Powerpoint Script

TONY

Question:

- Is it possible to predict a college basketball team's number of wins on a season, based on their statistics?
- Our goal was to predict a teams wins using machine learning methods

Data:

- We got our data from Kaggle
- There is two datasets, one for the seasons 2015-2019 and one from the 2019- 2020 season
- The data is individual team statistics, here is an explanation for each statistic or variable
- **'G':** Games Played
- **'W':** Wins
- **'ADJOE':** Adjusted Offensive Efficiency (An estimate of the offensive efficiency (points scored per 100 possessions) a team
- **'ADJDE':** Adjusted Defensive Efficiency (An estimate of the defensive efficiency (points allowed per 100 possessions) a team
- **'BARTHAG':** Power Rating (Chance of beating an average Division I team)
- **'EFG_O':** Effective Field Goal Percentage Shot
- **'EFG_D':** Effective Field Goal Percentage Allowed
- **'TOR':** Turnover Percentage Allowed (Turnover Rate)
- **'TORD':** Turnover Percentage Committed (Steal Rate)
- **'ORB':** Offensive Rebound Percentage
- **'DRB':** Defensive Rebound Percentage
- **'FTR':** Free Throw Rate (How often the given team shoots Free Throws)
- **'FTRD':** Free Throw Rate Allowed
- **'2P_O':** Two-Point Shooting Percentage
- **'2P_D':** Two-Point Shooting Percentage Allowed
- **'3P_O':** Three-Point Shooting Percentage
- **'3P_D':** Three-Point Shooting Percentage Allowed
- **'ADJ_T':** Adjusted Tempo (An estimate of the tempo (possessions per 40 minutes) a team would have against the team
- **'WAB':** Wins above bubble

ZANE

Model:

- We conducted a linear regression of wins on the team stats to see which stats are most associated with wins
- Wins above bubble was most associated, but we decided to include all of the statistics in the prediction model
- We chose to use logistic regression to make our predictions

Function Slide:

- We made two functions, one that loops through all of the NCAA and one looping through only the “Power” conferences
- The functions take past and current season data and splits into feature and target train test without randomization
- We assigned the past data as the feature and target train set, and the current data as the feature and targe testing set
- The functions then pass these into a logistic regression to train the model, then predicts the target (Wins) for the current season using the feature test set
- After the predictions are made the function creates a dataframe for easy viewing
- The functions also calculate accuracy scores for each conference, and creates one dataframe with the accuracy scores to compare each conference’s prediction accuracy

Outcome 1:

- Our prediction results were varying by conference, although most of the accuracy was between .10 and .20
- A good amount of the conferences however, had 0.00 for accuracy
- While many of the teams predicted wins didn’t match the actual wins, they were fairly close which is a good indication that the model is somewhat accurate

NICK

(GRAPHS)

Outcome 2 :

- This graph shows the average Wins above bubble (Cutoff between making the NCAA tournament and not making it). This graph shows that some conferences have a better chance at making the tournament than other conferences.

Outcome 3:

- This graph is the predicted wins it takes to make the NCAA tournament. The cutoff is zero but is not limited to zero. By this graph the estimated wins it takes to make the tournament is about 23 wins.

Outcome 4:

- These graphs show the Power Ratings by power conference (left) and by team in those conferences (right).

Outcome 5:

- Power rating just in the BIG 10.

Outcome 6:

- Predicted wins for BIG 10 teams.

Outcome 7:

- Power rating for teams who have made the Elite 8 or further in the past 6 years.

TONY

Pros and Cons/ Conclusion:

- Overall it's very hard to predict sports wins, as there are many factors that are difficult or impossible to accurately calculate
- Some of those include: luck, coaching ability, a team's ability to handle pressure, individual player injuries, and players or teams simply having "off nights"
- To improve our model we could add more statistics and add weights to certain statistics we believe affect the prediction accuracy the most
- This would be very difficult to determine, and in most cases would require us to come up with a completely new model to make the machine learning prediction