Darrell Nelson
IST 652
05/26/2019

Final Project Report

Data

Introduction

College athletics is one of America's favorite pastimes. Seeing young men and women hone their craft and compete at the highest level is a soap opera that the world just can't turn off. The NCAA, the nonprofit association that runs college athletics, takes in close to $8 billion a year. According to a Business Insider report, there are now 24 schools that make at least $100 million annually from their athletic departments. With so much media attention, hype, and money on the line schools are clamoring to find a competitive advantage. Many schools choose the path of least resistance and invest millions of dollars into their "amateur" sports teams to build multi-million-dollar facilities, coaches, and staff.

With all this money circulating around the college athletic scene there's no wonder why schools are habitually getting caught for paying their amateur athletes under the table. With such a huge incentive for the school/program to win, attracting talent is top priority. What if there is a way to influence top prospect to come to town without crossing the red line? The analysis laid out below looks into how social media affects preseason rankings and more importantly, recruitment.

The dataset used is a collection of Twitter data compiled by gaining access to the Twitter API through the python program tweepy. The dataset imported was a collection of 3,041 tweets that were released by the teams in the Pacific 12 conference (Pac-12). Each tweet was imported and listed as a json structure. The data was manipulated with numpy and pandas dataframes. There was no preprocessing required. The semi-structured data came with the appropriate labels and datatypes to facilitate the analysis.

Methods of Analysis

Overview

Twitter accounts of all football teams selected were scraped for the number of tweets and how many retweets and likes each tweet received. The schools were then ranked based on the amount of buzz they generated for their teams during the 2018 offseason (02/01/2018 – 07/24/2018). The school rankings were then compared to their pre-season and recruiting rankings. The pre-season rankings were taken from the Pac-12 media day on 07/25/2018. Media members are asked to vote for where they believe each team will place, and each vote has a number assigned to it with more points assigned to teams that will have the most success. Each team's points are then compiled into a list and ranked from highest to lowest. The recruiting rankings were taken from 247Sports.com for the 2018 Pac-12 year. Each recruit that committed to a team is assigned an overall rating, then the 247Sports analytics team uses a Gaussian distribution to rank all recruits on that team. The teams are then ranked in descending order of points.

Data Questions

- Does social media affect how many top football recruits a school gets?
    - Group the counts of the tweets, likes, and retweets each school's official twitter account generated/received during the offseason into a pandas dataframe. Create a column for retweets per tweet, how many people decided the content was so good that they should share it on their story. Create a column for number of likes per tweet, how many people saw and "liked" the post. The data is then sorted in descending order for each attribute. Team positions/rankings by Twitter are then compared to the recruiting rankings given by 247Sports.com. The overall distance between the predicted rankings and the real rankings for each team are calculated and the absolute value is taken. The feature that produced the lowest distance value is the feature with the fewest deviations from the true ranking.

- Is there a correlation between twitter activity in the offseason and pre-season rankings?
    - Preseason polls are released at the end of July or August. The same dataframe and methodology that was used above was used here with a replacement of the recruiting rankings with the Pac-12 media day pre-season rankings.

Output

Outputs will be sorted dataframes that contain the team names, number of tweets, likes, retweets per tweet, and likes per tweet. Two other dataframes: 1) Distance Metric vs Pre-Season Rankings and 2) Distance Metric vs Recruiting Rankings were also outputted. Bar graphs for each attribute ranked in descending order are outputted to the console. A print statement stating the feature that best approximates the top 3 finishers and the overall finishers in the conference with the distance metric for each are outputted to the console.

Program Description

A twitter developer account is needed from "https://developer.twitter.com/en.html" to gain access to the Twitter API. There are four requisite keys/tokens that are linked to the developer account that are removed from the program for security reasons. Tweepy, , pandas, datetime, and numpy libraries are imported. Tweets were searched based on the date range of the preseason, the twitter handle of the team, and the maximum number of tweets desired.

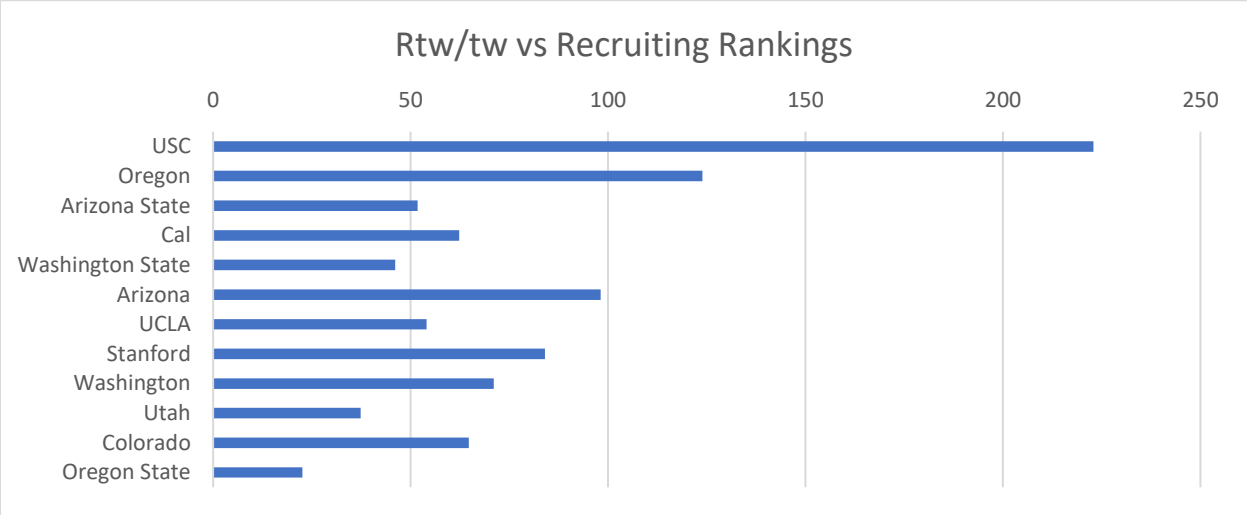
A function named "TwitterFunc" was created to take in a user name, timeframe, and max number of tweets and for each tweet it outputted the number of retweets and likes it had. These numbers were then put into a list and summed up, with the final output of the function being a list containing the total number of tweets found, number of total retweets, and total number of likes. Originally, a for loop was created to iteratively look up each school's Twitter handle and find the accompanying data. However, it resulted in multiple computer crashes. Instead, each handle was looked up manually and their data was placed into an Excel spreadsheet. The spreadsheet was imported back into python with the pandas read_excel function. From then a new sorted table was created for each attribute. Two new columns (retweet per tweet & like per tweet) were added to the original database and sorted as well.

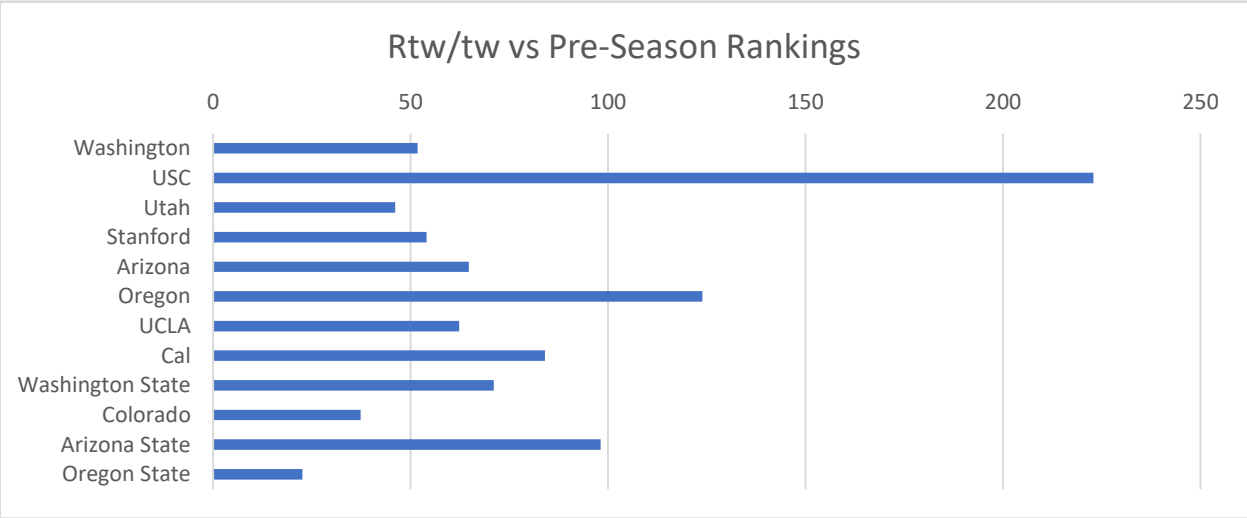
The exported file from Python is the dataframe pulled from the Twitter API that is composed of all 5 features discussed above for each team. The second exported file is the displacement table comparing recruiting rankings to the Twitter features. The third exported file is the displacement table comparing preseason rankings to the Twitter features.

Conclusions

Based on the distance metric and data provided, the attribute that best predicted the rankings for the top 3 teams in recruiting was the retweet per tweet feature. It was also the best feature when looking at the predictive performance for all 12 schools. For preseason rankings there were two metrics tied for the best: tweets and retweet per tweet. When viewing the best predictive performance for all 12 schools the retweet per tweet was the sole champion.

In all, the retweet per tweet metric does the best in predicting the rankings for both recruiting and preseason rankings. Comparing those metrics visually produces the below two graphs:





Of the two graphs above, recruiting rankings produces a clearer trend which cooraberates the fact that it had the lowest overall sum of displacemtent. It appears that recruits can be persuaded to join a team if they are popular on social media. The pre-season rankings which are released by the media appear to be more immune to social media fads. Perhaps because the media take their job of reporting and evaluating teams seriously, and with dailly access to all 12 teams they can compare them based off of practices.

I would strongly recommend for teams to highly invest in building up more buzz in their social media to attract more recruits. Even if the product on the field isn't great, showing teenagers that the organization relates to them and promotes the same type of things they like could be the difference between landing that 5-star recruit and that player going to your rival.