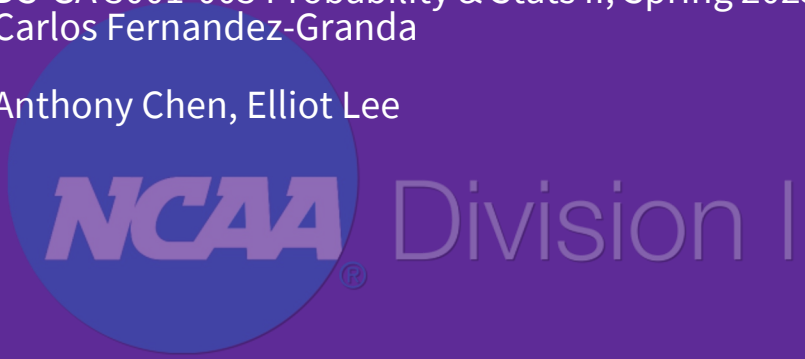
NCAA Division I College Basketball NBA Prospects

DS-GA 3001-005 Probability & Stats II, Spring 2023
Carlos Fernandez-Granda

Anthony Chen, Elliot Lee

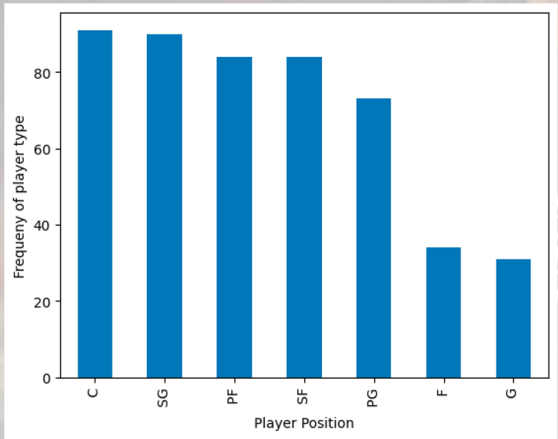


Background & Motivation

We wanted to see if we could use college basketball statistics of individual players who were selected in the **first round of the NBA Draft**, and predict whether they would be potential all-star caliber players.

For example, in these 3 images, De'Aaron Fox played for the University of Kentucky, and was drafted 5th overall in the 2017 Draft class by the Sacramento Kings, and was voted the most clutch player (most points in final minutes of the fourth quarter of a game) in the 2022-23 NBA season (6 seasons).

Dataset

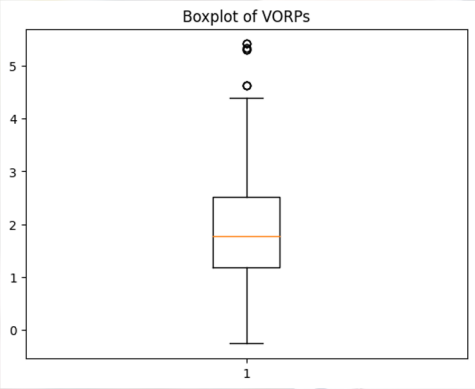


- Data was collected from BasketballReference, SportsReference, and through running nba_api, an API client for the NBA's website of players drafted from 2000-2018 in the first round
- Some players did not play in college, such as *Luka Dončić* (Euroleague) and *Lebron James* (High School), we left them out of our dataset.
- *Target value:* **VORP** (Value over replacement player)
    - This divided by number of seasons played gives us **Average Seasonal VORP**

Fun Fact: NYU hasn't been D1 since 1971

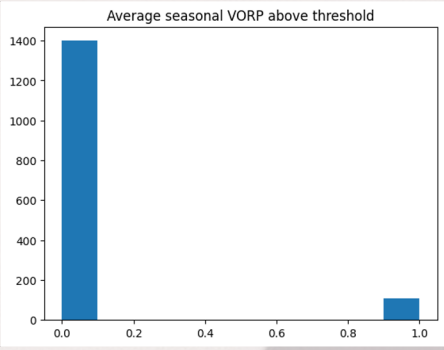
Dataset

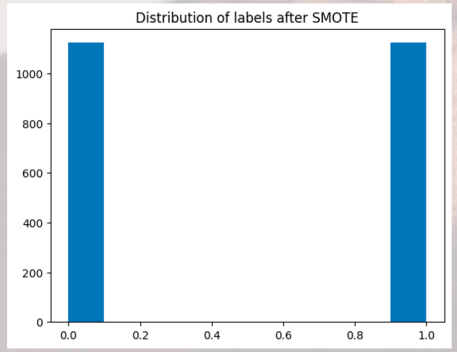
- We imputed missing basketball statistics with the **median**
  - Avoids diminishing the quality of our dataset due to skewness of outliers.
  - By conference, and then position.
- We also created our own binary labels (Good player:1, Bad player:0) based on an Average Seasonal VORP threshold equivalent to the average all-star player Seasonal VORP in our dataset.



Methods & Models

- **PCA** transform on highly correlated basketball statistics in our dataset to address multicollinearity
- **Linear Regression** (removed insignificant features with p>0.05, and applied regularization)
- **Random Forest** (ensemble method to minimize variance of predictions through averaging decision trees + pruning to add some regularization)
- Upsampling through SMOTE to combat data imbalance





Results & Evaluation

Linear Regression was evaluated using:

- Mean Absolute Error (MAE)
- R-squared

Random Forest was evaluated using:

- Precision
- AUC-ROC
- F1-score

Results & Evaluation - OLS

- MAE: 0.4639
- R-Squared: 0.119

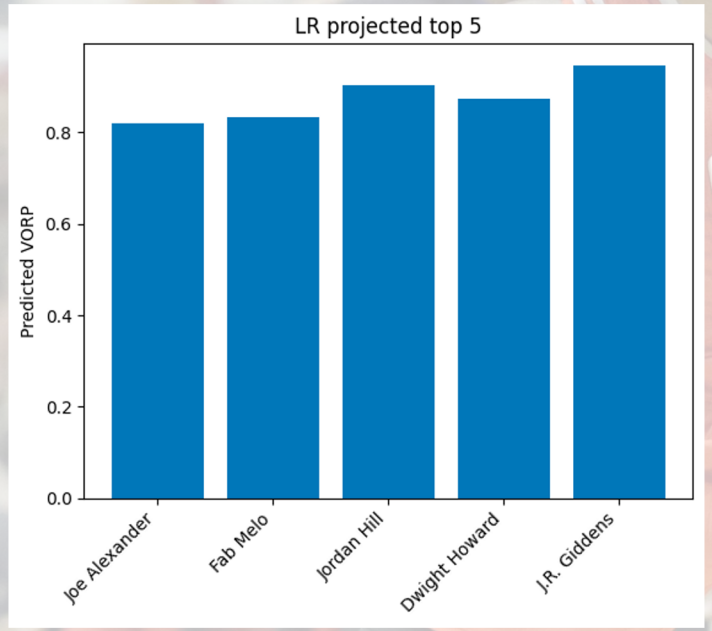
OLS Regression Results

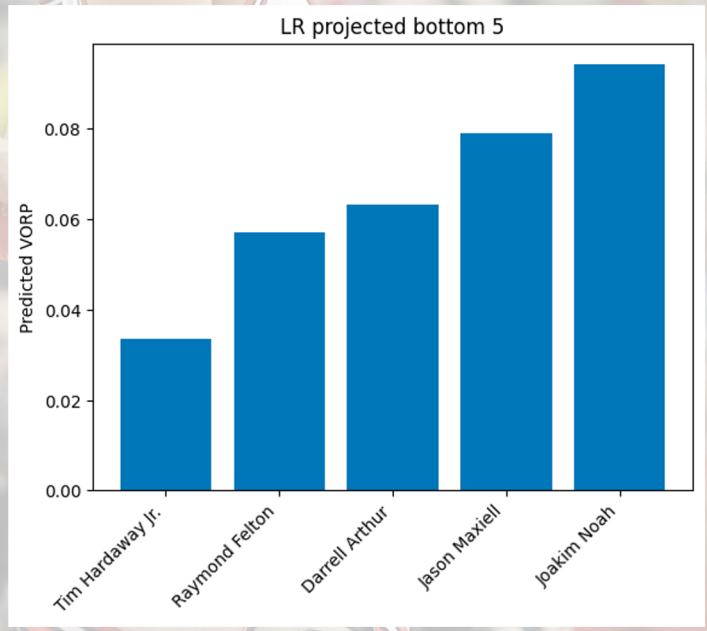
Dep. Variable:	Caliber	**R-squared (uncentered):**	0.119
Model:	OLS	**Adj. R-squared (uncentered):**	0.115
Method:	Least Squares	**F-statistic:**	30.25
Date:	Sat, 06 May 2023	**Prob (F-statistic):**	3.33e-55
Time:	22:46:21	**Log-Likelihood:**	-2266.0
No. Observations:	2246	**AIC:**	4552.
Df Residuals:	2236	**BIC:**	4609.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
pca_guard	-0.1069	0.019	-5.634	0.000	-0.144	-0.070
pca_big	0.0055	0.013	0.425	0.671	-0.020	0.031
pca_wing	-0.0300	0.012	-2.605	0.009	-0.053	-0.007
scoring_attr	0.0348	0.015	2.371	0.018	0.006	0.064
FG%	-0.2679	0.061	-4.426	0.000	-0.387	-0.149
2P%	-0.2022	0.039	-5.241	0.000	-0.278	-0.127
3P%	-0.0701	0.018	-3.874	0.000	-0.106	-0.035
BLK	0.0661	0.021	3.160	0.002	0.025	0.107
eFG%	0.1382	0.050	2.771	0.006	0.040	0.236
TOV%	0.0054	0.019	0.293	0.770	-0.031	0.042

Omnibus:	1089.615	**Durbin-Watson:**	0.352
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	133.317
Skew:	-0.191	**Prob(JB):**	1.12e-29
Kurtosis:	1.870	**Cond. No.**	13.9

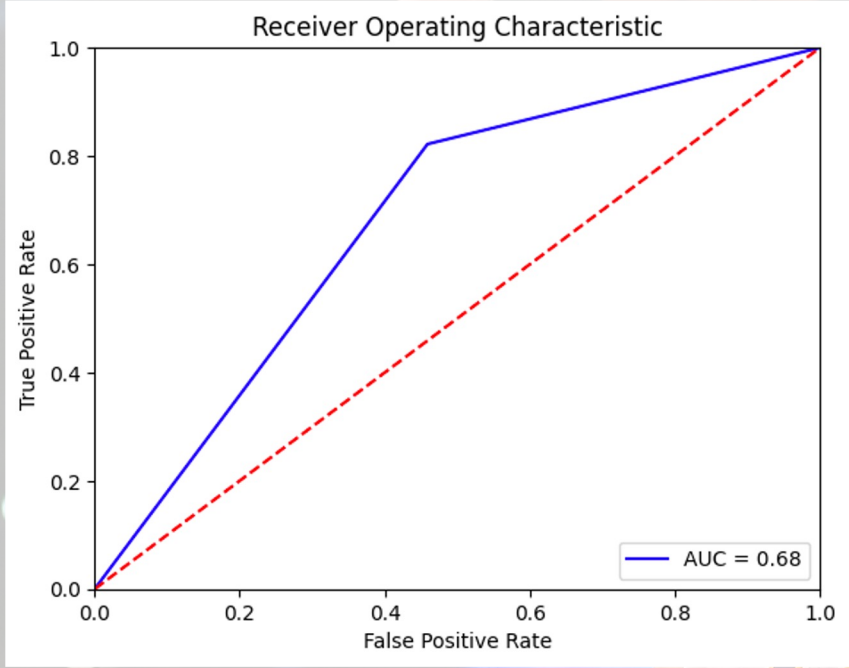
Results & Evaluation - Linear Regression



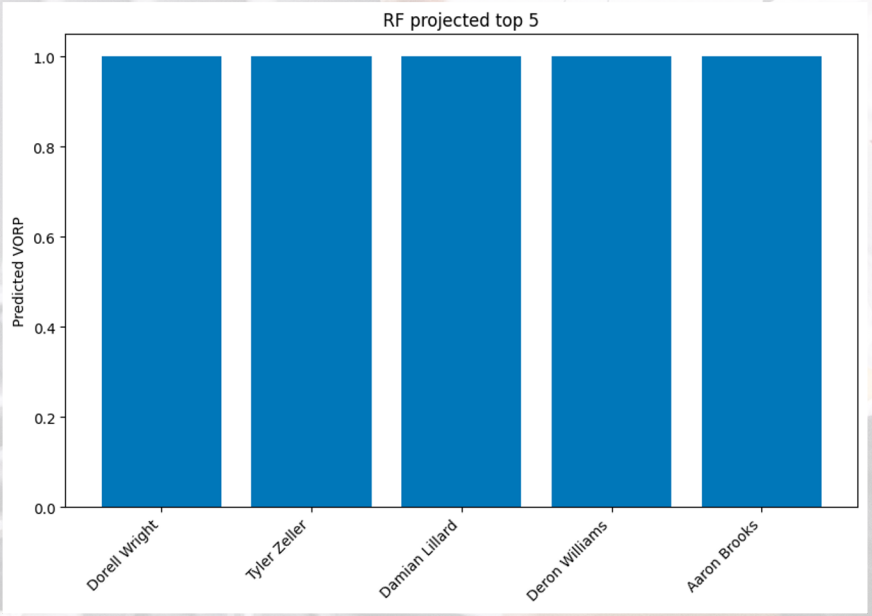


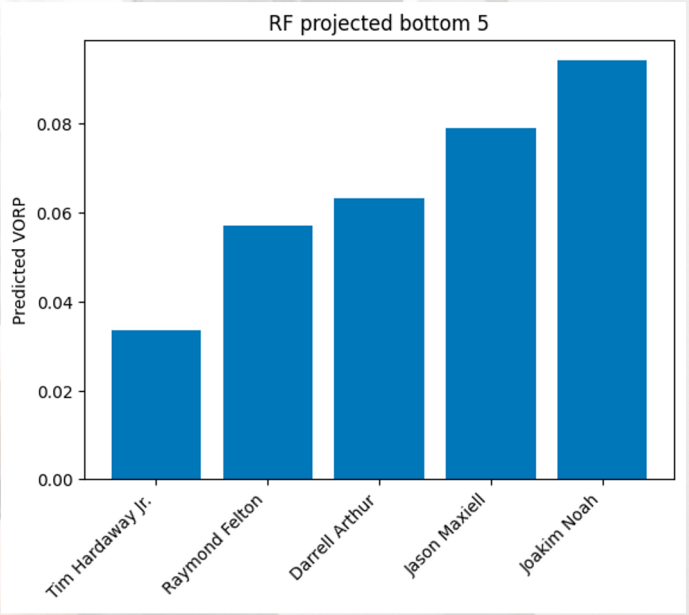
Results & Evaluation - Random Forest

- **Precision**: 0.1153
- **F1-score**: 0.1923076923076923



Results & Evaluation - Random Forest





Conclusion & Future Considerations

- In general, higher usage players that take a lot of shots tend to get drafted higher
- Sports data is hard to predict on!
  - High variance
  - Small sample size
  - Constant covariate shift

- Considerations:
  - Increase our dataset (Requires: money and/or manual scraping and annotation)
  - Gather better features (Requires granular in game metrics that are behind paywalls)

Thank you for a great semester!

Other data science project ideas: Whether a hall of fame player will finish higher education?