Are You Ready For Some Football?

Jeff Dunagin

DSC 530

Final Project

My EDA uncovered several things about the data. The first thing I did was to plot the distribution of several variables. I will note that all of these variables from running backs only for the 2021 season. The variables I plotted were total fantasy points, total yards, total touchdowns, total usage (passes thrown to + rushing attempts), and passes thrown to MINUS rushing attempts. We found a similar distribution for the first 4 variables: they appeared exponential or pareto, with a small value for many and a large value for few. The last variable was bell-curve like with a negative skew.

I then used the last variable to split the running backs into two groups, based on how much they received vs ran the ball. I compared the two groups with a PMF, and I found that one group accounted for almost all of the high scorers.

I then made a few CDFs for total points and usage, which contained much of the same information as the histogram. As it appeared to be a Pareto distribution, I fit a pareto distribution to the total usage curve. This revealed differences between the actual data and an idealized model.

I used a scatter plot to test the relationship between usage and points next. There was a very high correlation between the two, with a Pearson’s r of 0.97 and R^2 of 0.94. This motivated me to test usage against a new related variable, efficiency. It turns out that those two variables did not have a relationship, as the p-value showed it to be not significant.

Finally, I did a regression analysis to learn which component of the fantasy football score (yards, touchdowns, receptions), were most correlated with the final score. The answer turned out to be yards.

A lot was analyzed in this project, but I felt like I could have explored the relationship between different variables more. Lots of time was spent on analyzing the distribution of the variables themselves, but that isn’t necessarily a bad thing. In that vein, I didn’t have to make and unwarranted assumptions about my data, since I was mostly uncovering it.

If I had more regression techniques to test relationships under my belt, this analysis could have been better. The relationships between these variables are extremely complicated, and I don’t think linear regression would get me very far. A future project would tackle this with a wider range of tools for analysis.