MIST 7635
Spring 2024

Group Members
Katelyn Bond
Jim Flannery

Project Description

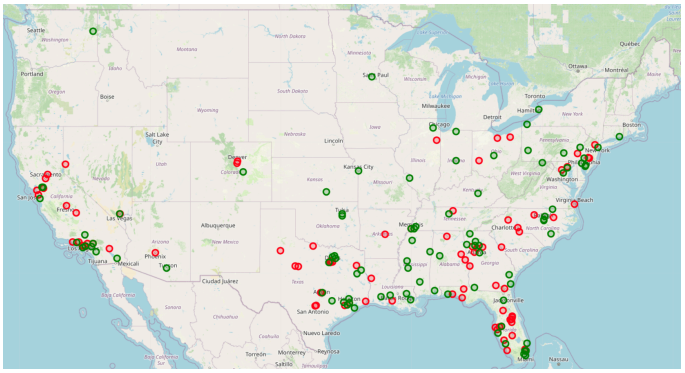
The goal of our analysis is to predict the likelihood of a high school football player making it to the NFL, post-college.

Type of Problem

This is a classification problem. We are predicting two classes - “NFL Draft Pick” or “Not an NFL Draft Pick”.

The data we have to predict this outcome ranges from categorical data such as “All Conference Selection” to numerical data around height, weight, and performance data (rushing yards, carries, etc).

All the players we are examining are running backs so that we have a consistent set of data to compare players. We also worked to select players from across the country.



Three Potential Methods for Solving

Method 1
We can start with a logistic regression model. This should help with identifying the features which have the largest impact on reducing error in the model. Deciding between a lasso or ridge model will be critical to determining the best accuracy.

Method 2
Next, we will try a random forest model as there are a lot of features (a wide data set) and crafting a decision tree might help us quickly identify which features are critical. We might compare the results of a random forest approach to an XGBoost model.

Method 3
We will also utilize unsupervised clustering. This helps us explore potential underlying patterns in the data that were not previously known. Clustering may also help identify subgroups of players who have similar characteristics.

Specifics about training/testing procedures
We will split our data into a training and testing set (75%/25%). In addition, we will need to make decisions about our data in regards to missing data points. Setting missing features to zero might not be the appropriate way to represent missing data.

Steps to Provide Accurate Results

In pre-processing, we will take a number of steps to ensure accurate results. First, we will use the same training and testing data across all three models. We will also use a consistent cross-validation procedure in all models. We will track the training loss to prevent overfitting during cross-validation.

The proposed use case of this data is to help college recruiters select potential high-profile players. Therefore, we will need to make a decision for our classification model about the relative importance of precision vs. recall (or a balance between the two) as a metric while running our testing data. It is our assumption that we'd want to aim for higher recall as we'd prefer to find all potential NFL athletes and recruit some who won't make it. However, if a school is using the tool to decide scholarships, we'd probably want to aim for higher precision.