Fantasy Football Fanalytics

MIS 749 – BUSINESS ANALYTICS

MUDDU APPANERVANDA • KARL-LUDWIG BUECKLE •STEVEN DITSWORTH • HARSHAL SANAP

04/25/2018

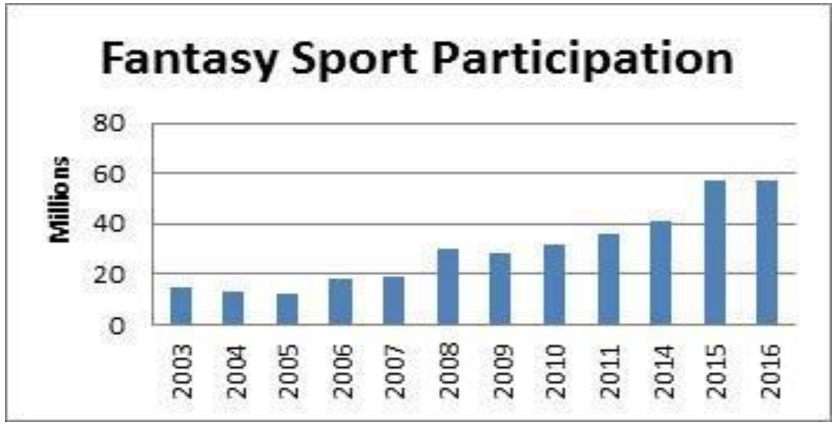
Outline

1. Situation Analysis
2. Executive Summary & Project Goals
3. Data Preparation & Pre-Processing
4. Regression Model
5. Clustering Analysis
6. Classification Model
7. Conclusion & Recommendations

Situation Analysis

Fantasy Football as an Industry

- Fantasy Football currently a $7 billion industry
- 59 million players yearly
- Annual spending of $ 556 per player



DraftKings - the leading commercial internet operator of the fantasy football betting system has approached our team to build a prediction model for fantasy football points of individual players.

- Retain their position as leading commercial Fantasy Football betting system by improving their model
- Maintain competitive advantage over competitors Fan Duel
- Optimize their betting odds for users

Executive Summary & Project Goals

A prediction model to rank Fantasy Football Players can be sold profitably to betting companies (DraftKings, FanDuel) and private users of Fantasy Football

Draft Kings wants to be leaders in Fantasy Football by:

1. Being able to predict the statistics of players based on the games played.
2. Estimate the high performing and low performing players given their background (college)
3. Estimate in advance the winning probability of a game

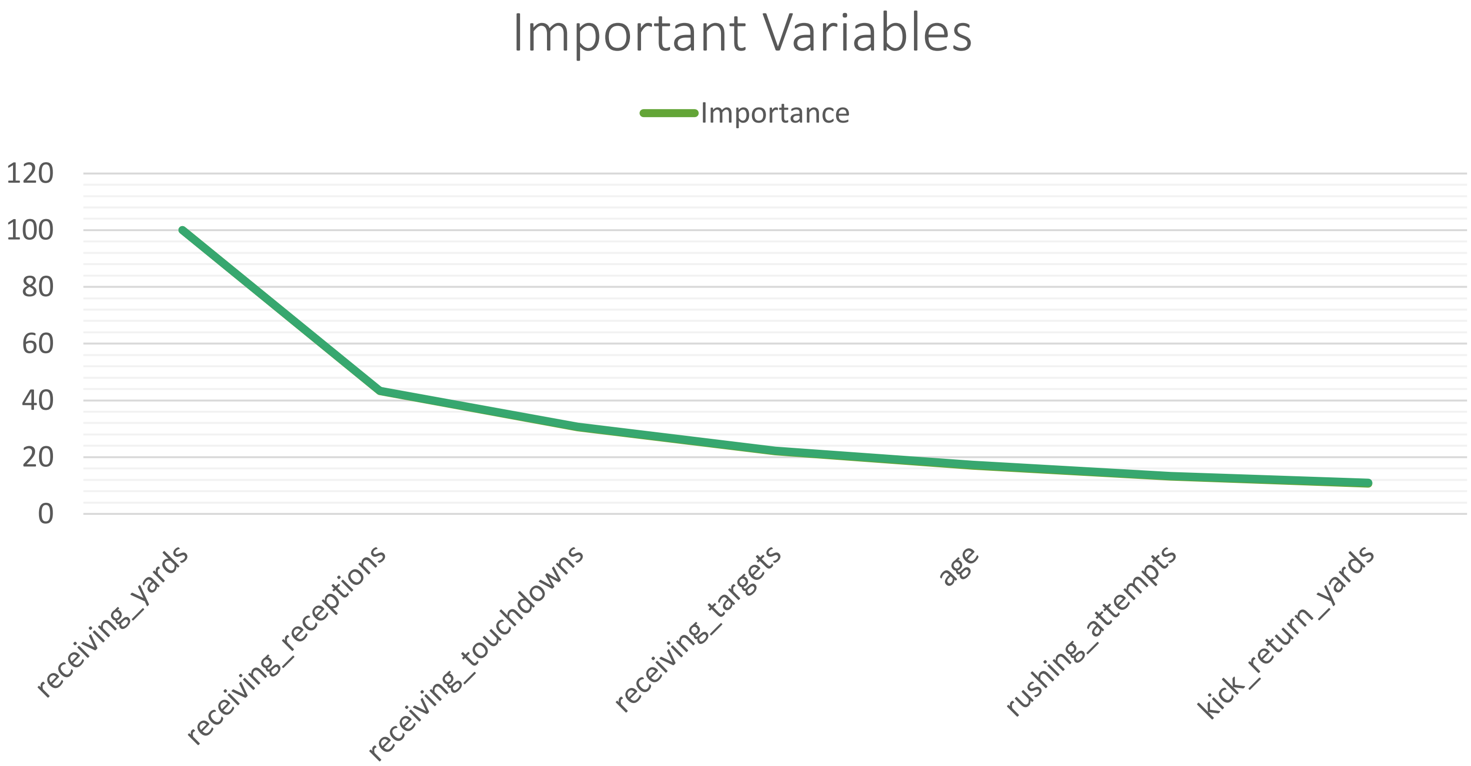
Project Goals/ Approach:

Our project predicts the following three parameters

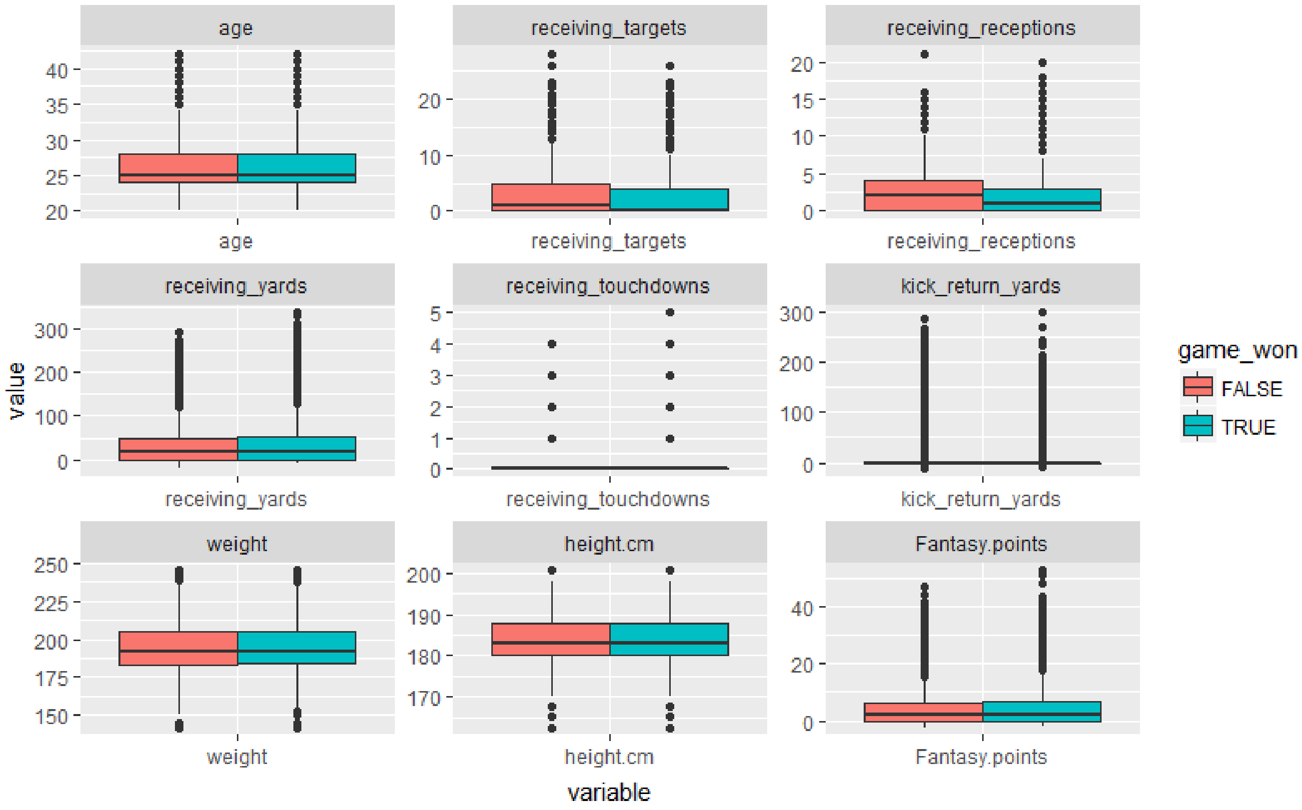
- **Fantasy points** based on played games (**regression**)
- **Fantasy score** of upcoming player's based on their college (**clustering**)
- **Winner of game** based on the given game parameters (**classification**)

Data preparation

```
'data.frame':	99362 obs. of  48 variables:
 $ player_id                    : num  5 5 5 5 5 5 5 5 5 5 ...
 $ year                         : num  2015 2015 2015 2015 2015 ...
 $ game_number                  : num  5 6 7 9 12 13 14 15 16 17 ...
 $ age                          : num  24 24 24 24 24 24 25 25 25 25 ...
 $ team                         : Factor w/ 38 levels "ARI","ATL","BAL",..:
 $ game_location                : Factor w/ 3 levels "A","H","N": 2 2 1 2 1
 $ opponent                     : Factor w/ 40 levels "ARI","ATL","BAL",..:
 $ game_won                     : Factor w/ 2 levels "FALSE","TRUE": 2 2 1
 $ player_team_score            : num  24 27 10 16 27 28 30 8 13 35 ...
 $ opponent_score               : num  10 20 29 18 23 7 20 38 20 18 ...
 $ passing_attempts             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ passing_completions          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ passing_yards                : num  0 0 0 0 0 0 0 0 0 0 ...
 $ passing_rating               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ passing_touchdowns           : num  0 0 0 0 0 0 0 0 0 0 ...
 $ passing_interceptions        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ passing_sacks                : num  0 0 0 0 0 0 0 0 0 0 ...
 $ passing_sacks_yards_lost     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ rushing_attempts             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ rushing_yards                : num  0 0 0 0 0 0 0 0 0 0 ...
 $ rushing_touchdowns           : num  0 0 0 0 0 0 0 0 0 0 ...
 $ receiving_targets            : num  0 0 0 7 3 1 3 2 0 4 ...
 $ receiving_receptions         : num  0 0 0 4 1 0 3 1 0 2 ...
 $ receiving_yards              : num  0 0 0 57 16 0 33 5 0 14 ...
 $ receiving_touchdowns         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ kick_return_attempts         : num  0 1 1 0 0 0 0 0 0 0 ...
 $ kick_return_yards            : num  0 30 22 0 0 0 0 0 0 0 ...
 $ kick_return_touchdowns       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ punt_return_attempts         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ punt_return_yards            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ punt_return_touchdowns       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ defense_sacks                : num  0 0 0 0 0 0 0 0 0 0 ...
 $ defense_tackles              : num  0 0 0 0 0 0 0 0 0 0 ...
 $ defense_tackle_assists       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ defense_interceptions        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ defense_interception_yards   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ defense_interception_touchdowns: num  0 0 0 0 0 0 0 0 0 0 ...
 $ defense_safeties             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ point_after_attemps          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ point_after_makes            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ field_goal_attempts          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ field_goal_makes             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ punting_attempts             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ punting_yards                : num  0 0 0 0 0 0 0 0 0 0 ...
 $ punting_blocked              : num  0 0 0 0 0 0 0 0 0 0 ...
 $ weight                       : num  195 195 195 195 195 195 195 195 195
 $ height.cm                    : num  185 185 185 185 185 ...
 $ Fantasy.points               : num  0 0 0 5.7 1.6 0 3.3 0.5 0 1.4 ...
```



Data preprocessing



Preprocessing:

1. **Remove variables** that are constant or serial numbers.
2. Discarded 4 variables that had **missing values** for more than almost 50% of data.
3. **Filtering (dplyr):**

Age >35 – (185 records)
Receiving targets >15 – (293 records)
Receiving_receptions >11 (100 records)
Receiving_yards>200 (126 records)
Receiving_touchdowns >2 (240 records)
Kick_return_yards >180 (121 records)
Weight <150 (121 records), Height.cm >200 (80 records)

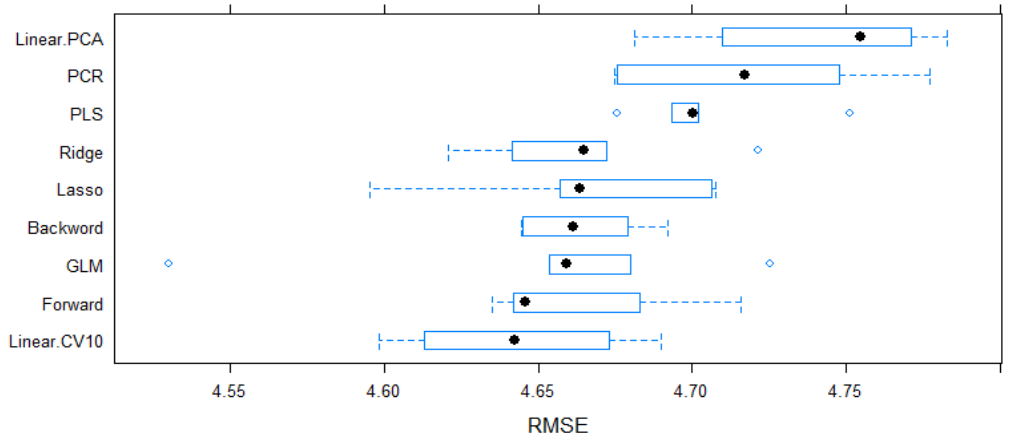
4. **Removing correlated variables:**

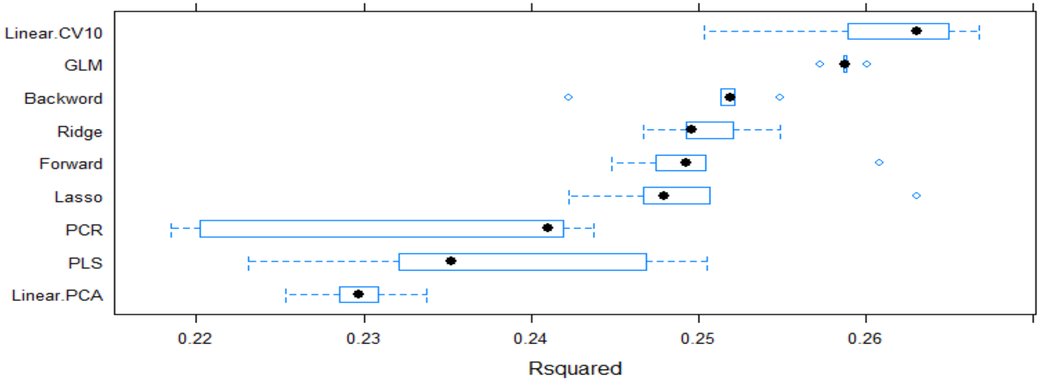
Highly correlated variables were removed by putting a threshold of more than 0.65

5. **Removing near zero variance**

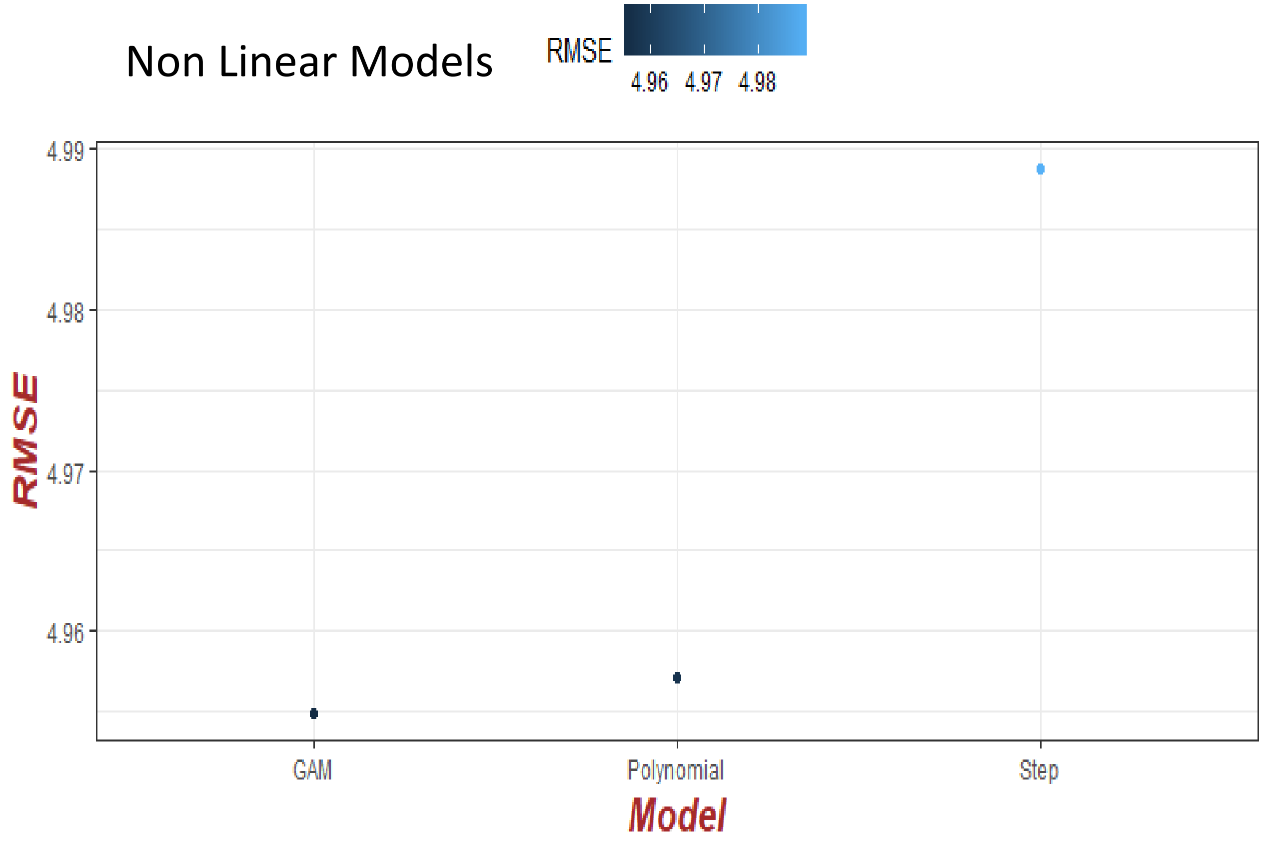
Regression Modeling

Linear Models





Non Linear Models

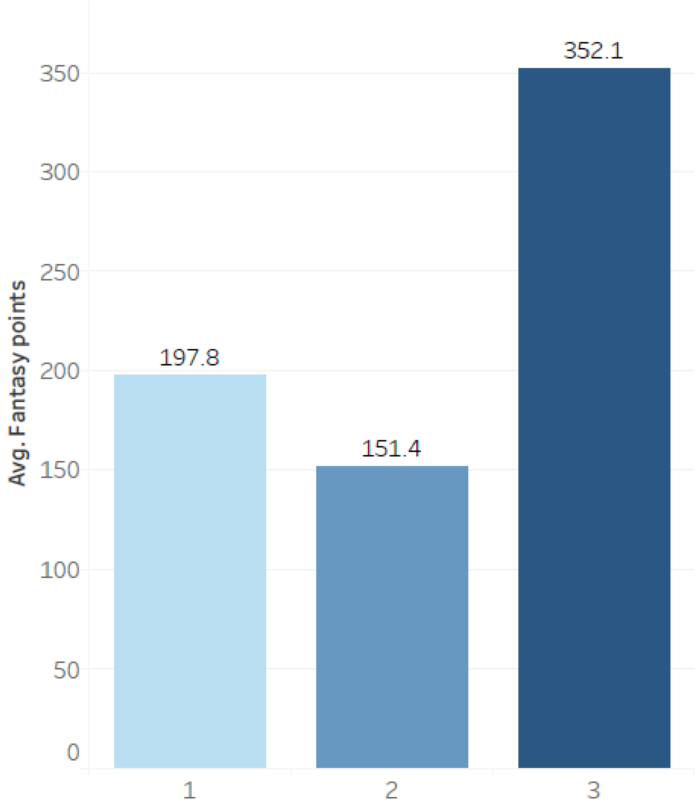


Clustering Analysis

College Names

1		2		3
Abilene Christian	Beloit	Alcorn St.	Tennessee	Arkansas
Acadia (Canada)	Bethune-Cookman	Arizona	Utah St.	Clemson
Adams St.	Bishop	Arizona St.	Washington	Florida St.
Air Force	Boise St.	Auburn	West Texas A&M;	Fresno St.
Akron	Boston Col.	Clemson		James Madison
Ala-Birmingham	Boston Univ.	Colorado		La-Monroe
Alabama	Bowling Green	Colorado St.		Minnesota
Alabama A&M;	British Columbia	Georgia		Nevada
Alabama St.	Brown	Houston		Northwestern
Albany State (GA)	Buffalo	Illinois		Notre Dame
Alcorn St.	BYU	Indiana		Ohio St.
Allegheny	Cal Poly-San Luis Obispo	Iowa St.		South Florida
American Int.	Cal State-Fullerton	Kansas St.		Trinity (TX)
Amherst	Cal State-Northridge	Kentucky		USC
Anderson (IN)	California	Long Beach St.		Vanderbilt
Angelo State (TX)	California (PA)	Marshall		Washington
Appalachian St.	California-Davis	Maryland		West Texas A&M;
Arizona	California-Riverside	Mississippi St.		
Arizona St.	California-Santa Barbara	Montclair St.		
Ark-Monticello	Cameron	Murray St.		
Ark-Pine Bluff	Carson-Newman	Nicholls St.		
Arkansas	Catawba	Notre Dame		
Arkansas St.	Central Arkansas	NW Missouri St.		
Arkansas Tech	Central Florida	Oklahoma		
Auburn	Central Michigan	Oregon		
Austin	Central Missouri St.	Pittsburgh		
Austin Peay St.	Central State (OH)	Prairie View		
Baker	Cheyney	San Jose St.		
Ball St.	Chowan	SMU		
Baylor	Cincinnati	Temple		

Average Fantasy Points



Classification analysis

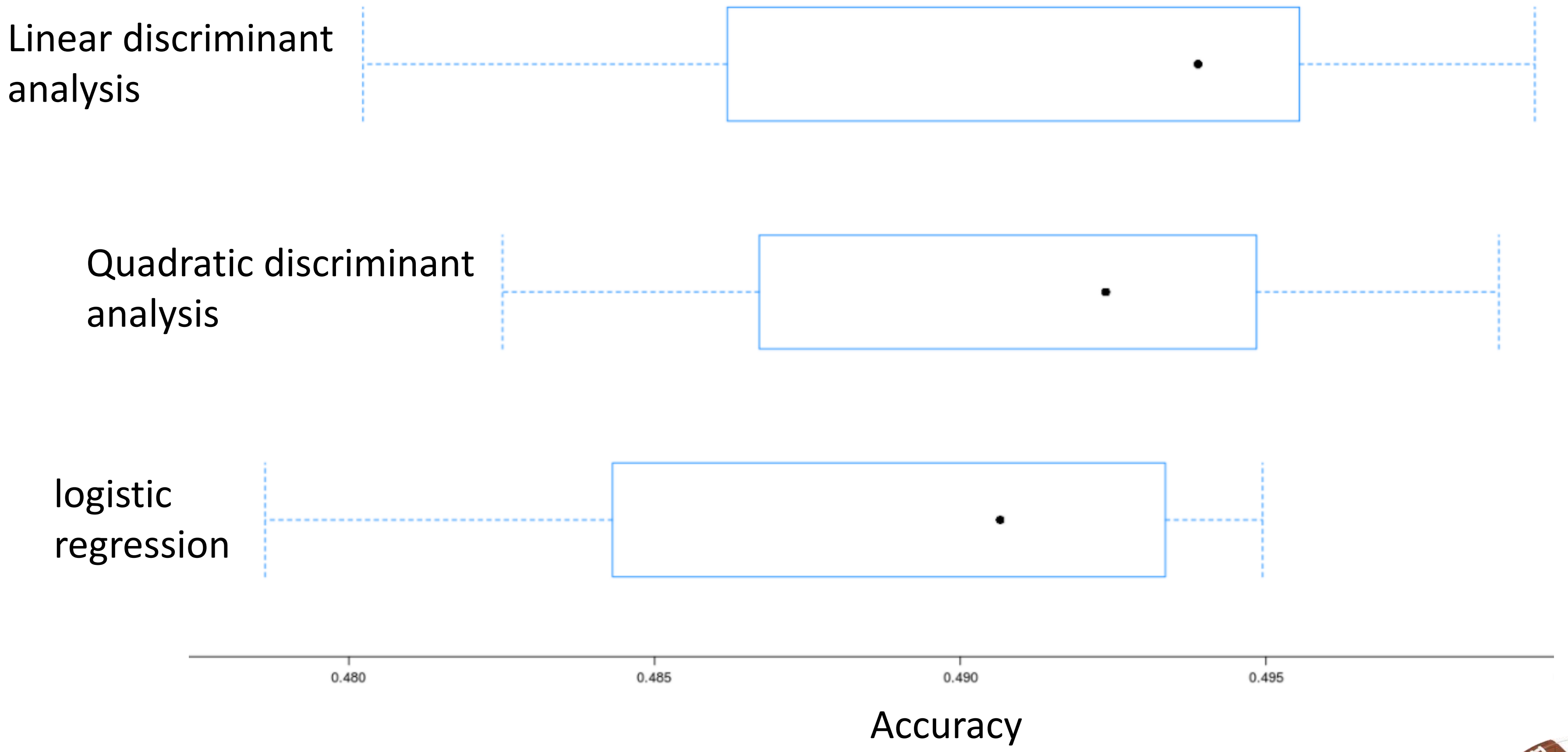
Using Classificaiton to predict Winner

	Loss	Win
Predicted loss	27958	19162
Predicted win	19033	27283
True Pos. Rate	59.33%	
True Neg. Rate	58.91%	
Correct Predictions	59.12%	

-Simple Classification using preprocessed Data Set on WR.

- With more data that we currently don’t have we could improve this models Accuracy
  - Weather
  - Injury Reports
  - Team Power Rankings
  - Team Standings

Adv. Classification Techniques W/Caret



Decision Trees and Visualization

Used all 48 preprocessed variables
- RMSE = 10
- Data Split on REC/TD only

Removed Highly Correlated Variables
- RMSE = 22
- Data Split on Game Stats

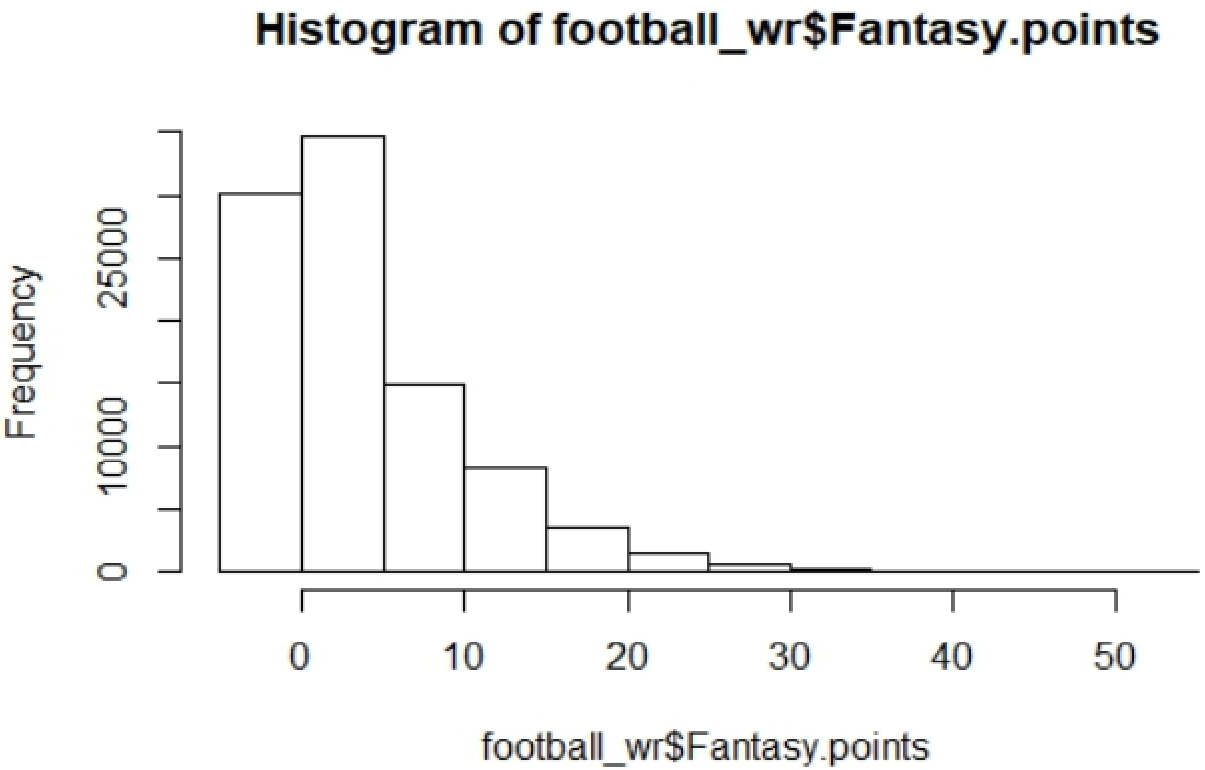
Removed Game Stats
- RMSE = 38
- Data Split on Players only

-Ultimately tells us that individual players make the most difference when you are trying to determine fantasy points.

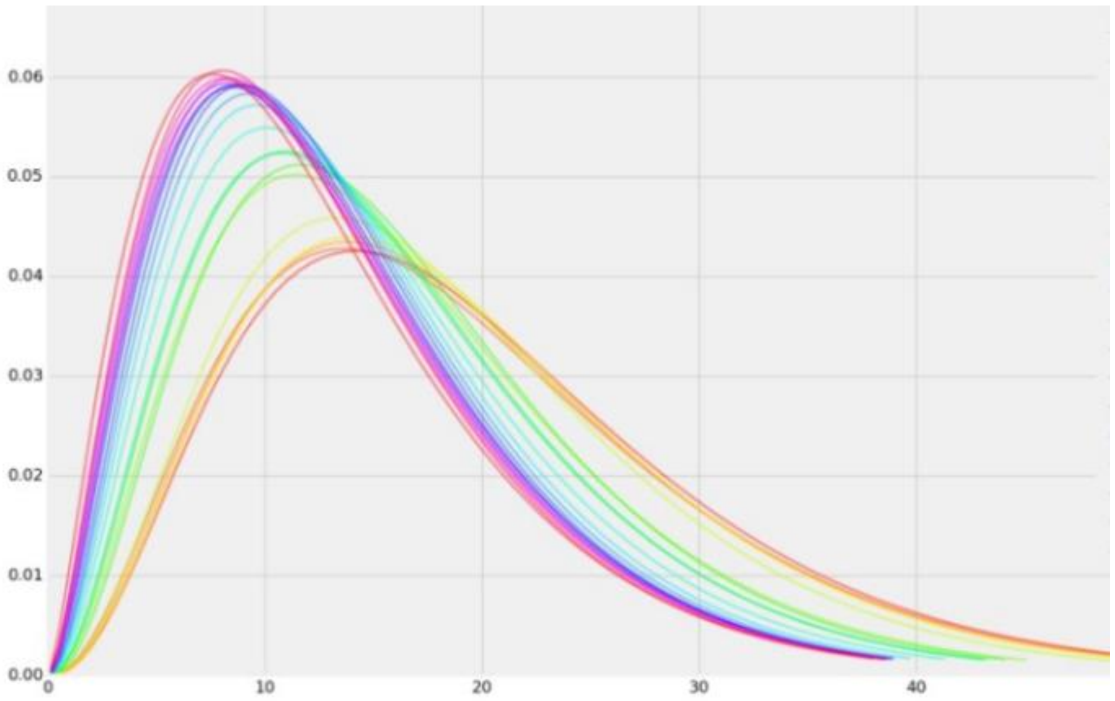
-Each player has a **probability distribution of points** they are likely to score

-Not a great model for predicting, but did give insights on data

Probabilities of Fantasy Points



Top 20 Player Distribution



Key Findings

Objective 1: **Fantasy points** based on played games (**regression**)

Status: Current model accuracy is around **30%**, we plan to improve it to around 50% by following the recommendations discussed later

Objective 2: **Fantasy points** of upcoming player's based on their college (**clustering**)

Status: Successful estimate of player performance given the college

Objective 3: **Winner of game** based on the given game parameters (**classification**)

Status: With about 60% accuracy we can predict who will win, but after we incorporate more data, we can improve this model

Difficulties in predicting sports outcomes

- Sports outcomes mix of skills and good and bad luck (Paper by Raquel Aoki et al.)

  → in between pure skills (Chess) and pure luck (Lottery)
- Despite the money involved, there are no accurate prediction methods (and may never be found)

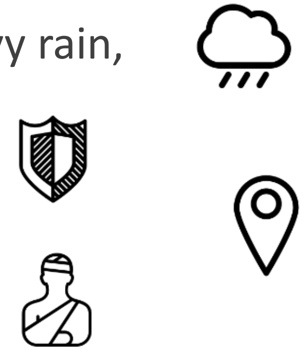
- Research found that by removing
  - 50% of teams in NBA or
  - 20% of teams for soccer

  ... makes a season a random tournament

- Teams favored by bettors win
  - 50% of the time in soccer
  - 60% of the time in baseball
  - 70% of the time in football & basketball

Recommendation and future work

- Improve the regression model accuracy using Random Forest & ensembling techniques
- We estimate that the existing variables are insufficient to explain the variability in data
- Incorporate additional variables into prediction model like
  - Weather conditions for game time (heavy rain, wet, dry/ temperature/ humidity)
  - Opponents performance in defense
  - Locations (Altitude)
  - Injuries



- Make a model for more global fantasy point tendency from historical data for players performance. (e.g. running backs performance drop)

Thank You!