NBA Player Selection

By: Miriam Brem, Aidan Conley, Lucas Huang, and Ryana Rajesh

Introduction and Research Question

For our Final Project, we chose to look at the NBA Player Dataset to study how basic background characteristics relate to when players are selected in the NBA draft.

Our research question is: How are attending different U.S. colleges or entering the non-college or international paths, and draft year associated with a player's draft position, after accounting for physical characteristics (height and weight) and origin (country)? We chose this question of interest to determine how much these factors can help us determine a player's draft number before looking at amateur stats and/or film of the player.

This question is relevant because if historical patterns show that draft position is systematically related to college affiliation, draft era, and physical attributes, NBA teams may be able to improve the efficiency of the early stages of the scouting process. If we can identify these relationships, it could allow teams to allocate scouting resources more strategically, potentially saving time and money while maintaining effective talent evaluation.

Description of Key Variables & Data Context

The first of our key variables of interest include *college* which is the name of the college the player attended. Next we have *draft year* which is simply the year the player was drafted. We have *draft number* which is the number at which the player was picked in his draft round. We have player *height,* which is the height of the player measured in centimeters. We have player *weight* which is the weight of the player in kilograms. Lastly, we have *country*, the name of the country in which the player was born.

Analyses Approach

Logic for Selecting Covariates:

In an effort to select the appropriate model, we utilized the following covariate selection logic, continuing to add more and more covariates to see if we can account for more of the variance in draft position and improve our predictive power.

Model 1: $\beta_0 + \beta_1 \text{college}_i + \varepsilon_i$

We began with Model 1 as our baseline. We regressed the draft position on only the *college* variable. For this, we used “None” (players with no listed US college) as our reference group.

Model 2: $\beta_0 + \beta_1 \text{college}_i + \beta_2 \text{year}_i + \varepsilon_i$

This model was created to add *draft year* as a covariate. This was meant to account for how the NBA draft has changed over time after 1989.

Model 3: $\beta_0 + \beta_1 \text{college}_i + \beta_2 \text{year}_i + \beta_3 \text{height}_i + \beta_4 \text{weight}_i + \varepsilon_i$

The third model adds *heigh*t and *weight* to control for the player's physical size.

Model 4: $\beta_0 + \beta_1 \text{college}_i + \beta_2 \text{year}_i + \beta_3 \text{height}_i + \beta_4 \text{weight}_i + \beta_5 \text{country}_i + \varepsilon_i$

Model 4 is where we add *country* of origin as our final covariate. This model balances a good fit with interpretability.

Model 5: $\beta_0 + \beta_1 \text{college}_i + \beta_2 \text{year}_i + \beta_3 (\text{college}_i \text{ x } \text{year}_i) + \beta_4 \text{height}_i + \beta_5 \text{weight}_i + \beta_6 \text{country}_i + \varepsilon_i$

This model allows each college to have its own draft-year trend so the impact of draft year differs by college in a non-parallel lines model

Selecting a Model

In an effort to address potential overfitting and to formally select our model, we are going to compare the in-sample fit by utilizing R^2 values.

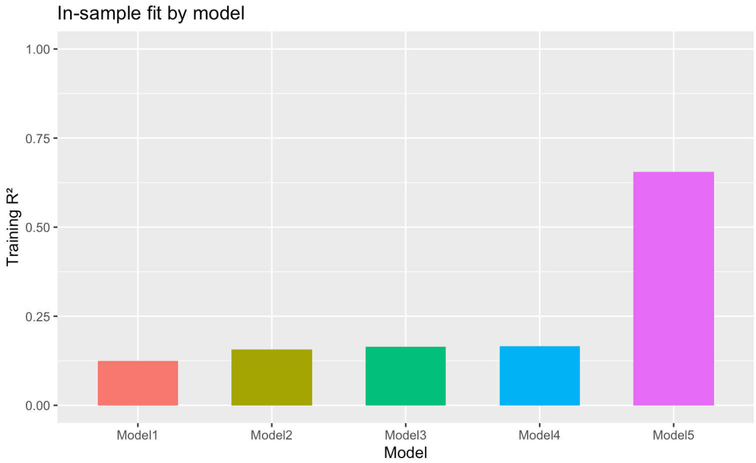


Figure 1: Training R^2 for each Model

As we can see in **Figure 1** (and Appendix 2A), models 1 through 4 all have a fairly low R^2 and Adjusted R^2 that increases slightly with the addition of new variables. Since all of these models have a relatively low R^2, they only explain a small share of draft positions variation.

On the other hand, Model 5's R^2 jumps to around 0.65, meaning it explains roughly 65% of the variation in draft position on the training data. This is much higher than any of the other models. However, this model uses thousands of parameters which indicates classic overfitting.

Thus, to assess overfitting, we used a five-fold cross-validation to assess each model's predictive performance and compared the RMSE for each model.

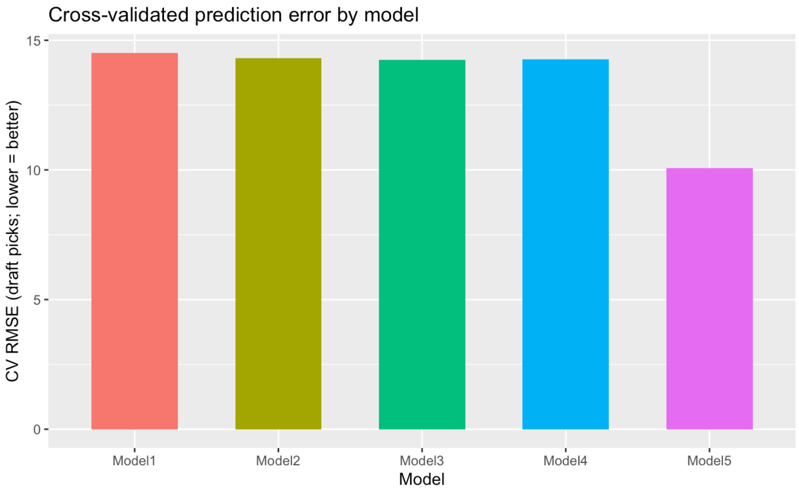


Figure 2: Cross-Validated Prediction Error by Model

As we can see in **Figure 2** (and Appendix: Table 2), most of our models have an RMSE of approximately 14, which indicates predictions by that model are off by about 14 spots on average. Values only slightly decrease from Model 1 to Model 4. This shows that adding draft year, height, weight and country only modestly improves predictive accuracy compared with the simplest college-only model.

Model 5 however, drops to an RMSE of 10, meaning it predicts roughly 4 draft positions more accurately than the simpler models.

However, it achieves this at the cost of huge complexity and instability since it does this with thousands of parameters. So, while it looks good on the training data, cross-validated error barely improves relative to the jump in complexity. We suppose that the model is too closely tailored to the sample we used and it starts learning random noise instead of real patterns. The model almost “memorizes” the dataset which makes training fit statistics look great, but would predict new seasons or different NBS samples poorly. Thus, we treat it as overfitting and use Model 4 for interpretation.

Assumptions and Diagnostics

In order to draw any conclusions using our models, we have to assess the data and see if the assumptions underlying our analyses are met.

To evaluate the validity of our results and methods, we must assess the key assumptions underlying Model 4. Specifically, we must examine independence of errors, linearity, normality of residuals, mean 0 errors, and homoscedasticity. The diagnostics below address each of these assumptions in turn:

Independence of Errors

This data is inherently non-independent. In order to draft one player first, others have to be picked second. As a result, the errors are dependent.

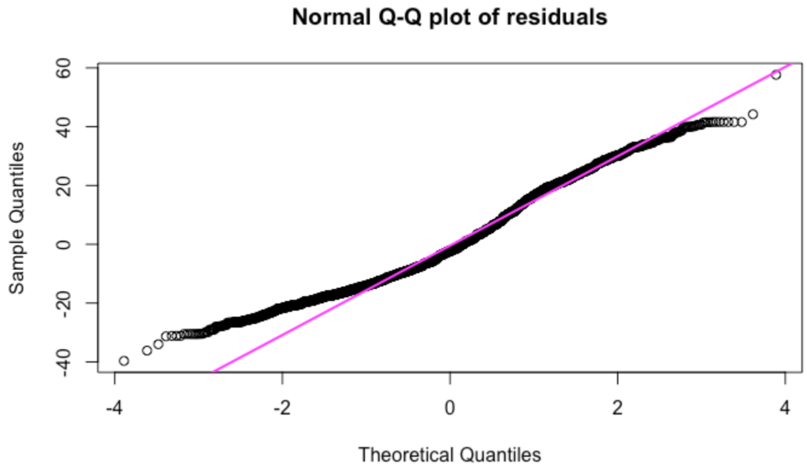


Figure 3

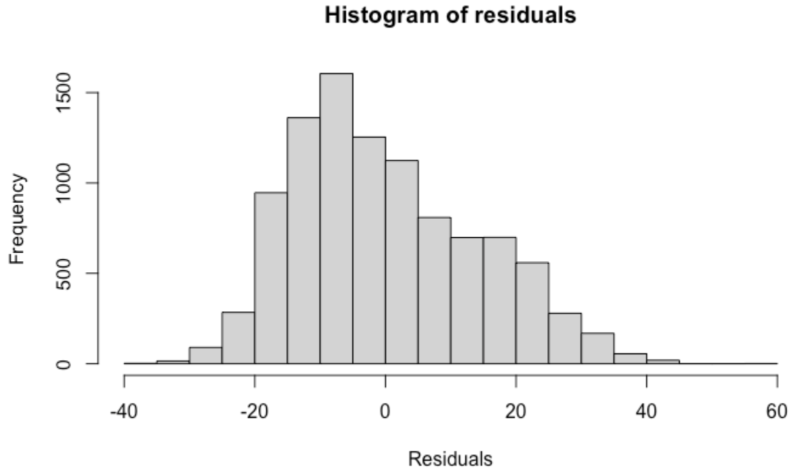


Figure 4

Normality of Residuals

Figure 3 displays a Q-Q plot showing residuals that are close to the guideline in the center of the plot and deviating in both tails. Additionally, **Figure 4** shows a residual histogram that is roughly bell-shaped with a slight right skewed. Although there are some deviations in the shape of **Figure 4** and the extremes in **Figure 3**, due to our large sample size, we can state that the residuals are approximately normal, with some slight deviations that we acknowledge as limitations.

Linearity

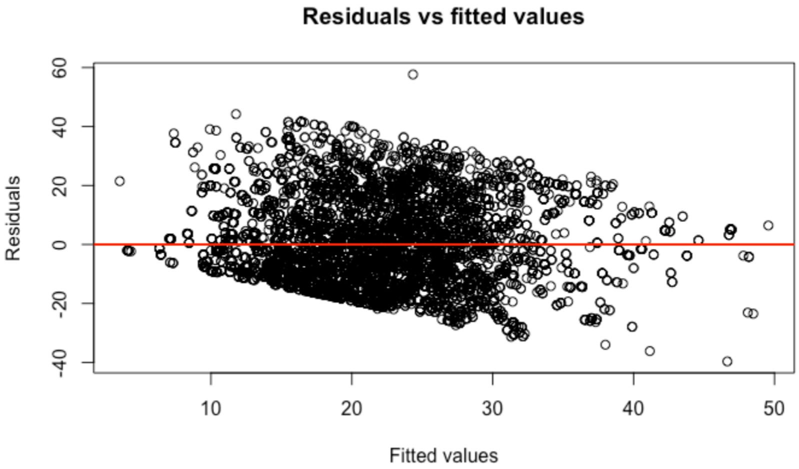


Figure 5

Figure 5 shows a Residuals vs Fitted plot with no strong curved pattern. There is a slight tilt and some funneling on the lower left side. However, since this is only impacting a small portion of the graph, it does not appear to be particularly extreme. We acknowledge that the residuals are not perfectly random and some mild structure remains, nevertheless, since there is no real curved pattern, we do not think that there is any major violation of linearity.

Mean 0

As shown on **Figure 5**, the residuals are roughly centered around 0, so the mean-zero assumption is reasonable.

Homoscedasticity

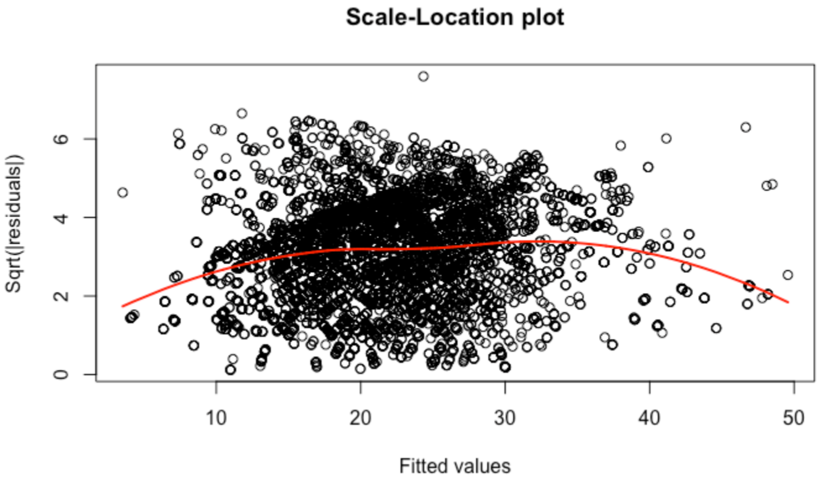


Figure 6

Figure 6 shows a scale location with a slight curved pattern. This means that there is more variance in the middle fitted range, and it indicates mild heteroskedasticity. Thus, it likely only has a small impact on the data, nevertheless, the assumption is not met.

High Leverage and Influence Points

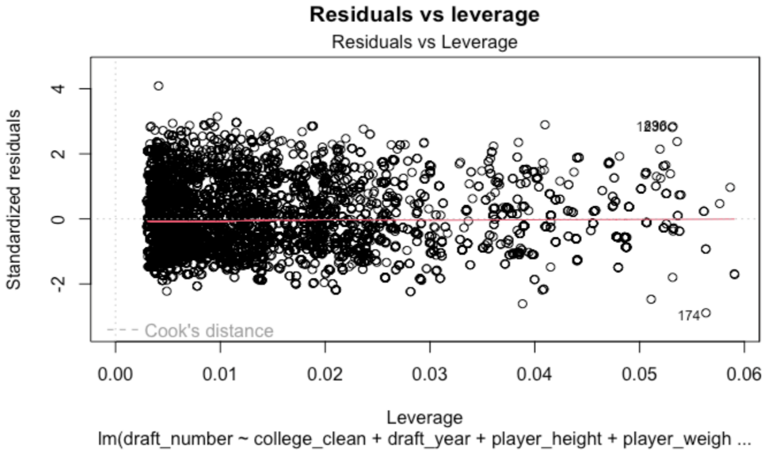


Figure 7

Although this is not one of the assumptions, as a general diagnostic check, **Figure 7** is used to verify that there are no extreme high-leverage points with large residuals. This means that there are no highly influential observations affecting our results.

Results Description

College Path vs. Draft Position

Model 4 serves as the primary basis for interpreting substantive associations between predictors and draft position. The outcome is draft position and the reference category for college is “None” (players without a listed U.S. College). In this model, each college’s coefficient represents the average difference in draft position between players from that college and players in the “None” group, holding draft year, height, weight, and country constant. A negative coefficient indicates that players from that college are drafted earlier (lower pick numbers) than comparable players with no U.S. college path.

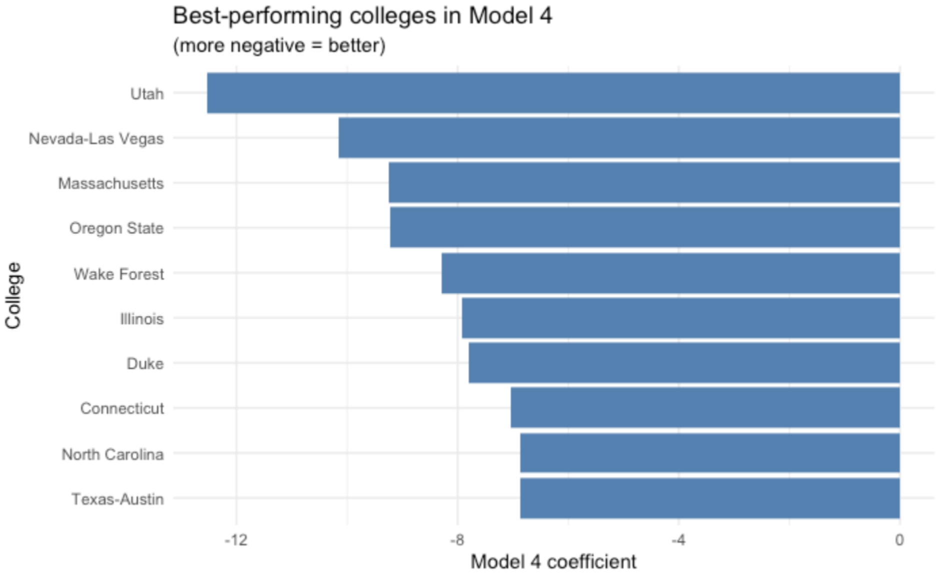


Figure 8: Top 10 Colleges by Estimated Effect on Draft Position by Model 4

Figure 8 shows the top ten colleges with the most negative Model 4 coefficients. University of Utah appears to be our top college with a coefficient of approximately -12.5, meaning players from this university are predicted to be drafted 12-13 picks earlier, on average, than similar players in the “None” group. University of Nevada, Las Vegas (UNLV) closely follows with coefficients around -10, with the University of Massachusetts (UMass) and Oregon State after, with coefficients around -9. The alumni from these schools are selected roughly one-third of a round (~30 picks per round) ahead of comparable non-college prospects.

Several other prominent programs also rank highly, including Duke, University of North Carolina (UNC), and University of Connecticut (UConn) show coefficients near -7, indicating that players from these schools enjoy an average advantage of about seven draft spots relative to non-college prospects after adjustment.

Our model revealed that many other colleges displayed coefficients close to zero, implying no meaningful difference in average draft position from the non-college baseline once covariates were controlled. The pattern revealed that only a select group of programs - typically those with stronger historical recruiting backgrounds - are associated with earlier selections in the draft rounds.

Draft Year and Draft Position

While our variables have a modest fit, they cannot fully predict a player's draft number on their own without additional player performance statistics. However, there are still a few significant effects that our models demonstrate based on our predictive variables alone.

We found that adding draft year in Model 2 increased the adjusted R^2 (see Appendix: Table A1) from about 0.117 to 0.146, indicating that year/time effects matter to some extent for draft position. In our selected Model 4, draft year continues to be a significant predictor as some years are associated with slightly earlier or average picks relative to the reference year, even after adjusting for college, size and country. This supports our initial thought of treating draft year as a control variable to account for changes in the NBA context and scouting practices over time.

When college interacts with exact years in Model 5, the training R^2 value jumped dramatically to 0.620 (see Appendix: Table A1 or Table A2), but the cross-validation showed no real improvement in predictive performance, and many of the interaction coefficients were poorly estimated (large negative and positive numbers) or undefined. Although the timeline and draft years do have an effect on the college associations, the precise year-by-college interactions created too many unique combinations for a dataset of this size.

Physical Characteristics

As mentioned, Model 3 adds physical characteristics like player height and weight. In Model 3 we find that taller players tend to be drafted earlier with a small but statistically significant coefficient (-0.14). This means that players are drafted around -0.14 picks earlier per additional centimeter. This aligns with our initial descriptive height-draft scatter plot in **Appendix: Figure B1**, where the regression line slopes downward but the point cloud still remains wide at each height.

However, once height is included, weight is no longer statistically significant. This suggests that conditional on height, additional weight does not meaningfully shift average draft position in this model. This makes sense logically as weight is typically correlated with height.

Origin (Country)

The country indicator in Model 4 implies that after controlling for college, draft year, height and weight, USA-born players are drafted about 2.5 picks earlier on average than comparable Non-USA players. Our initial descriptive boxplot (see **Appendix: Figure B2**) showed that distributions for USA and Non-USA players overlap greatly, though the median for USA players is lower than that of Non-USA players. The difference is modest in magnitude but statistically detectable in a large sample, as shown by our Model results.

Overall Model Fit for our Chosen Model (Model 4)

The adjusted R^2 for Model 4 is approximately 0.155, meaning that the combination of college path, draft year, height, weight and country explains about 15-16% of the variation in draft position (see **Appendix: Table A1 or A2**). The cross-validated RMSE of roughly 14.26 draft picks (see **Appendix: Table A2**) indicates that even with these predictors, the typical prediction error still remains large.

The Era Model

We still wanted to observe the effects of college x year interactions, but found that we could not do so due to the complexity of Model 5. In order to observe these effects, we created an era model.

Era Model: $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{college}_i + \boldsymbol{\beta}_2 \text{year}_i + \boldsymbol{\beta}_3(\text{college}_i \text{ x era}_i) + \boldsymbol{\beta}_4 \text{height}_i + \boldsymbol{\beta}_5 \text{weight}_i + \boldsymbol{\beta}_6 \text{country}_i + \boldsymbol{\varepsilon}_i$

The era model simplifies Model 5 by grouping draft years into three separate periods, the 1990s, 2000s, and 2010s, but still including the college x era interactions in a non-parallel model. This allows the association between a college and draft position to differ across broader eras without estimating a separate interaction for every single draft year, which turned out to be a significant problem in our original Model 5.

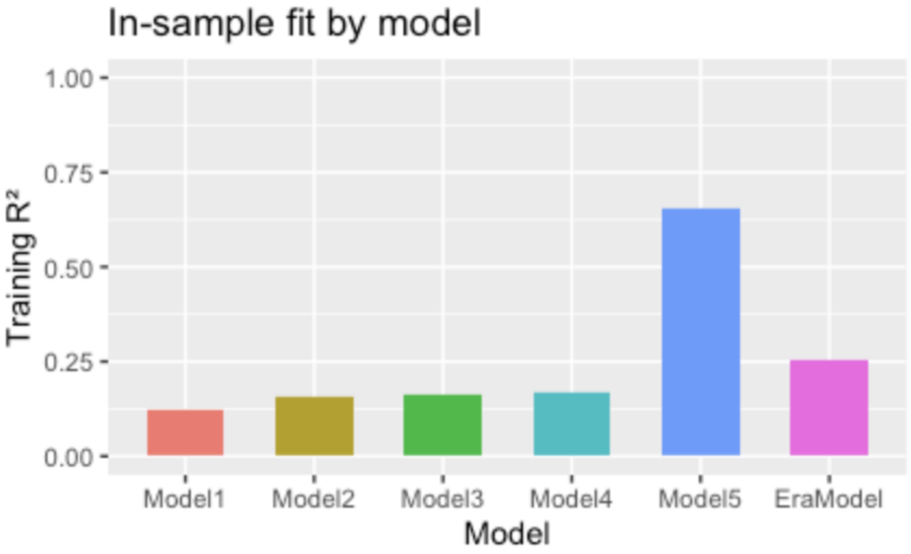


Figure 9: Training R^2 for each Model Including Era Model

This model uses substantially more parameters than Model 4, but much fewer than Model 5 (roughly 282 total parameters instead of 3165, see **Appendix Table A2**). This model also achieves a slightly higher R^2 than Model 4 but still lower than that of Model 5 as shown by **Figure 9**. This indicates that allowing college associations to change over eras does slightly improve predictive performance without the extreme overfitting seen in Model 5.

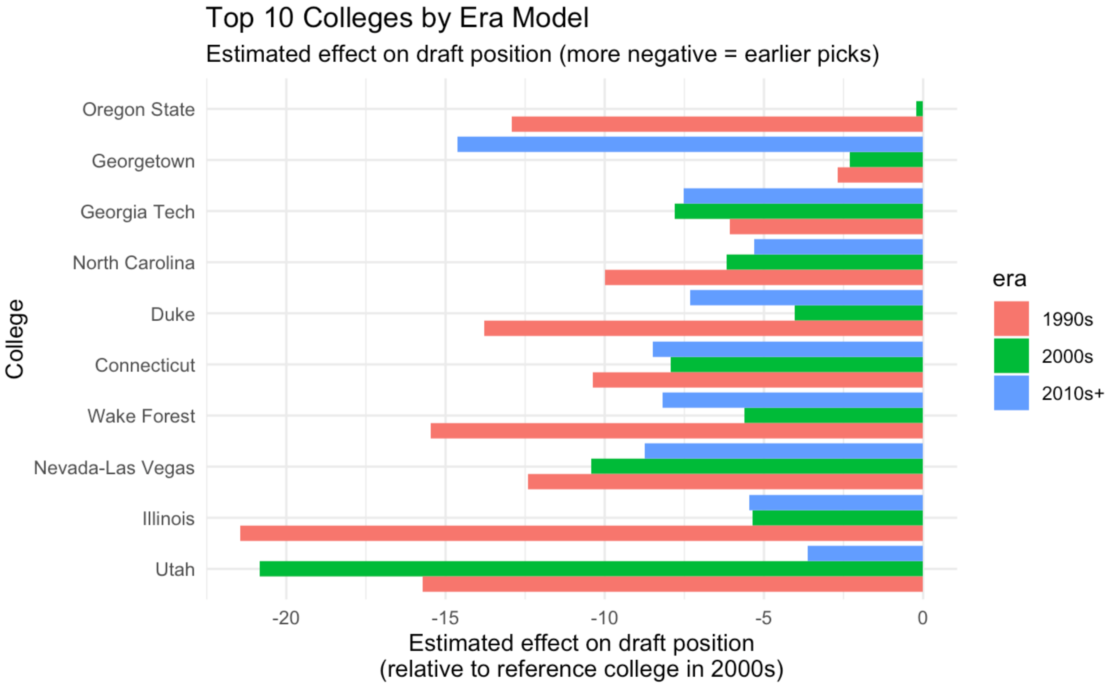


Figure 10: College Effects on Draft Position by Era

This suggests that the “strength” of a college program is not constant over time. Some programs appear to have particular eras of success where they consistently produce earlier-picked players. **Figure 10** shows the college effects on draft position by era, showing that even the strongest programs do not maintain a constant draft advantage over time. Utah was ranked first from Model 4, and here we can see that they had a large advantage in the 1990s and 2000s but the effects shrink substantially in the 2010s. Other programs, such as Georgia Tech and North Carolina have more stable effects across all eras.

Conclusions

Concerns and Limitations

A key limitation of our analysis is that both the models and the dataset capture only a narrow slice of what drives NBA draft decisions. Model 4 explains about 15-16% of the variation in draft position but still has a cross-validated RMSE of roughly 16 picks, so even our best linear specifications leave most of the outcome unexplained. Many of our assumptions are also only partially satisfied, so our results must be interpreted with caution. Furthermore, our models remain a simplified linear explanation of this dataset. With further research, it would be ideal to focus on more nonlinear or interaction heavy patterns that our simple models were unable to capture. When trying to fit a not-necessarily parallel lines model to our data (model 5), we were met with uninterpretable coefficients since some college-year cells had few players. We adjusted for this issue by observing college-era rather than college-year in Era Model, however, that was still only a modest increase in predictive power from Model 4.

On the data side, we could not observe detailed amateur performance metrics, which would likely account for a large percent of variation in draft outcomes. As a result, our coefficient estimates should be interpreted as describing broader associations conditions on a limited set of background variables, rather than a full causal explanation of why players are drafted in specific rounds.

Key Takeaways

In this project, we examined how college pathway and draft year are associated with NBA draft position after accounting for physical characteristics and country of origin. We found that college is indeed related to draft outcomes, with players from certain schools being selected earlier on average than comparable players with no U.S. college background or from different schools. However, college alone explains limited variation in draft position. Adding draft year, height, weight, and country modestly improved model fit, indicating that these background characteristics provide some explanatory power but leave much of the draft process unexplained.

We selected Model 4 as our primary model because it balances interpretability and predictive performance. While Model 5's fully interacted college–draft year model achieved a much higher in-sample R^2, cross-validation suggested that this improvement largely reflected overfitting rather than meaningful gains in predictive accuracy. Future work could improve this tradeoff by grouping draft years into broader eras (e.g., pre-2000, 2000–2010, post-2010), allowing for structural changes in the draft while avoiding the instability of year-by-college interactions. Additionally, given the mild violations of independence, homoscedasticity, and normality, resampling approaches such as bootstrapping could be used to construct more reliable confidence intervals without relying heavily on strict parametric assumptions. Together, these extensions would allow for more robust inference while maintaining a parsimonious and interpretable modeling framework.

Appendix A

Model Comparison

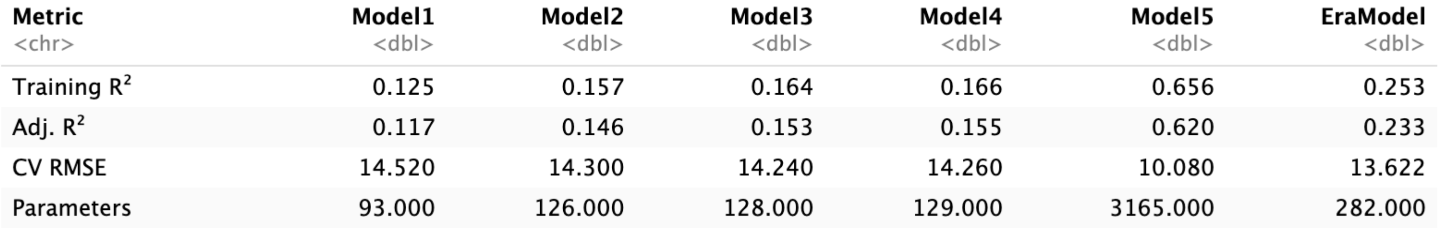
Table A1

Comparing Models and Model Compositions

Model	**Predictors Included**	**Adjusted R2**	**Residual SE**	**F-Stat**	**P-value**
1	College only	0.117	14.45	15.33	< 2.2 e-16
2	College + draft year	0.146	14.21	14.64	< 2.2 e-16
3	College + draft year + height + weight	0.153	14.15	15.18	< 2.2e-16
4	College + draft year + height + weight + country	0.155	14.13	15.31	< 2.2e-16

5	College x draft year + height + weight + country	0.620	9.48	17.88	< 2.2e-16
Era	College x era + height + weight + country	0.234	13.47	12.67	<2.2e-16

Table A2
Comparing Models with Cross-Validations



Metric	Model1	Model2	Model3	Model4	Model5	EraModel
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Training R^2	0.125	0.157	0.164	0.166	0.656	0.253
Adj. R^2	0.117	0.146	0.153	0.155	0.620	0.233
CV RMSE	14.520	14.300	14.240	14.260	10.080	13.622
Parameters	93.000	126.000	128.000	129.000	3165.000	282.000

Appendix B

Descriptive Visualizations of Key Variables

Figure 1B

Draft Position by Country of Origin

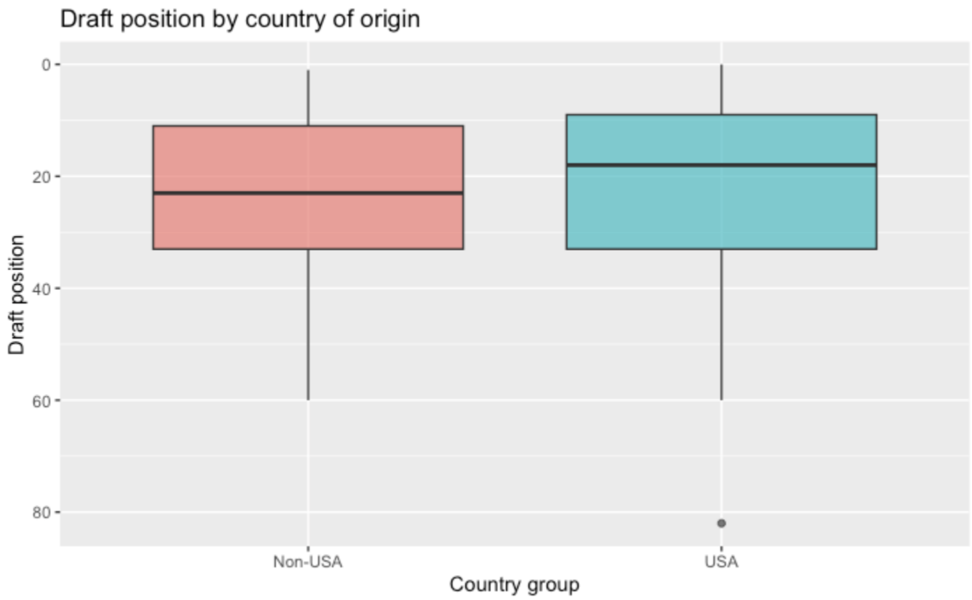
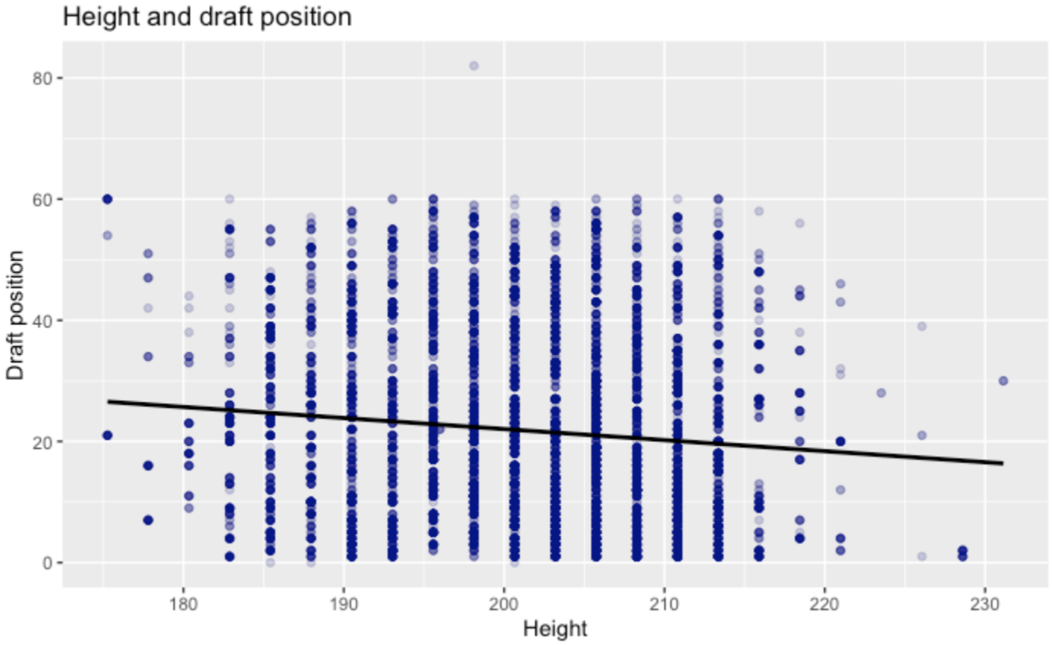


Figure 2B

Draft Position by Height

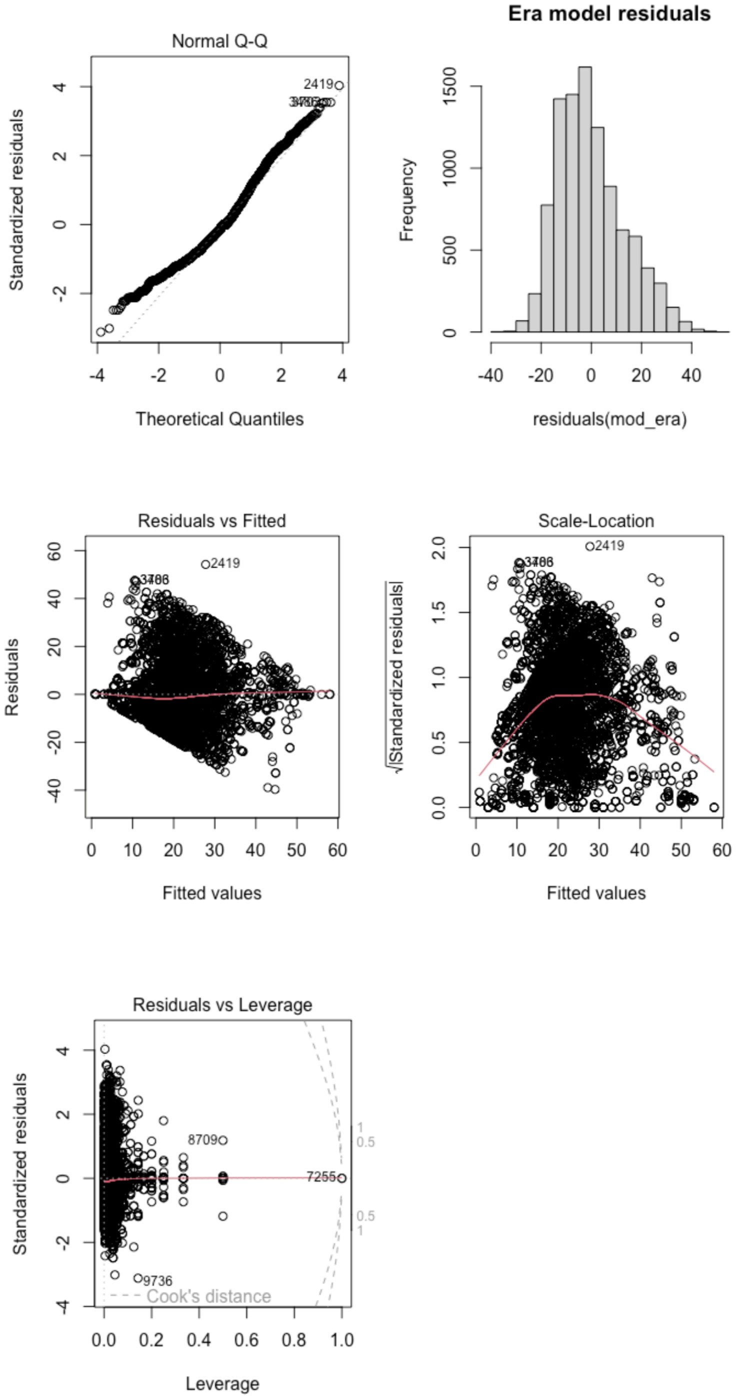


Appendix C

Assumptions of the Era Model

Figure 1C

Plots to Verify Assumptions for the Era Model: Q-Q Plot (Figure 1Ca), Era Model Residuals (Figure 1Cb), Residuals vs. Fitted (Figure 1Cc), Scale-Location Plot (Figure 1Cd), Residuals vs Leverage (Figure 1Ce)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
2419
-4
-2
0
2
4
Era model residuals
Frequency
residuals(mod_era)
0
500
1000
1500
-40
-20
0
20
40
Residuals vs Fitted
Residuals
Fitted values
2419
60
40
20
0
-20
-40
10
30
50
Scale-Location
√|Standardized residuals|
Fitted values
2419
0.0
0.5
1.0
1.5
2.0
Residuals vs Leverage
Standardized residuals
Leverage
8709
7255
9736
Cook's distance
1
0.5
0.5
1
0.0
0.2
0.4
0.6
0.8
1.0