Predicting Drafted Quarterbacks

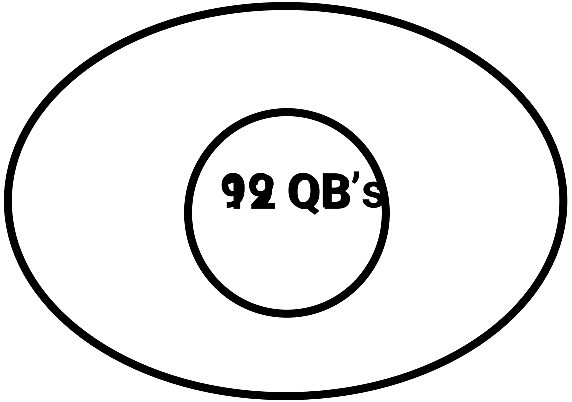
Using Machine Learning

Connor Levenson
Mihir Arya
Kennard Peters
Joe Kinderman

Apr 17

Inspiration

- It is inefficient for sports agents to study every player
- Use NCAA Statistics to determine players most likely to be drafted



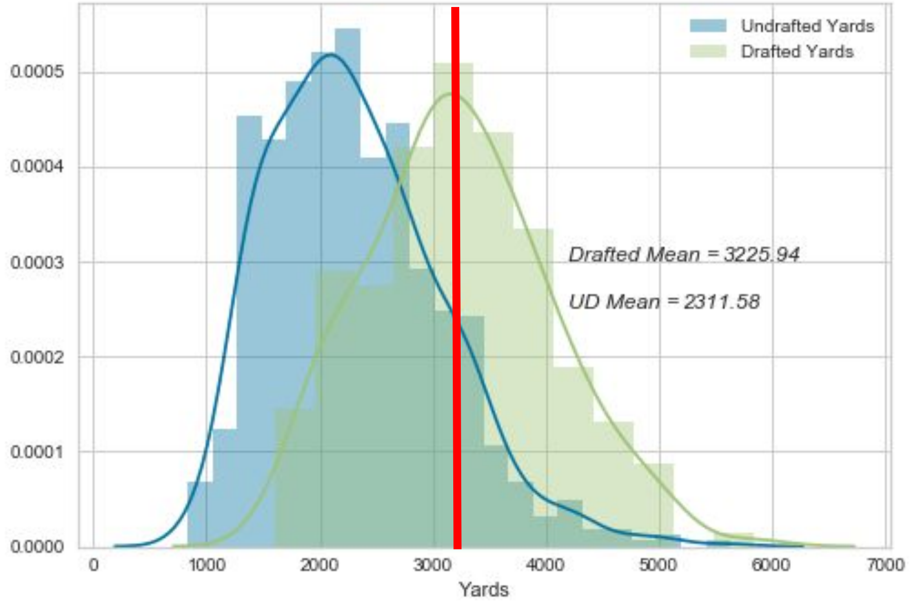
35,000 Pass Attempts

Data Gathering

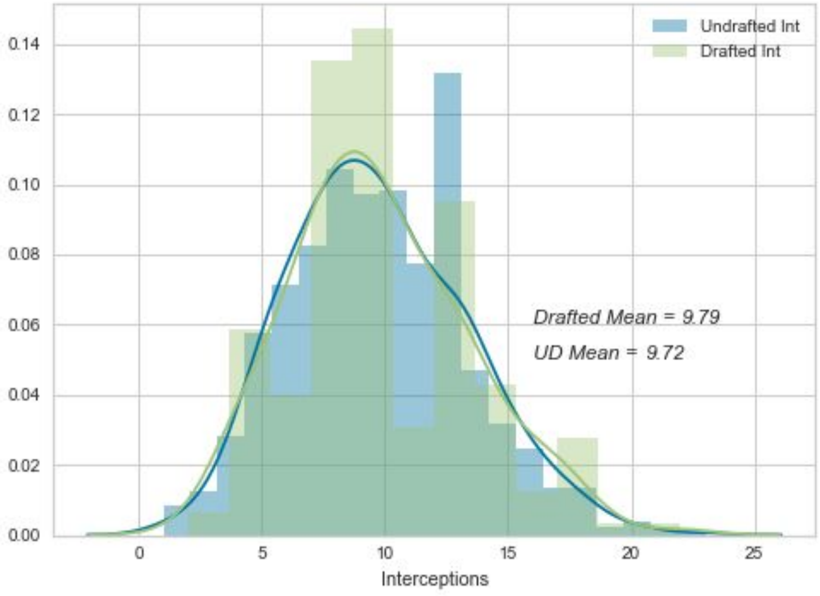
Rk	Player	School	Conf	G	Cmp	Att	Pct	Yds	Y/A	AY/A	TD	Int	Rate	TDC	Att/Int	Yds/G
1	Baker Mayfield	Oklahoma	Big 12	14	285	404	70.5	4627	11.5	12.9	43	6	198.9	7.5	67.33	330.5
3	Mason Rudolph	Oklahoma State	Big 12	13	318	489	65.0	4904	10.0	10.7	37	9	170.6	4.9	54.33	377.2
5	Logan Woodside	Toledo	MAC	14	264	411	64.2	3882	9.4	9.9	28	8	162.2	4.4	51.37	277.2
9	Danny Etling	LSU	SEC	13	165	275	60.0	2463	9.0	9.8	16	2	153.0	3.5	137.5	189.5
11	Sam Darnold	USC	Pac-12	14	303	480	63.1	4143	8.6	8.5	26	13	148.1	3.4	36.92	295.9

Data Exploration

CFB Yards



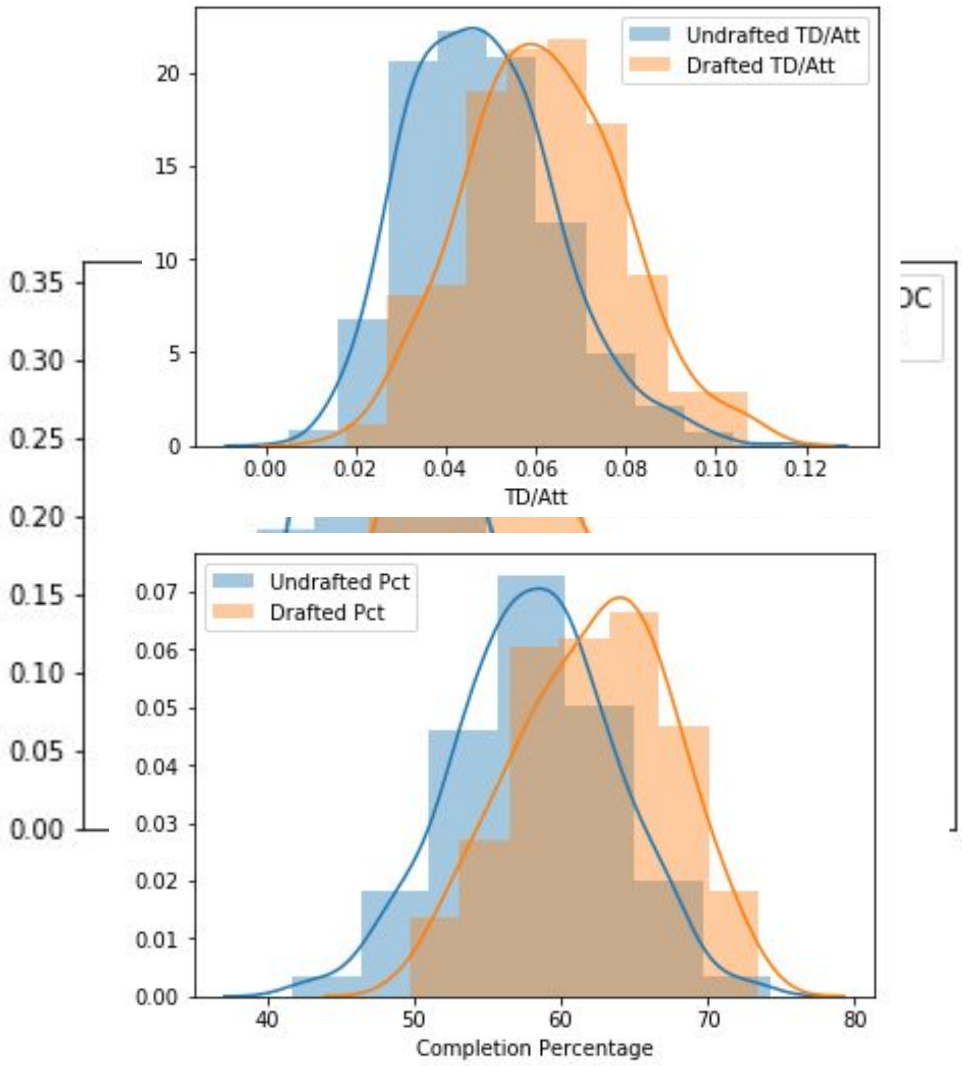
CFB Interceptions



Data Transformation

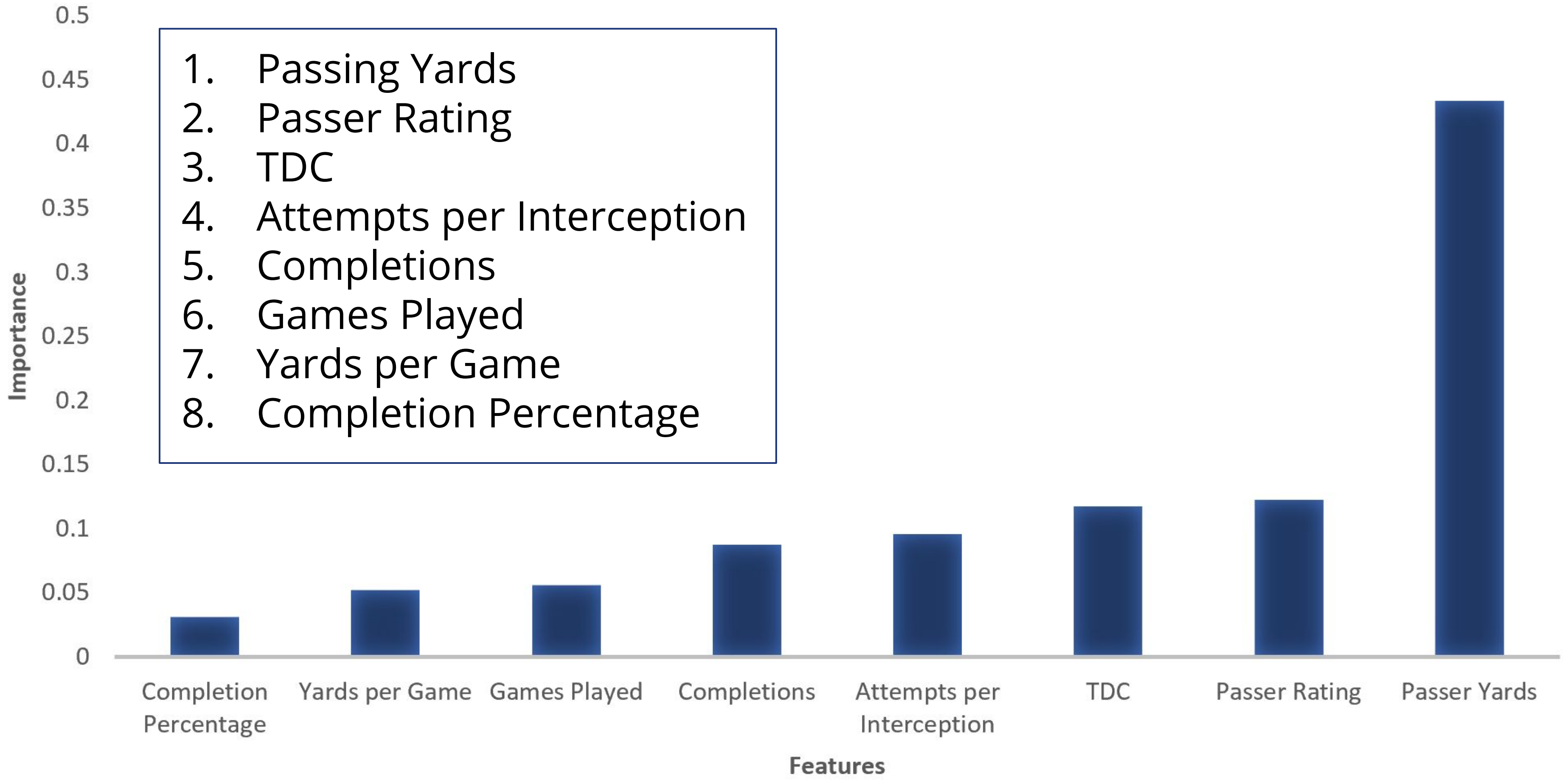
- Touchdowns per attempt indicates scoring ability
- Completion percentage signifies consistency
  - How do we combine these into one metric?

$$TDC = \frac{TD * Cmp}{(Att)^{2}}$$



FEATURE IMPORTANCE

1. Passing Yards
2. Passer Rating
3. TDC
4. Attempts per Interception
5. Completions
6. Games Played
7. Yards per Game
8. Completion Percentage



Model Selection

- Decision tree
- Random forest
- K-Nearest Neighbor
- Gradient Boosted Random Forest

Cross-Validation

Classifier	Mean weighted-F1
Decision Tree	0.79
Random Forest	0.83
K Nearest Neighbors	0.83
Gradient Boosting	0.84

Results

Classifier	Average Precision	Average Recall	F1 Score
Decision Tree	0.79	0.74	0.74
Random Forest	0.83	0.83	0.82
K Nearest Neighbors	0.83	0.82	0.82
Gradient Boosted Random Forest	0.84	0.83	0.83

2019 Draft

April

25

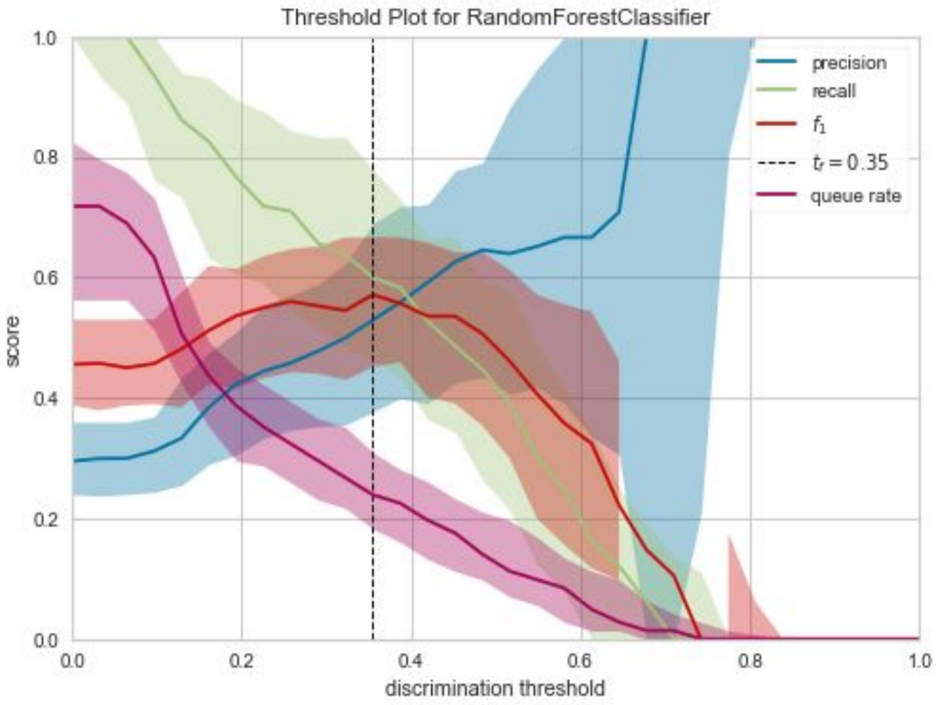
Player	College
Drew Lock	Missouri
Kyler Murray	Oklahoma
Dwayne Haskins	Ohio State
Ryan Finley	NC State
Will Grier	West Virginia
Brett Rypien	Boise State
Gardner Minshew	Washington State
Justice Hansen	Arkansas State
Jordan Ta’amu	Ole Miss
David Blough	Purdue

Questions?

Appendix

Random Forest

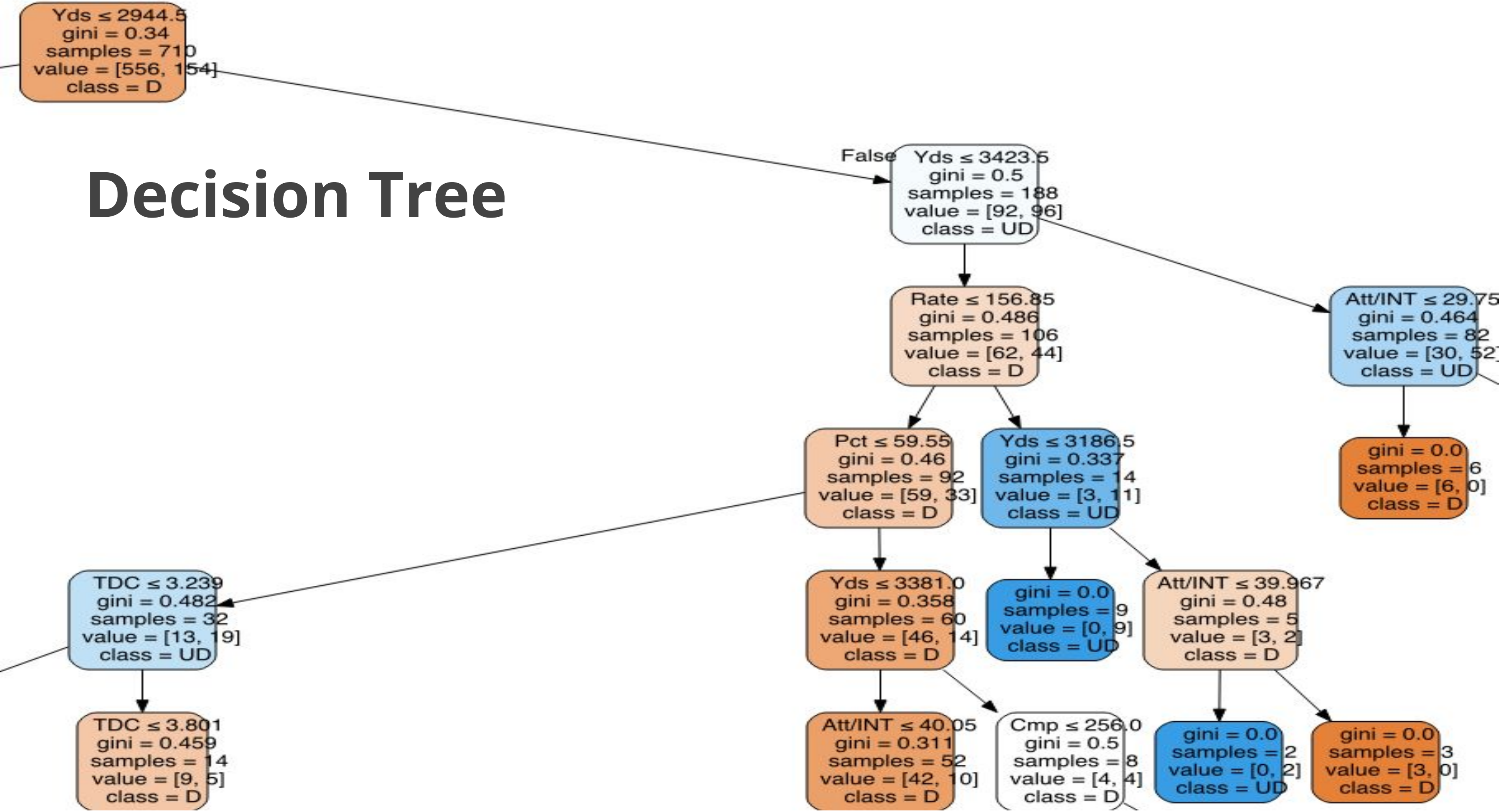
- Tried to create a lot of error from different places
- Cross-validation had some wacky results at first
  - Increased number of trees
  - decrease max number of features/depth to try to limit variance
- N_estimators = 20
- Max_depth = 4
- Max_features = 3



Gradient Boosted Random Forest

- Uses same bagging technique as random forest
- Does not build all trees in random forest at once
- Uses gradient boosting technique to fit on the errors
- Hyperparameters
  - Early_stopping_rounds = 10
    - Stops boosting after test accuracy decreases <u>and</u> train accuracy increases for 10 rounds
      - Prevents overfitting
  - Validated using manually split data
  - Max_depth = 3
    - More generalized

Decision Tree



Why Maximize F1?

- Agents have a fixed number of resources (a limit on the number of players they can represent)

- Representing a quarterback that does not get drafted means the opportunity cost of representing that quarterback was the ability to represent a quarterback who did get drafted

- An agent wants to minimize the amount of times he represents a quarterback that does not get drafted (in this case, false positive rate)

- And at the same time maximize the amount of times he represents a quarterback that does get drafted (true positive rate)

- Representing a quarterback who did not get drafted means he missed out on one who did