Nagpal

1. ABSTRACT

Football is a largely popular American sport, and within fans of the sport, betting and fantasy football are also very popular. Football is also very unpredictable. On a certain day, a player may perform better because of the weather. On another day, that player may perform poorly because of an injury. Similarly, yearly stats also fluctuate. A quarterback may throw for 5,000 yards one year, and the next year they may throw for 3,000 yards. Due to this, it is important to have a reliable application that can aid users of betting and fantasy football in which players they should put bet on, or choose for their fantasy teams. My motivation behind this project was to create something that could further betting and fantasy football, and even increase traction. My approach to doing this was putting together dataframes of yearly statistics from 75 quarterbacks, and seeing if my linear regression model could predict statistics for a future season. Reliable sources such as the NFL website, or ESPN, were used in order to put these statistics together. The main objective was to produce as accurate as possible stats, which came over time as more seasons were added, and the code was tweaked.

2. INTRODUCTION

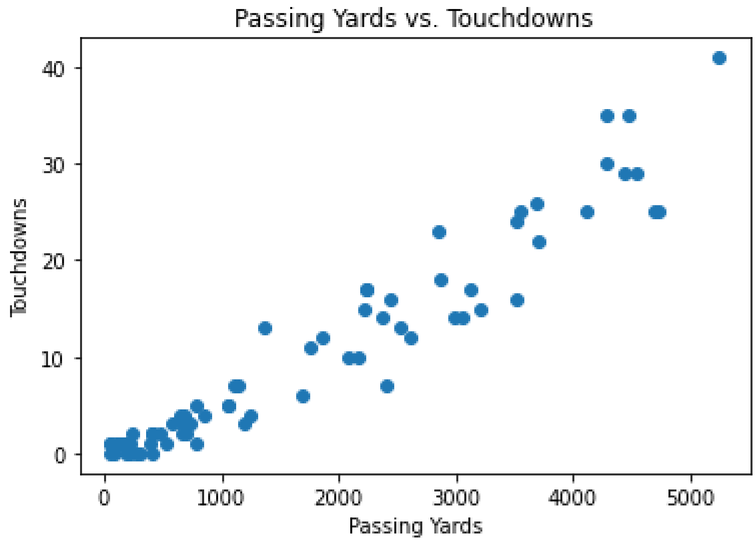
The objective for this research was to investigate the accuracy of machine learning models in predicting the performance of NFL quarterbacks, using variables such as previous season stats and team records. With the growing popularity of fantasy football and sports betting, there is demand for an application with the ability to help one succeed in creating their fantasy football team, or betting on the correct player. In order to complete this objective, stats from three consecutive seasons for the quarterbacks were compiled, trained and tested on a model using these stats. The model predicted passing yards, touchdowns, interceptions, passing attempts, passing completions, fantasy points, and quarterback rating. The accuracy of this model was then tested using mean squared error. With this application, a new way to approach sports betting and fantasy football drafting will be created. Football fans who like betting money on and playing fantasy football with friends have always wanted a system that could help them draft the perfect team and win their league. To begin, four spreadsheets were created. Three of them were used for previous seasonal stats, and one was used for team records across those three seasons. The model was trained on the 2019-2020 season, and the 2021-2022 season. It was tested with the stats for the 2021-2022 and 2022-2023 seasons to make sure the predicted stats were accurate. Then, the calculations for QBR and fantasy points were created with the formulas that the NFL uses to calculate these. Finally, new columns were added to the dataframe that would loop through each player's seasonal stats and then predict the stats for that player's 2023-2024 season. By testing my model with those two seasons, truthful results were ensured.

3. BACKGROUND

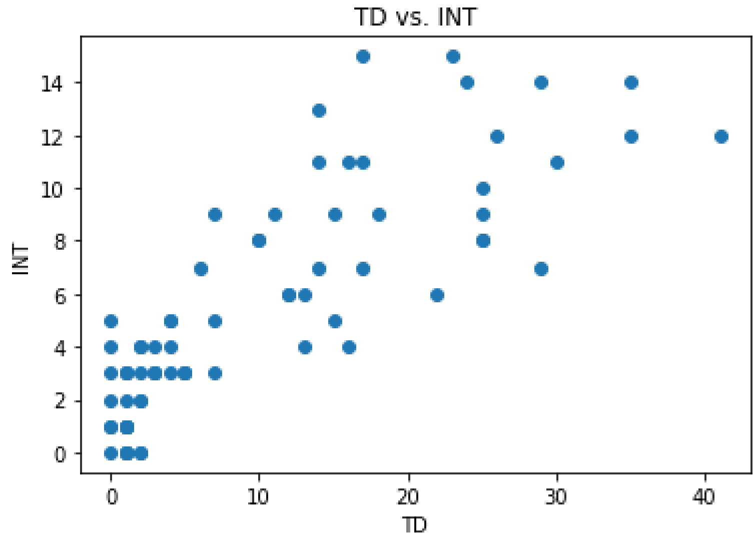
In the past, there have been people who have trained models to predict the outcome of football games. FiveThirtyEight, a company who has used and trained one of these models, uses an elo rating system in order to predict the winner of a football game, as stated in their article titled “How Our NFL Predictions Work”. Multiple factors are taken into account. These factors include home field advantage, rest and bye weeks, all quarterback ratings, weather forecast, team record, and other smaller factors. For each factor in favor of a specific team, a certain amount of points are added to that team. Those points go towards the chances of a team winning a certain game. For example, the team with a home advantage gets 48 points for being at home, and 4 points for every 1,000 miles the other team had to travel. Most of the weight for prediction was put on quarterbacks, and all quarterbacks were taken into account in case of the starting quarterback getting injured. Higher drafted rookies were given a higher rating, while undrafted free agents or late draft picks were given a much lower rating, affecting the prediction system. Then, the quarterback who is starting for a team gets their rating points added to the overall points a team has. Then, the team with the higher points is projected to win. In addition to predicting game outcomes, these models could also project which quarterbacks would start each game by incorporating suspensions, injuries, and rest games. As of right now, ESPN is the largest company that has predictions for players’ future seasons. As stated on their website, ESPN takes into account the points a team is projected to score in each of their games by reviewing their opponents for the season, how likely a player is to be utilized at their position, the length of the players career, how good they have played in the past, sports experts’ predictions, and talent of the defensive opponent.

4. DATASET

The dataset includes a total of 75 samples, each representing a different quarterback. There are nine features included in the dataset. team, player, passing yds, pass, touchdowns, interceptions, completions, attempts, and year. Passing yards represent the total number of yards a quarterback threw for during a specific year, touchdowns represents the total number of touchdowns thrown by a quarterback, interceptions represents the total number of interceptions thrown, completion represents how many passes a quarterback completed, attempts represents the amount of times a quarterback attempted to throw the ball, and year represents the year of the season, which could be 2021-2022, 2022-2023, etc. In order to keep track of all of the QB’s across multiple CSV’s, the files were merged on Player and Team into one dataframe which included statistics from all the mentioned seasons, for each quarterback instead of having three different dataframes for the three seasons. To better understand the dataset, different graphs of each feature were plotted. Graphs of tds vs ints, pass yds vs tds, and completions vs attempts were created.



Quarterback Passing Yards vs Touchdowns scatter plot



Quarterback Touchdowns vs Interceptions scatter plot

There was a noticeable trend that quarterbacks with more passing yards had a significantly higher amount of passing attempts than those with lower passing yards, so that was taken into account. Data was collected using the NFL and ESPN websites. There were no suitable datasets online for the project, so I decided to make my own datasets. Stats such as yearly fantasy points, average quarterback rating (qbr), rushing yards, and rushing tds were not used. There was no interest in calculated rushing yards and tds, and yearly fantasy points and average qbr have formulas for them, so formulas for those were used instead. The formulas for QBR and fantasy points used was as follows:

QBR=
$$((data1['Comp\_2023\_24'] - 30) / 20 + ((data1['Pass\ Yds\_2023\_24'] / data1['Att\_2023\_24']) - 3) * 0.25 + (data1['TD\_2023\_24']) * 0.2 + 2.375 - (data1['INT\_2023\_24'] * 0.25)) * 100 / 6 / 3.$$

Fantasy Points =
$$data1['Pass\ Yds\_2023\_24'] * pass\_yd\_pts + data1['TD\_2023\_24'] * pass\_td\_pts + data1['INT\_2023\_24'] * int\_pts.$$

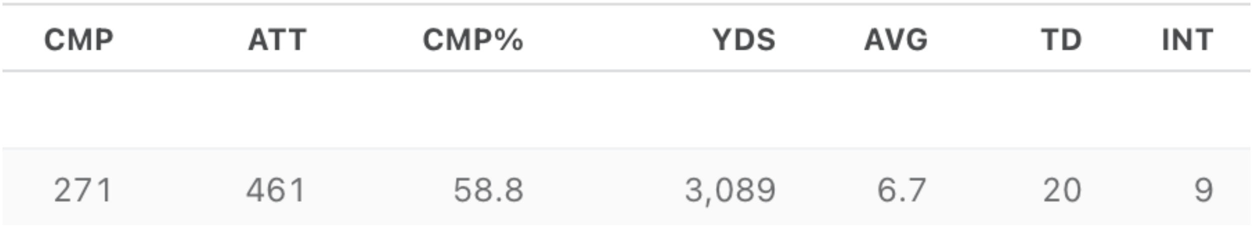
All of the stats predicted for the 2023-24 season of a player were used to calculate QBR and fantasy points. For the training set, statistics from seasons 2019-2021 and 2021-2022 were used. For the testing set, statistics from the 2021-2022 and 2022-2023 seasons were used. The set used a 70/30 split. The training set consisted of 150 samples, and the testing set consisted of 150 samples as well.

5. METHODOLOGY/MODELS

For this project, the performance of NFL quarterbacks was predicted using machine learning models. Specifically, linear regression was utilized to predict quarterback statistics based on their previous season's performance. Data from three NFL seasons (2020, 2021 and 2022) was gathered and split into training and testing sets to evaluate the performance of the models. For the linear regression model, a simple architecture with five inputs was used, which were the stats from the seasons (passing yards, touchdowns, interceptions, completions, and attempts) and outputted were the same five inputs, except the inputs were the predicted stats for the 2023-24 season. The model was trained on the 2020 and 2021 data and evaluated on the 2021 and 2022 data. RNN models were looked into in order to incorporate team records into predicting stats. The RNNs would learn the relationship between the stats a QB had for a certain season and the record their team had, adjusting accordingly to predict the QB's future stats. For example, if a QB's team has a poor record of 3-14, the RNN might adjust the predicted QB stats to be lower than what it would have predicted, assuming that the QB will have less opportunity to throw passes, and might make mistakes due to playing from behind, facing tougher opponents, or not having a very well structured team and wide receivers to throw to. Conversely, if the QB's team has a strong record of 14-3, the RNN might adjust the predicted QB stats to be higher, assuming that the QB will have more opportunities to throw passes to talented wide receivers and put up big numbers against other weaker teams.

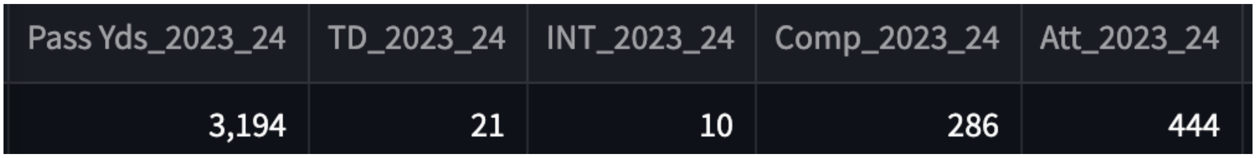
Here is how the models' projections stacked up to actual stats:

Josh Allen's actual stats from a season -



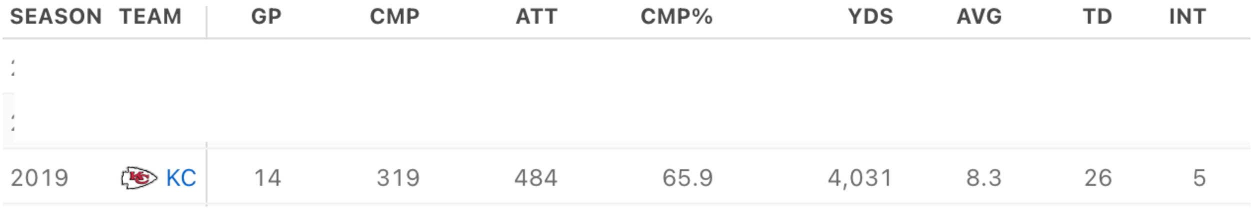
CMP	ATT	CMP%	YDS	AVG	TD	INT
271	461	58.8	3,089	6.7	20	9

Josh Allen's projected stats using the model -



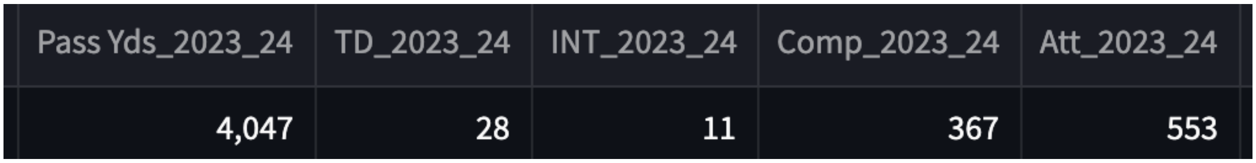
Pass Yds_2023_24	TD_2023_24	INT_2023_24	Comp_2023_24	Att_2023_24
3,194	21	10	286	444

Patrick Mahomes actual stats -



SEASON	TEAM	GP	CMP	ATT	CMP%	YDS	AVG	TD	INT
2019	KC	14	319	484	65.9	4,031	8.3	26	5

Patrick Mahomes projected stats -

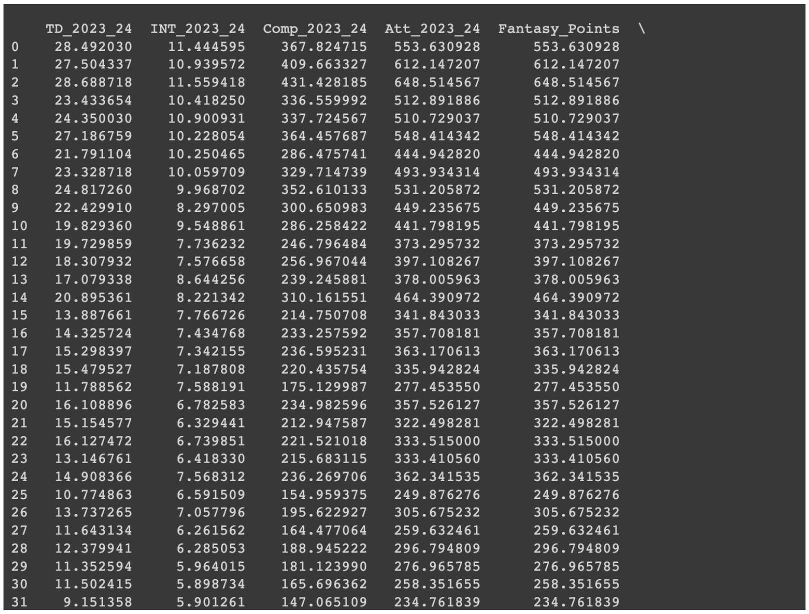


Pass Yds_2023_24	TD_2023_24	INT_2023_24	Comp_2023_24	Att_2023_24
4,047	28	11	367	553

The model was accurate to a certain extent, however there are still some things that need to be fine tuned when predicting the stats.

6. RESULTS AND DISCUSSION

The objective of this project was to evaluate how machine learning could predict the stats of a certain quarterback based on stats from previous seasons. The results and objective were achieved using linear regression, as seen by this table where the 2023-24 stats of the quarterback are printed:



```
    TD_2023_24  INT_2023_24  Comp_2023_24  Att_2023_24  Fantasy_Points  \
0    28.492030    11.444595    367.824715   553.630928      553.630928
1    27.504337    10.939572    409.663327   612.147207      612.147207
2    28.688718    11.559418    431.428185   648.514567      648.514567
3    23.433654    10.418250    336.559992   512.891886      512.891886
4    24.350030    10.900931    337.724567   510.729037      510.729037
5    27.186759    10.228054    364.457687   548.414342      548.414342
6    21.791104    10.250465    286.475741   444.942820      444.942820
7    23.328718    10.059709    329.714739   493.934314      493.934314
8    24.817260     9.968702    352.610133   531.205872      531.205872
9    22.429910     8.297005    300.650983   449.235675      449.235675
10   19.829360     9.548861    286.258422   441.798195      441.798195
11   19.729859     7.736232    246.796484   373.295732      373.295732
12   18.307932     7.576658    256.967044   397.108267      397.108267
13   17.079338     8.644256    239.245881   378.005963      378.005963
14   20.895361     8.221342    310.161551   464.390972      464.390972
15   13.887661     7.766726    214.750708   341.843033      341.843033
16   14.325724     7.434768    233.257592   357.708181      357.708181
17   15.298397     7.342155    236.595231   363.170613      363.170613
18   15.479527     7.187808    220.435754   335.942824      335.942824
19   11.788562     7.588191    175.129987   277.453550      277.453550
20   16.108896     6.782583    234.982596   357.526127      357.526127
21   15.154577     6.329441    212.947587   322.498281      322.498281
22   16.127472     6.739851    221.521018   333.515000      333.515000
23   13.146761     6.418330    215.683115   333.410560      333.410560
24   14.908366     7.568312    236.269706   362.341535      362.341535
25   10.774863     6.591509    154.959375   249.876276      249.876276
26   13.737265     7.057796    195.622927   305.675232      305.675232
27   11.643134     6.261562    164.477064   259.632461      259.632461
28   12.379941     6.285053    188.945222   296.794809      296.794809
29   11.352594     5.964015    181.123990   276.965785      276.965785
30   11.502415     5.898734    165.696362   258.351655      258.351655
31    9.151358     5.901261    147.065109   234.761839      234.761839
```

In this image, everything that needed to be predicted (passing yards, touchdowns, interceptions, etc) was printed. The motivation behind this project was to make it easier for fans of sports betting or fantasy football to bet on players or select players for their fantasy team more accurately, which it should achieve. Although it does seem to work as intended, there still needs to be data compiled from more previous seasons in order to ensure the program works as accurately as possible. With just three seasons, it is hard to have very accurate predictions. Similarly, some limitations are that the model does not include rookies and only includes 75 quarterbacks, however that can be solved by adding statistics for rookies and more quarterbacks in the future.

7. CONCLUSION

There can definitely be some uncertainties in the data, as only 3 seasons worth of statistics were included, and no further variables that could impact performance were added. In spite of that, the project did complete the research objective, which was to analyze how a model could predict future season statistics based upon past season statistics. In the end, the model was able to use these past season statistics to output predictions for the 75 quarterbacks in the dataframe, which was the goal. This project could influence others to create similar applications for different sports that they like, such as soccer, basketball or tennis. This could allow for more accurate sports betting, and raise interest in a certain player if they are predicted to perform better, in all sports worldwide. In the future, many more seasons should be added to make the outputed data more accurate, and in addition to that add more variables such as team record, weather, and injuries should also be added in order to make the data as accurate as possible.

8. ACKNOWLEDGEMENTS

Thank you to my Inspirit mentor Eric Bradford for working with me and helping me whenever I had any questions or concerns regarding my project.

9. REFERENCES

ESPN. (n.d.). *NFL Player Stats*. ESPN.com. Retrieved April 7, 2023, from https://www.espn.com/nfl/stats/player

FiveThirtyEight. (2023, January 9) *How our NFL Predictions Work.* FiveThirtyEight.com. Retrieved April 7, 2023, from https://fivethirtyeight.com/methodology/how-our-nfl-predictions-work/