Logistic Regression

Group 2

02 12 2019

1. Introduction

The aim of this project is to predict whether a College Football Player will be drafted into the NFL. This ReadMe lines out, how the Logistic Regression model was used to classify the players into the to categories "Drafted" and "Not Drafted".

2. Logistic Regression

The basic idea of a Logistic Regression is to find the logistic function that is best fitted to model a binary, dependent variable. In our case this is the "Drafted" variable. In the end we want to be able to tell whether a player is drafted (1) or not drafted (0). R provides the glm function that allows us to find the coefficients for all the features.

3. Our Approach

We have built different models for the different player positions, as well as for different data samples. The specifics of this are explained in the following sub-chapters.

3.1 Player Positions

First of all, our data only includes the offense player positions quarter back (QB), running back (RB) and wide receiver (WR). We have decided to do so because the performance of offense players is easier to quantify than the one of defensive players. We then decided to trin different models for all players together, and based solely on the data for QBs, RBs and WRs respectively. As the calculations in the discussion/ conclusion part of our documentation show, the models trained on the specific positions perform less well than the one trained on the data of all the players together. This does not only apply to the Logistic Regression model, but is true in general.

3.2 Sampling

Since in our data only a small portion of the players are drafted, we have sampled our data in different ways. In total we trained the models for five data samples. The first one includes the unsampled data, that includes all players that have played in 10 or more games. We have taken out the players with less than 10 games for the simple reason that a player is naturally less likely to be drafted if he has played only few games (e.g. because good players will play more often). The second data set has been oversampled, the third undersampled and the fourth dataset includes under- and oversampling. The fifth and last sample has been sampled according to the Smote method.

In the end we have run a Logistic Regression model for all combinations of player positions and data samples. This amounts to 20 (4 positions including all together * 5 data samples) models.

3.3 Cross-validation

In order to make sure that the model strikes a balance between bias and variance (bias-variance trade-off) we cross-validated all the models 10-fold. This will ensure that the model is not overfitted while at the same time providing accurate predictions.

4. Performance

Compared to the other models we have built, the Logistic Regression is not really a high performer. Of course, this is partly due to the fact that our data is imbalanced (a vast majority of the players have not been drafted), but other models manage to handle this situation in a better way. Furthermore, the models for the individual positions are overall better fitted to the training data, but present more of an overfit than the models trained with the Together data.