Predicting the NFL Draft Order based on team statistics

Greg Vargas

2/9/21

Springboard Data Science Capstone

https://www.ncaa.org/about/resources/research/football-probability-competing-beyond-high-school

Odds of playing in the NFL

7.3 %

- 1,006,013 High School players

1.6%

- 73,712 NCAA players
- Approximately 16,380 players eligible for NFL Draft

15%

- 254 Draft picks
- Overall median ‘percentage of games started’ by players selected in 2010 NFL Draft is 15%

https://www.ncaa.org/about/resources/research/football-probability-competing-beyond-high-school
https://www.forbes.com/sites/prishe/2015/05/22/tracking-nfl-draft-efficiency-how-contingent-is-success-to-draft-position/?sh=8ec1a457495b

Can the team's statistics predict which the position order of the next draft?

- By scraping public data, will the statistics of the past season shed light on who the team will draft the next year?
- Without getting pulled in by the news of stand out college football players, can the team pick number or position be determined by regression or classification?

Webscraping methodology

Team year summary

								Points				Top Players				Off Rank		Def Rank		Overall Rank				Simple Rating System				
Year	Lg	Tm	W	L	T	Div. Finish	Playoffs	PF	PA	PD	Coaches	AV	Passer	Rusher	Receiver	Pts	Yds	Pts	Yds	T/G	Pts±	Yds±	out of	MoV	SoS	SRS	OSRS	DSRS
2020	NFL	Arizona Cardinals	8	8	0	3rd of 4		410	367	43	Kingsbury	Murray	Murray	Drake	Hopkins	13	6	12	13	17	13	11	32	2.7	-0.1	2.6	1.5	1.0

		Tot Yds & TO						Passing							Rushing					Penalties						Average Drive				
Player	PF	Yds	Ply	Y/P	TO	FL	1stD	Cmp	Att	Yds	TD	Int	NY/A	1stD	Att	Yds	TD	Y/A	1stD	Pen	Yds	1stPy	#Dr	Sc%	TO%	Start	Time	Plays	Yds	Pts
Team Stats	410	6153	1083	5.7	21	8	381	387	575	3916	27	13	6.5	211	479	2237	22	4.7	136	113	868	34	179	40.2	11.7	Own 29.1	2:36	6.22	34.4	2.30
Opp. Stats	367	5631	1054	5.3	21	10	363	365	570	3623	26	11	5.9	207	436	2008	13	4.6	118	104	841	38	172	37.2	10.5	Own 28.0	2:56	6.3	32.7	2.02

Team week summary

										Score		Offense					Defense					Expected Points		
Week	Day	Date				OT	Rec		Opp	Tm	Opp	1stD	TotYd	PassY	RushY	TO	1stD	TotYd	PassY	RushY	TO	Offense	Defense	Sp. Tms
1	Sun	September 13	4:25PM ET	boxscore	W		1-0	@	San Francisco 49ers	24	20	29	404	224	180	1	18	366	243	123		6.46	-3.88	-0.46
2	Sun	September 20	4:05PM ET	boxscore	W		2-0		Washington Football Team	30	15	22	438	278	160	1	19	316	199	117	2	8.92	0.18	6.91

Team Starting Roster

Pos	Player	Age	Yrs	GS	Summary of Player Stats	Drafted (tm/rnd/yr)
	Offensive Starters					
QB	Kyler Murray	22	Rook	16	349 for 542, 3,722 yards, 20 td, 12 int, & 93 rushes for 544 yards and 4 td	Arizona Cardinals / 1st / 1st pick / 2019
RB	Kenyan Drake	25	3	8	123 rushes for 643 yards, 8 td, & 28 catches for 171 yards and 0 td	Miami Dolphins / 3rd / 73rd pick / 2016

Team Draft Results

							Misc					Passing					Rushing			Receiving			
Rnd	Player	Pick	Pos	Yrs	From	To	AP1	PB	St	CarAV	G	Cmp	Att	Yds	TD	Int	Att	Yds	TD	Rec	Yds	TD	College/Univ
1	Kyler Murray	1	QB	2	2019	2020	0	1	2	30	32	724	1100	7693	46	24	226	1363	15				Oklahoma
2	Byron Murphy	33	CB	2	2019	2020	0	0	1	9	31												Washington

https://www.pro-football-reference.com/teams/crd/2019.htm

Scraped data was put into an SQLite Database

Team Season Data
- Examples:
- Margin of Victory/Strength of Schedule
- Wins/Losses
- Aggregate Season yardage

Team Weekly Data
- Examples:
- Weekly Opponent
- Weekly Game Stats

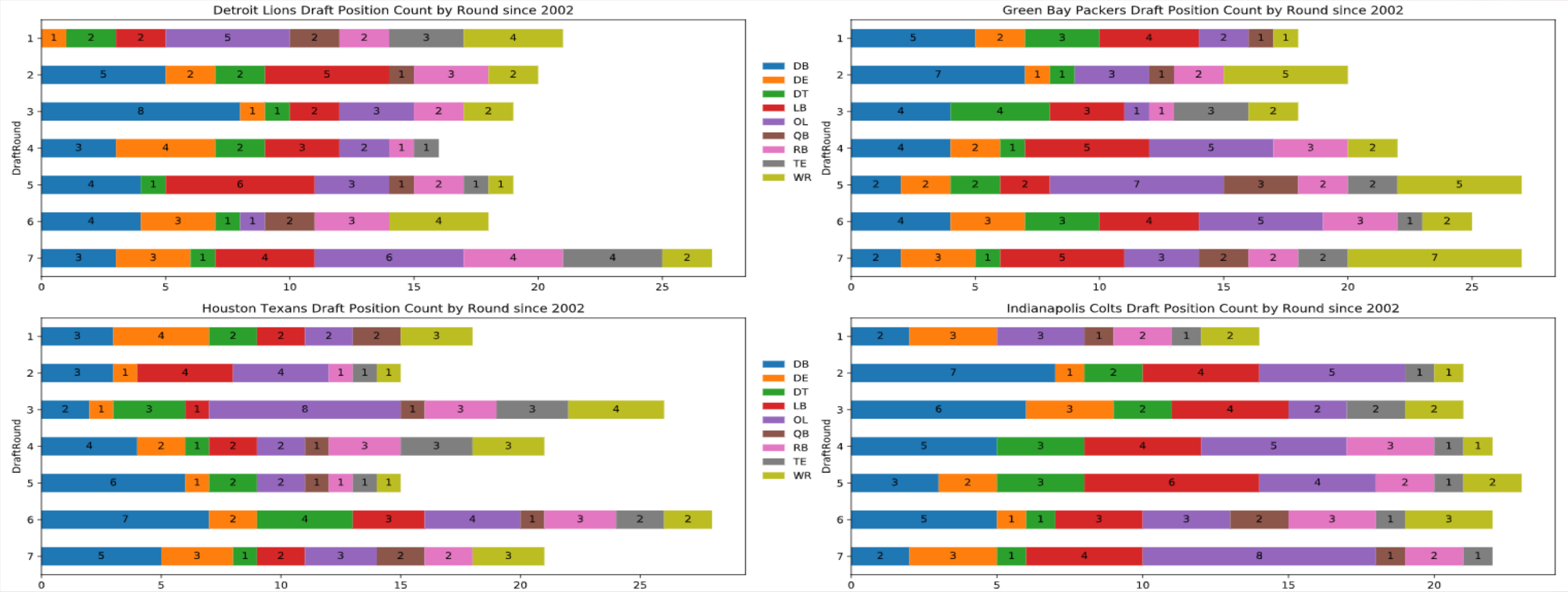
Starter Data
- Examples:
- Starting Position
- Years and Games as a starter

Draft Data
- Examples:
- Draft Position
- Draft Round Pick
- Team pick number

SQLite Database

Using Machine Learning to capture team Draft Trends

- Can a Machine Learning Model predict capture the nuance of each team’s drafting strategy by looking into historical data?



Framing of target decides type of ML model

Regression

- What round will the team select a Quarterback? What round will a Defensive Back be chosen?
- Models used:
  - Linear Regression
  - Regularized Regression
    - Ridge/Lasso/Elastic Net
  - Random Forest Regression

Classification

- If it is the team's second pick, given the stats from last year, which position will they draft?
- Models used:
  - Random Forest Classifier
  - Support Vector Classifier

Reasons for Choosing Regression Models

- Linear Regression
  - Serves as the baseline of a simple model tackling the problem
- Regularized Regression
  - The size of the coefficients, as well as the magnitude of the error term, are penalized
  - Complex models are discouraged, primarily to discourage overfitting.
- Random Forest Regression
  - Not prone to overfitting by pruning the number of estimators

https://www.statisticshowto.com/regularized-regression/
https://builtin.com/data-science/random-forest-algorithm

Reasons for Choosing Classification Models

- Random Forest Classifier
  - Can natively handle categorical variables, so there is no need to one hot encode the categorical features with high cardinality
  - Uses subsets of the data, so good for problems with high dimensionality
- Support Vector Classifier
  - SVC is effective in high dimensional spaces
  - SVC is effective in cases where the number of dimensions is greater than the number of samples

https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706

Models will be run on multiple datasets

- Original Datasets created from Exploratory Data Analysis
  - Team Aggregated Season Statistics
    - Second dataset with average position ranks included
  - Team Week Statistics
    - Second dataset with average positional ranks included
- Additional Datasets created to improve performance
  - Dataset with categorical variables bucketized to reduce cardinality
  - Dataset with coaching information removed to remove main source of categorical cardinality

Performance of Regression Models

Regression Models

By looking at the Mean Absolute Error, we can see that on average, the model missed the draft order selection by 2 picks

	R2	MSE	RMSE	MAE	best regressor
year	0.003490	6.803684	2.608387	2.203176	ElasticNet(alpha=1.0, copy_X=True, fit_interce...
yearAV	0.000286	6.996596	2.645108	2.220828	Lasso(alpha=0.15264179671752318, copy_X=True, ...
week	0.027038	6.751551	2.598375	2.174834	Lasso(alpha=0.0009540954763499944, copy_X=True...
weekAV	0.038667	6.563252	2.561884	2.143305	Lasso(alpha=0.0011513953993264468, copy_X=True...
yearnocoach	0.010850	6.734904	2.595169	2.183651	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...
yearnocoachAV	0.003437	6.907644	2.628240	2.184574	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...

Regression Models with PCA

Principal Component Analysis was used to remove any collinearity between features. Even with dimensional reduction, MAE is not improved

	PCA components	R2	MSE	RMSE	MAE	best regressor
year	50	0.006157	6.564303	2.562090	2.148639	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...
yearAV	50	-0.000032	7.133128	2.670792	2.246177	Lasso(alpha=0.0011513953993264468, copy_X=True...
week	50	0.012887	6.832590	2.613922	2.188608	Ridge(alpha=0.007543120063354615, copy_X=True,...
weekAV	50	0.016964	6.642123	2.577232	2.155551	Lasso(alpha=0.004291934260128779, copy_X=True,...
yearnocoach	50	-0.005583	7.053212	2.655788	2.221354	Ridge(alpha=0.6866488450042998, copy_X=True, f...
yearnocoachAV	2	-0.004378	7.119864	2.668307	2.228804	Ridge(alpha=0.15264179671752318, copy_X=True, ...

Performance of Classification Models

- With such low scores in the year and yearAV datasets, the week dataset was chosen to see how it compared to a dummy classifier. This was done to see if the trained model performed better than chance

	PCA components	Accuracy	f1
year	2	0.205402	0.084775
yearAV	15	0.201543	0.067612
week	15	0.269116	0.241012
weekAV	15	0.268848	0.251925
yearnocoach	30	0.196873	0.065078
yearnocoachAV	30	0.200579	0.067021

How does the best dataset compare to a dummy classifier?

- We can see an increase in the accuracy of the model over the best dummy classifier
- The train and test scores were similar, which shows that the model is not overfitting

```
Cross-Validation best parameters:  SVC(C=1, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='scale', kernel='poly',
    max_iter=-1, probability=False, random_state=None, shrinking=True,
    tol=0.001, verbose=False)
Cross-validation mean test score: 0.29585346113896405
                  score
stratified     0.133555
most_frequent  0.201656
uniform        0.112306
      PCA components  Train Accuracy  Test Accuracy  Train F1   Test F1
week              30         0.37681       0.299667  0.352613  0.273753
[[2725  165  147  450  767   26  104   59  283]
 [ 784  413   34  198  465   15   72   40  199]
 [ 721   86  271  194  397   14   62   27  190]
 [ 921  106   94  826  635    4   94   54  228]
 [1166  148   76  397 1629   34  106   44  231]
 [ 389   45   45  179  257   53   27   34  111]
 [ 811  127   70  227  443   20  254   31  212]
 [ 558   77   44  134  236   12   71  140  102]
 [1121  116   96  331  515   23   72   40  712]]
              precision    recall  f1-score   support

          DB       0.30      0.58      0.39      4726
          DE       0.32      0.19      0.24      2220
          DT       0.31      0.14      0.19      1962
          LB       0.28      0.28      0.28      2962
          OL       0.30      0.43      0.36      3831
          QB       0.26      0.05      0.08      1140
          RB       0.29      0.12      0.17      2195
          TE       0.30      0.10      0.15      1374
          WR       0.31      0.24      0.27      3026

    accuracy                           0.30     23436
   macro avg       0.30      0.23      0.24     23436
weighted avg       0.30      0.30      0.27     23436
```

Takeaways/Further Research

- The data does not contain the information needed to have predictive power.
  - What data is needed to add the desired predictive power?
- Minimal feature engineering was done, with additional engineering potentially increasing model accuracy
- Looking into the Problem statement, to see if the correct problem is defined