Connor Pandiscio

Professor Carter

MA 346

June 29, 2021

MA 346 Final Report

My original goal for this project was to incorporate my basic understandings of machine learning into predicting outcomes of players for the upcoming fantasy football season. More specifically, looking into whether or not a particular player would fall off and have a bad season.

To do this I found four datasets from pro-football reference.com which contained data on football players statistics for a given year. The nice thing about the datasets was they happened to have a column for fantasy football scoring called points per reception which is the same scoring I used in my fantasy football league. This was perfect for my analysis portion because I was able to use this as a good indicator as to whether or not a player had a good fantasy year. However, when I first loaded the datasets into my python notebook there was clearly some cleaning that needed to be done. For instance, the missing values in the datasets needed to be converted to zero in order for my logistic regression model to work. Also, I had to rename the column names accordingly so it was clearer to anyone exploring the dataset, what statistic they were looking at (See Deepnote link for more data cleaning done in this project). Once the four datasets were cleaned, I was ready to start the analysis preparation.

For the analysis preparation I first concatenated all four of the datasets so that it was easier to perform my analysis. After doing this I worked to create a Boolean column called Good\_season, with 1 representing a player having a good season and 0 representing a player having a bad season. I used the points per reception column to differentiate a good season as PPR greater than 95 and a bad season as the opposite. Next, I made two dataframes, one called df\_training and the other called df\_validation. Df\_training contained a random 60% subset of the data and was used to fit the model. Df\_validation contained the other random 40% of the data and was used to test our model.

The steps to performing the analysis can be found in my Deepnote link which is located in my GitHub repository README.md link, but I will walk through what the analysis worked to do. The analysis worked to use the football statistic columns excluding PPR to predict whether or not a player would have a good season for the current year. Then once it predicted the outcome, we could compare how accurate it was to the Good\_season column already differentiating what a good season was for a player. My models score was .92874 for the training data and .94435 for the validation data. I realized that the reason the model scored so highly was because it was using data for the current year to predict whether a player would have a good season for that same current year. It was not what I intended to do and though I had trouble training my model to predict future performance outcomes using the future years Good\_season value, I look forward to fixing this and perfecting my model for the upcoming fantasy football season.

The most interesting finding I found from conducting this project came from learning more about the precision and recall in my specific model. In my model’s example precision is the number of times the model predicted a player would have a good season and in fact did have a good season, divided by the total amount of players predicted to have a good season correct or not (True Positives + False Positives). More specifically, precision tells us the probability that when the model predicts a good season it will be a good season. In my model’s example recall is the number of times the model predicted a player would have a good season and in fact did have a good season divided by the total number of players that in fact had a good season (True Positive + False Negative). More specifically, recall is the probability that the model will predict a good season. As an example, to contrast these two metrics consider a model that always predicted a good season no matter what. The recall metric would be 1 because you predicted all good seasons. However, the precision metric would be terrible because of the large number of predictions of players that were sought to have a good season but in fact did not. In reality, you want a mixture of both metrics and this is what the F1 metric works to do by forming a geometric average of the two.

In conclusion, I really enjoyed learning more about one possible thing machine learning can do. In the coming months before the upcoming fantasy football season, I will be working to improve my model and updating my repository.