Capstone Project : In-depth Analysis (Machine Learning)

Regression Models.

My first step was to decide what type of machine learning model to apply to my dataset. Considering that the purpose of my project is to predict fantasy points scored for players and doesn't incorporate classifying players into different categories, the most appropriate machine learning technique to apply to my dataset is a regression model. Having decided on regression analysis being the best fit for my project I then needed to determine the most appropriate type of regression analysis to apply a; Linear, Random Forest, or Lasso regression model.

While implementing the different models I adjusted the hyper parameters for all the models and would then run the ".score" method to test if the change I made resulted in a higher or lower score.The Linear regression model had the fewest hyper parameters and the best score using defaults. I was able to improve the score by adjusting the "train_test_split()" parameters. I found that a test_size=0.275 and a random_state=411 resulted in the highest score. However, this was not the case when tuning the parameters for the Random Forest and Lasso models. I found that adjusting the "n_estimators" parameter to equal 2000 resulted in the highest score, this was accompanied with the "train_test_slpit()" "random_state" parameter being equal to 42. The Lasso Regression model had the most hyper parameters to tune and after some adjustments I was able to create a much more effective, higher scoring, model. I found that adjusting the "alpha" to equal 0.06777, the "max_iter" to equal 50000, and the "tol" to equal 0.00009999 resulted in the highest score, this was accompanied with the same "train_test_split()" parameters as the Linear model.

In conclusion, after testing all the types of regression models and fine tuning their hyper parameters, I found that the Linear Regression model performed best, followed very closely by the Lasso Regression model. Theses two models only differed in scoring by 0.01067 percentage points. The results for the Linear Regression model were 0.9988641, the results for the Lasso model were 0.9987574, and the results for the Random Forest model were

0.9569426. With the gap being so slim between the Linear and Lasso models I chose to display my Lasso Regression Model code to display how I tuned the parameters. This code is specifically modeled/tuned to project fantasy points scored for the Quarterback position, the models for the Running Backs, Wide Receivers and Tight Ends are tuned differently to maximize the models score.

Lasso Regression Model code.

```
# Creating Lasso Model
        model_Lasso = Lasso(alpha=0.06777,  max_iter=50000,
            tol=0.00009999)
# Creating train test splits on data
        X_train, X_test, y_train, y_test = train_test_split(X_qb, y_qb, test_size=0.275,
random_state=411)
# Fitting the model
        model_Lasso.fit(X_train, y_train)
# Creating variable for predicting QBs
        top20_qbs = X_qb[:20]
# Predicting top 20 QBs
        _top20_qbs = model_Lasso.predict(top20_qbs)
# Creating DataFrame to hold and visualize the prediction results
        y_top20_qbs = pd.DataFrame(y_top20_qbs)
        qb_names = qb['Player'][:20]
        frames = [qb_names, y_top20_qbs]
        qb_names_pred = pd.concat(frames, axis=1)
        qb_names_pred.columns = ['Player', 'Predicted FantPt']
# Using Linear Regression Model and displaying the results
        qb_names_pred.sort_values(['Predicted FantPt'], ascending=False)
```

Hypothesis development

Working through this project I hypothesized that touchdowns and yards (passing, rushing, or receiving) would be the best predictor for future performance. While doing my

correlation analysis my results supported my hypothesis. The two highest correlated statistics to fantasy points scored where touchdowns and yards, respectively.Touchdowns with a 0.928808 correlation and yards with a 0.900263 correlation. The next highest correlator was completions with a 0.834737 correlation. These data apply to the Quarterback position.

The data supports this throughout the other positions (Running Back, Wide Reciever, and Tight Ends). However, the correlation is much lower. This is due to their ability and frequency to score points in a variety of ways. For example, the highest correlator for the Running Back position is TD.3, or total touchdowns. This stems from the running back having a higher frequency in rushing and receiving touchdowns, as well as scoring points for receptions. So the highest correlator for running backs is at 0.788136 with rushing yards not far behind with a correlation of 0.747687.

Where my hypothesis failed to find support was in the Wide Receiver and Tight End positions. Both positions highest correlator was the Yds.2 stat, or receiving yards. The next highest correlators were receptions and targets, respectively. An in depth analysis of these results proved that receiving yards and receptions makes more sense to be the best predictor of future fantasy performance. First, scoring touchdowns is much more volatile for the receiving positions compared to the Quarterback and Running Back positions. Second, with this being true, the more reliable source of points for the receiving positions comes from their receiving yards and receptions. Considering that the data is based on the points per reception scoring system, the receptions are much more valuable for the receiving positions than touchdowns.

Conclusion of Machine Learning Analysis

After an in-depth analysis of the machine learning techniques I was able to find the most appropriate, best fitting regression models for each specific position, as well as further develop my hypothesis. Having analyzed the results from the process of creating machine learning models I found that my original hypothesis held true for the Quarterbacks and Running Backs position, however, was not the case for the Wide Receiver and Tight

Ends position. Even though this goes against my original hypothesis it makes logical sense considering the scope of the scoring system and the high variability in receiving touchdowns for a player from year to year. It would be appropriate that the highest correlators to scoring for a receiving position comes in the form of receiving yards and receptions.