DESCRIPTION

In the included folder is a number of scripts and datasets. One folder holds the datasets used for our analysis. Note that only the cleaned and integrated datasets are included due to the size and number of our raw datasets. The other folder holds the scripts used to process and analyze the data.

First, the raw data is downloaded and scraped in a series of R scripts and Python jupyter notebooks. The code files are numbered in the order they are supposed to be run in. They should all be run in the same folder and should have their input and output directories set to the same location across all the code files.

Once the data has been downloaded the datasets should be stored in a folder holding all of the raw data. This is also a good point to create folders that will be used later. These include a folder for cleaned data and a folder for model results. Next step is to integrate the different files into a more manageable number of datasets. Using the three separate R scripts, a few intermediate datasets are created out of the many, smaller sets. These datasets are then cleaned to replace missing values with imputed, average values and columns that are clearly irrelevant to the wide receiver position are omitted. Using game data, each player’s NFL value is calculated. The final dataset is saved as ‘NFL\_Featureselection\_wo\_imputeflags.rdata’.

The script correlation\_pca.R determines the most influential features. EPA\_Distribution\_cutoff.R looks at the NFL player performance distribution and decides appropriate cutoff values that determine whether we consider a player to be good or not good. This good/bad binary indicator is then used to classify players using the classificationModel.R. The last analysis step is to then cluster similar players in clusterPlayers.R.

The results of this analysis are then visualized for easy, intuitive understanding. First, a visualization showing what states players grew up in allows recruiters to target specific states for Wide Receiver talent. Next, a visualization of top players allows recruiters to easily see top performers. The clustering visualization shows similar players and would allow for scouts to compare new players to existing, known talent.

INSTALLATION (How to install and setup your code)

Install Python

Install the following Python libraries: pandas, numpy, json, datetime, http.client, time, timeit, pickle, os

Install R

Install the following R packages: tidyverse, randomforest, glmnet, stats, corrplot, rocr, e1071, car, scales, reshape2, cluster, fpc, openxlsx, ggplot2, DAAG, DataExplorer, nflscrapr, cfpscrapr, rpart, rpart.plot, rpart.utils, zoo

EXECUTION (How to run a demo on your code)

All preprocessing steps have been run and the results are available in the ‘modelResults’ folder. If interested in observing how these are derived, execute the following steps:

1. Ensure that Python, Jupyter, and R are installed on your machine and the required packages are also installed.
2. Use the scripts in the folder labeled ‘ImportRawDataScripts’ to load required raw data
   1. Be sure to run the codes in the numbered order
   2. You will need to edit the input and output directories for each file
   3. Skip to Step 3 if raw data is already downloaded
3. Put all raw data into a folder. The recommended name for this folder is ‘CODE/Data/rawData’. (Note: We already have the folder created and contains a file "Players\_missing\_highschool\_hometown.xlsx" that is part of raw data but not created from scripts above)
4. Use the following scripts to integrate, clean, and calculate additional metrics. The scripts must be run in the below order.
   1. Script 1: cleanNfl\_Draft\_1.R  *change lines 13-15 to reflect your folder setup*
   2. Script 2: CleanCFB\_2.R *change lines 12-14 to reflect your folder setup*
   3. Script 3: CleanCFB\_3.R *change lines 11-13 to reflect your folder setup*
   4. If cleaned data is already provided skip to step 5
5. Step 4 will create a number of intermediate files. The file ‘NFL\_Featureselection\_wo\_imputeflags.rdata’ will be the main data file used in later steps
6. The important features must be determined using the script ‘correlation\_pca.R’. The results from this step have been incorporated into the following scripts. There is no functional need to run this script, but can run if interested in observing this analysis
   1. *Change lines 11-13 to reflect your folder setup*
7. We must define who is considered a good or bad NFL player. The script ‘cleanNfl\_Draft\_1.R’ has already calculated a metric that calculates a player's game value. Now we need to define a cutoff value that will determine whether we designate a person is a good or bad NFL player based on that metric (script ‘EPA\_Distribution\_cutoff.R’). The results from this step have already been incorporated into the following scripts but can run if interested in observing
   1. *Change lines 14-15 to reflect your folder setup*
8. Use the college/high school dataset to develop a Random Forest model that predicts NFL performance by running classificationModel.R
   1. *Change lines 27-29 to reflect your folder setup*
9. Use the features used in the Random Forest dataset to cluster similar players by running clusterPlayers.R
   1. *change lines 13-15 to reflect your folder setup*

If interested in visualizing or interacting with the user interface, please navigate to the following website: <https://github.gatech.edu/pages/tmendez3/cs6242_team201_project/>