College Quarterback Draft Predictions

Using Logistic Regression

Robert Roche
rr696@drexel.edu
CS 383 Final Project

***Abstract*—This project aims to use Machine Learning to predict which round a college quarterback will be drafted based on a variety of statistics, and using logistic regression.**

I. Introduction

The NFL is one of the most popular sports organizations in America. Each spring, new prospects are drafted from college football teams into the NFL where they must prove themselves to become the next generation of superstars. Teams invest millions of dollars into new players in order to bring fresh, explosive talent to their rosters. Being able to understand which round a player may belong in could be critical for a team to strategize when they could draft a player. For the purposes of this project we are only looking at the Quarterback position as it is widely regarded as the most important and is a highly monitored position in every NFL draft.

II. Related Work

For years various sports shows and analytics experts have attempted to make draft predictions. There are many ML approaches to problems similar to this however, I have no found any exact projects attempting to calculate which round a player belongs to necessarily. There are plenty of examples of full draft predictions, predictions of how a player will perform in the NFL, and other draft related predictions.

III. Methodology

In order to make these predictions I will be using methods we have discussed and practiced in class, namely logistic regression.

A. Logistic Regression

Logistic regression is a classification algorithm that will allow me to predict the likelihood a QB will belong to each round in the draft and I can make the prediction based on the greatest likelihood.

B. Python

I will be using a Python Notebook to complete these calculations as there are Logistic Regression tools in the sklearn machine learning libraries. This will make computations easier.

C. Data

I will be gathering my data from Kaggle.com, a website containing a wide variety of datasets. After downloading, I will eliminate unnecessary features and standardize the data to ensure the scales of the features are properly set for my computations.

Figure 1: Sample of the dataset

Player	Age	GamesPla	Completio	Attempts	Yards	Touchdow	Interceptic	RushAtten	RushYards	RushTouch	Heisman	Round
EJ Manuel	23	43	600	897	7741	47	28	298	827	11	0	1
Geno Smit	22	44	988	1465	11662	98	21	245	342	4	0	2
Mike Glenn	23	36	646	1069	7411	63	31	111	-281	3	0	3
Matt Barkl	23	47	1001	1562	12327	116	48	132	-113	6	0	4
Ryan Nass	23	47	791	1312	9190	70	28	242	168	5	0	4
Sean Renfr	23	42	898	1389	9465	51	40	153	-167	9	0	7
Andrew Lu	23	38	713	1064	9430	82	22	163	957	7	0	1
Robert Gri	22	41	800	1192	10366	78	17	528	2254	33	1	1

IV. Experiments and Results

A. Experiment

After some research I realized I needed to create seven binary logistic regression models, one for each round. This is because logistic regression is a binary classifier and thus, each round needs to be learned in a one vs all environment. To do this I created seven different sets of training and testing data, and used sklearn's logistic regression fit methods to get the optimal thetas. I then created a list of the resulting thetas for each of the seven rounds. To make my predictions I iterated through each test sample, multiplying each test sample by the thetas and passing the result through the sigmoid function to gain the likelihood for the given test sample being in each target class. I then used the maximum likelihood to classify the sample.

B. Results

The results of this part of the experiment were not entirely conclusive. The accuracy of the models combined thetas was only about forty-nine percent. Although this seems disappointing, it is not unexpected. Since this experiment only utilized player data, we cannot make a fully accurate prediction on whether a player will be selected in a given round. This is because we are only considering one member of the transaction, the player. The need of each team also influences the draft position of the players. Some teams

prioritize getting certain positions over others. Some teams have bias towards certain players, and some teams will panic if too many players in a position group are drafted and they may trade ahead and take the player earlier than they were projected to be taken. There are many external factors that influence when a player is drafted. However, in a scenario where all teams have the same needs, and they simply take the next best player on the board regardless of position, this model would make much more accurate predictions.

C. One vs All Predictions

Although the combined thetas resulted in only a fifty-fifty success rate, the one vs all model predictions alone found great success. These fell between eighty-five and ninety-five percent accuracies. In other words, if you wanted to know if a given player was going to be drafted in one individual round, this particular program would be able to predict very accurately whether a player would be taken in that round.

V. Concolusion

During this experiment I discovered using logistic regression to predict the round a given player will be taken in the NFL draft is not optimal. This is because there is more to a player's draft position than simply his stats. In a purely mathematical world where players were drafted based on only statistics, this model would work far better. However, trying to use logistic regression to predict if a player will be drafted in a given round works far better than the former. I also discovered using PCA on this smaller dataset actually inhibited the algorithm's learning ability. This was most likely because the dataset was quite simple to start and PCA was over-simplifying the data.

VI. Future Work/extensions

In the future I would love to revisit this problem. Upon doing some research I found some other interesting potential methods for helping create a more accurate model, however the scope of this project did not allow me to explore those avenues. In addition, I believe more data collection, and more features in conjunction with a component analyzing the needs of each team and their positions in the draft could truly enhance this project into something quite useful.

References

Monga Aseem. "NCAA College Quarterback Data". Kaggle.com. av8ramit. September 2019. https://www.kaggle.com/av8ramit/ncaa-college-quarterback-data

"sklearn.linear_model.LogisticRegression". sci-kitlearn.org. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html