NFL Draft

Luke Newman

Project Design

Every year the NFL Draft happens and about 250 college football players get recruited to a team to play professionally. The draft consists of 7 rounds, the first round where the more desired players get drafted and the seventh being the least desired. My goal is to first build a machine learning model that accurately predicts which players will be drafted and then use my model to help NFL recruiters find potential players for their team. Machine learning can narrow down the search from 15,000 players to about 300 and can also search by position, if a team is looking to fill certain positions on their team.

To create this model, I used train_train_split to separate my data into three categories. Training data, used to train my model, validation data, used to score different model's performance, and test data, used to give a final performance score on my best model. Now since this is a binary classification problem, I used classification algorithms like K-nearest-neighbors, RandomForestClassifier, SVC, LogisticRegression, and GradientBoostingClassifiers. Also, as stated above my data was highly imbalanced. Out of 22,000 players only 600 of them were
Drafted. I used oversampling on the training data to handle this imbalance, specifically Synthetic Minority Oversampling Techniques (SMOTE). My metrics to judge performance of different models was ROC AUC Score and used the confusion matrix to gain insights. My best model was GradientBoostingClassifier or XGBoost specifically. To tune hyper parameters, I used both Random Search and Grid Search. First, I use RandomizedSearchCV to find a good set of parameters over a large range of values and then used GridSearchCV to tune the best parameters even further around the output given from RandomizedSearchCV. Finally, using the parameters given from grid searching XGBoost gave me a ROC AUC Score of 0.935. on the test data. Now my model was complete.

Now looking into the confusion matrix, I wanted to gain insights onto what observations my model was misclassifying. The false negatives, players who were drafted but my model predicted undrafted, tended to be higher round picks. The false positives, players who were undrafted, but my model predicted would be drafted, were all players with good stats and quite a few Juniors. Although it is uncommon for Juniors to get

drafted it is possible, so it seemed like my model was good at finding talented younger players. These false positives, I would propose to be potential late round draft picks and the Juniors to be players to keep an eye out for next year's draft or recruit them a year early. Finding late round picks can have a huge impact on the success of your NFL team because they're low risk high reward players. The average salary of a 7th round pick is less than $200,000 and given the chance to play might prove himself quite an asset. Many late round picks have become the stars of their team like the 6th round pick Tom Brady.

Tools

- Python
    - Pandas, Sklearn, Seaborn, XGB, mlxtend, imblearn
- PowerPoint

Data

I webscraped data from two sites. Cfbstats.com has data on every college football player ranging from 2010-2019. My data was from years 2016-2018. Pro-refence-football.com has a list of every player who was drafted so I scraped that data and matched the 2019 draft to the 2018 college football year and so on. My features included the year in school the player was, position, height, weight, touchdowns, points, rush yards, pass yards, receiving yards, interceptions, solo tackles, sacks, attempted passes, completed passes, passing touchdowns, passing interceptions. I binned positions together which have similar positions on the field.

Further Work

For further work, I want to gather more features including what college they played for and how many wins they have. Also, it could be easier to create a different model on each position giving even more accurate predictions. NFL recruiters could find players based on positions they need to fill. Also, I would want to calculate the error made using a machine learning model to select late round picks versus a human choosing. Which one finds stronger players in the late round picks.