Part 1: NBA Player Position Classification

1 Introduction

Traditionally, there are five basketball positions: point guard (PG), shooting guard (SG), small forward (SF), power forward (PF), and center (C). When created, these labels differentiated players roles. Centers, for example, grabbed rebounds and blocked shots. Over time, rule changes and the rise of analytics have allowed the NBA to develop into a high speed game with emphasis on the 3-point shot. The lines dividing positions have become increasingly blurred. These changes have sparked debate about the validity of the five-position system. Some argue that the traditional labels are archaic and should be redefined; others argue that they should be abolished [1][2][3]. In this report, we use classification techniques to weigh in on this debate. Specifically, we investigate how well the five-position system captures playing styles, both historically and today.

2 Data

Our data is comprised of per 36 minute NBA statistics for players in a given season from *Basketball-Reference* and span the following seasons: 1993-94 to 1997-98 and 2013-14 to 2018-19 [4]. There is a single response, Position (Pos). There are 28 predictors, most of which are traditional statistics measuring blocks, rebounds, assists, steals, fouls, and shooting on a per 36 minute basis. The remainder are per season statistics or descriptive variables like player name, age, rank, team, games played, and minutes played (Appendix Table B.1).

In this analysis, we build models using two datasets representing different basketball eras (Era 1 and Era 2). Because we analyze player statistics across multiple seasons, we refer to each observation as a player-season. We assume that all player-seasons are independent. Era 1 contains 2,119 player-seasons from 1993-94 to 1997-98 and Era 2 contains 2,476 player-seasons from 2013-14 to 2017-18. We also create a third dataset (2018) containing 530 player-seasons to be used as an additional testing set for our final Era 1 and Era 2 models.

3 Preprocessing

We check the response, Pos, and notice that it is a factor with 14 levels. 9 correspond to combinations of positions (i.e. PF-SF), which refer to players who played two positions in a given season. We exclude these player-seasons, ensuring that Pos has 5 levels (with the baseline being C). Next, we plot the distribution of Minutes Played (MP) across player-seasons for Era 1 and Era 2. A significant number of players had less than 400 MP in a given season (Appendix Figure B.1). We exclude these player-seasons, as players with insignificant on-court time are unlikely to produce robust statistics that we can use to determine position. After this and the previous step, we have 1,581 observations in Era 1, 1,828 in Era 2, and 374 in 2018.

We remove Age, Team, and Rank because they are unrelated to Pos. We also remove efficiency statistics (i.e. shots made) as we do not want to find the best players, but rather all players in a certain position. Next, we assess multicollinearity (Appendix Figure B.2). We deem a correlation too high if it exceeds 0.7 and remove the following variables: Offensive Rebounds, Defensive Rebounds, Games Played, Games Started, Total Points, and Field Goal Attempts.

We ensure that our datasets do not contain any missing values and that each dataset is balanced (Appendix Table B.2). Next, we split Era 1 and Era 2 into training and testing sets according to a 70%-30% split. The split is randomised to ensure that all seasons within a particular era are represented in the training sample. Last, we standardise training and testing sets because our predictors are on different scales. Minutes Played is measured on a per season

basis, while the remainder are measured on a per 36 minute basis. We use this standardised data to train and test all models for accurate model comparison.

4 Analysis

We start with linear methods, which are typically simple and interpretable. The first is naive Bayes, a simple approach used in classification settings. Next, we try multinomial logistic regression, an extension of logistic regression allowing for multiple classes in the response. We assess the need for variable selection by incorporating a lasso penalty. We then try a variety of discriminant methods, such as linear discriminant analysis, quadratic discriminant analysis, mixture discriminant analysis, and regularised discriminant analysis.

We also explore more flexible approaches to see if relaxing the linearity assumption improves performance. While typically less interpretable, these methods may have better accuracy. We try k-nearest neighbours, a widely applied nonlinear approach. Next, we explore tree-based methods, which are conveniently built to handle multi-class qualitative predictors. We begin with classification trees, then move to bagging, double bagging, random forest, and boosting. Last, we explore support vector machine (SVM) algorithms, although they do not extend naturally to the multi-class setting. First, we fit a support vector classifier using the one-versus-one approach. We then try implementing SVMs with a radial kernel using the one-versus-one and one-versus-all approaches. See Appendix A for a more detailed discussion of our method implementations.

5 Results

We use the following metrics to assess our models: misclassification error rate (MER), accuracy, precision, recall, F1 and kappa. Note that precision, recall, and F1 are per-class metrics. In order to determine the overall scores for our methods, we compute their macro-averages. Detailed results can be found in Appendix Tables B.3 and B.4.

5.1 Comparison of Era 1 and Era 2 Results

The most powerful Era 1 method is the SVM with a radial kernel using the one-versus-one approach. This model achieves an accuracy of 73.5% and has the highest scores for all of our performance measures. The kappa is 0.668. According to Landis and Koch (1977), this signifies substantial agreement between our predicted classes and actual observations. Era 1 also performs very well on other nonlinear approaches, achieving accuracies of 72.8%, 72.4%, and 72.2% for the SVC, RF, and KNN algorithms respectively. The most powerful linear approach is the MLR, which achieves an accuracy of 71.6%.

Table 1: Top Performing Methods for Era 1 Data

Method	MER	Accuracy	Precision	Recall	F1	Kappa
1v1SVM (Radial Kernel)	0.265	0.735	0.735	0.729	0.727	0.668
SVC	0.272	0.728	0.727	0.722	0.720	0.660
RF	0.276	0.724	0.722	0.721	0.718	0.655
KNN	0.293	0.707	0.705	0.703	0.701	0.634
MLR	0.284	0.716	0.711	0.709	0.709	0.644

The RF significantly outperforms all other methods for our Era 2 dataset and results in an accuracy of 72.1%. Additionally, this method has the highest precision, recall, F1, and kappa. The kappa of 0.651 indicates substantial agreement between predicted and observed classes. The next strongest methods are bagging and SVM with a radial kernel using the one-versus-one

approach, which achieve accuracies of 69.8% and 68.7%, respectively. Like Era 1, the strongest linear method is the MLR, which results in an accuracy of 67.9%.

Table 2: Top Performing Methods for Era 2 Data

Method	MER	Accuracy	Precision	Recall	F1	Kappa
RF	0.279	0.721	0.715	0.717	0.715	0.651
Bagging	0.302	0.698	0.692	0.694	0.693	0.622
1v1SVM (Radial Kernel)	0.313	0.687	0.678	0.681	0.678	0.608
SVC	0.317	0.683	0.675	0.679	0.675	0.603
MLR	0.321	0.679	0.671	0.675	0.671	0.599

Comparing the Era 1 and Era 2 results, one can see that Era 1 achieves slightly better performance across the board. However, both datasets respond well to RF and SVM approaches. MLR, a more interpretable approach, performs relatively well for both datasets.

5.2 Final Model Selections

This analysis involves competing strategies for assessing changes in playing styles over time: (1) to explore the characteristics that define each position and understand the ways they have evolved; and (2) to compare the performance of Era 1 and Era 2 models in classifying the positions of modern NBA players. For (1), we want a model that is accurate and interpretable. We therefore choose MLR because it is the best performing linear method for both Era 1 and Era 2. This approach is appealing because it allows us to explore the impact of individual predictors on each position. For (2), we want the models with the highest accuracy, as they will have the best chance of performing well on our 2018 data. We therefore select the SVM with a radial kernel using the one-versus-one approach for Era 1 and the RF for Era 2.

6 Discussion

6.1 Changes in Playing style

We calculate variable importance for our Era 1 and Era 2 MLR models and find that Total Rebounds, Blocks, Steals, and Assists contribute most to both models (Appendix Table B.5). Next, we look at the coefficients and transform them into relative risk ratios (Appendix Tables B.6 and B.7). Rebounding is clearly a hallmark trait for centers, as are assists for point guards. While there are not any substantial differences between the models, we notice some subtle shifts. For example, the odds ratios associated with 3-point shot attempts drops from 2 in Era 1 to about 1-1.5 in Era 2. This could demonstrate the increasing trend for players in all positions to shoot 3-pointers.

We investigate further by looking at the confusion matrices for both models and notice the same pattern (Appendix Table B.8). A significant number of centers are misclassified as power forwards, and shooting guards misclassified as small forwards (and vice versa). So while some positions do have clear distinctions (i.e. C and PG), a significant amount of overlap exists. This provides some evidence that perhaps we cannot assess how playing styles have changed because the five-position system does not accurately capture differences in playing styles.

6.2 Performance on 2018 - 19 Data

Next, we test our Era 1 SVM with radial kernel (one-versus-one approach) (Era 1 Model) and our Era 2 random forest model (Era 2 Model) on the 2018 dataset. The misclassification error rates are 0.476 and 0.286 for the Era 1 and Era 2 models, respectively. It is not surprising that the Era 2 model performs well on the 2018 data as the two datasets are very close in time. In fact, the Era 2 model performs almost as well on 2018-19 players as it did on the Era 2 testing

set. The Era 1 model, in comparison, performs poorly. This provides some evidence that NBA playing styles today are in fact different than they were 20 years ago.

7 Conclusion

In this analysis we have applied machine learning algorithms to datasets representing two distinct eras of basketball. Our Era 1 models generally outperformed the Era 2 ones, suggesting that perhaps the positions labels better described playing styles 20 years ago. However, we are unable to obtain any accuracy above 73.5%. Further, by interpreting the results of our Era 1 and Era 2 MLR models, we cannot draw conclusions about how the characteristics defining each player position have changed over time. Testing our most accurate models from each era on our 2018 data, we see that there has been a significant change in playing styles over time. Overall, we think there is plenty of reason to question the ability of these position labels to capture the different playing styles in the NBA today. Further investigation is certainly warranted.

References

[1] Alagappan, M. (2012) From 5 to 13: Redefining the Positions in Basketball. *MIT Sloan Sports Analytics Conference.* Available from: http://www.sloansportsconference.com/content/the-13-nba-positions-using-topology-to-identify-the-different-types-of-players/.

[2] McMahan, I. (2018) How (and why) position-less lineups have taken over the NBA playoffs. *The Guardian.* Available from: https://www.theguardian.com/sport/blog/2018/may/01/how-and-why-position-less-lineups-have-taken-over-the-nba-playoffs.

[3] Dakhil, M. (2017) Position-Less Basketball Taking Shape in the NBA. *The Jump Ball.* Available from: https://thejumpball.net/2017/08/18/position-less-basketball-taking-shape-in-the-nba/comment-page-1/.

[4] Basketball-Reference. *NBA Player Stats: Per 36 Minutes.* Web. Available from: https://www.basketball-reference.com/ [Accessed 18th November 2019].

[5] Landis, J. & Koch, G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics.* 33(1), 159-174. Available from: https://www.jstor.org/stable/2529310.

Appendices

A Further Details on Method Implementations

Multinomial Logistic Regression (MLR)

We start by training two MLR models on our Era 1 and Era 2 training data. We use the likelihood test to whether we should remove the intercept. We can see that the p-value is significant, therefore, we can reject the null hypothesis that the coefficient of the intercept is zero. Hence, we should not remove the intercept.

Listing A.1: Summary of Likelihood Ratio Tests for Era 1 and Era2

```
Likelihood ratio test

Model 1: Pos ~ MP + X3PA + X2PA + FTA + TRB + AST + STL + BLK + TOV +
    PF
Model 2: Pos ~ (MP + X3PA + X2PA + FTA + TRB + AST + STL + BLK + TOV +
    PF) - 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  44 -749.76
2  40 -914.75 -4 329.99  < 2.2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

Looking at the coefficient estimates for our Era 1 model (Table A.1), we find that Two-Point Shot Attempts (X2PA), Free Throw Attempts (FTA), and Turnovers (TOV) may not contribute significantly. For our Era 2 model (Table A.2), we find that the same is true for FTA and TOV.

Table A.1: Significance of Coefficients for MLR of Era 1

	Dependent variable:			
	PF	PG	SF	SG
	(1)	(2)	(3)	(4)
MP	−0.130	−0.837***	−0.359	−0.574**
	(0.174)	(0.324)	(0.224)	(0.262)
X3PA	0.976***	0.742	1.063***	0.799*
	(0.370)	(0.467)	(0.392)	(0.418)
X2PA	0.007	0.252	0.280	0.373
	(0.161)	(0.374)	(0.224)	(0.288)
FTA	0.238	−0.280	0.236	0.392
	(0.159)	(0.370)	(0.223)	(0.280)
TRB	−0.121	−7.907***	−2.329***	−5.933***
	(0.185)	(0.785)	(0.306)	(0.512)
AST	0.091	3.511***	0.277	1.095**
	(0.363)	(0.606)	(0.466)	(0.533)
STL	0.860***	2.663***	1.596***	2.088***
	(0.190)	(0.325)	(0.240)	(0.279)
BLK	−1.198***	−4.513***	−1.756***	−2.648***
	(0.148)	(0.856)	(0.251)	(0.478)
TOV	−0.266	−0.026	−0.263	−0.181
	(0.175)	(0.380)	(0.246)	(0.302)
PF	−0.377**	−1.564***	−1.024***	−1.322***
	(0.168)	(0.412)	(0.241)	(0.330)
Constant	2.508***	−4.273***	3.341***	1.168**
	(0.376)	(0.848)	(0.396)	(0.476)
Akaike Inf. Crit.	1,587.513	1,587.513	1,587.513	1,587.513

Note: *p<0.1; **p<0.05; ***p<0.01

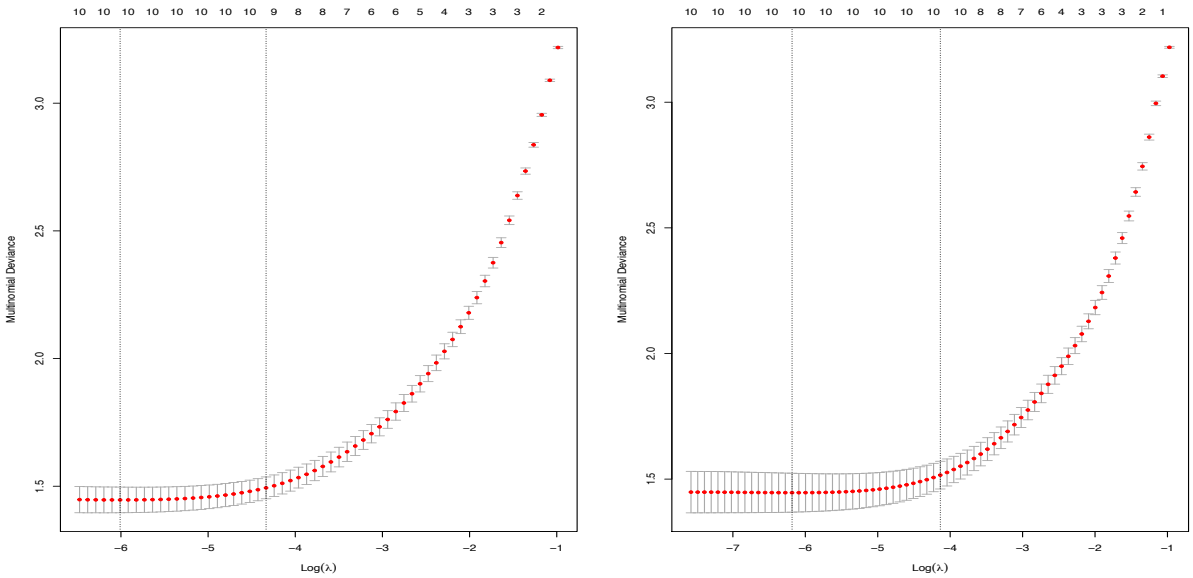
Table A.2: Significance of Coefficients for MLR of Era 2

	Dependent variable:			
	PF	PG	SF	SG
	(1)	(2)	(3)	(4)
MP	0.162	−0.511*	0.114	−0.257
	(0.154)	(0.275)	(0.207)	(0.230)
X3PA	0.476***	0.170	0.153	0.578**
	(0.173)	(0.338)	(0.246)	(0.277)
X2PA	−0.049	−0.073	−0.667**	0.091
	(0.155)	(0.381)	(0.269)	(0.315)
FTA	−0.146	0.121	0.395	0.069
	(0.160)	(0.345)	(0.264)	(0.302)
TRB	−0.832***	−8.625***	−3.789***	−6.700***
	(0.198)	(0.668)	(0.377)	(0.489)
AST	−0.430	3.733***	−0.397	0.629
	(0.278)	(0.529)	(0.394)	(0.443)
STL	0.389***	1.835***	1.426***	1.719***
	(0.139)	(0.281)	(0.200)	(0.229)
BLK	−1.122***	−3.598***	−1.867***	−2.413***
	(0.146)	(0.619)	(0.264)	(0.366)
TOV	−0.157	−0.531	−0.044	0.012
	(0.204)	(0.425)	(0.319)	(0.365)
PF	−0.266*	−0.140	−0.596***	−0.570**
	(0.152)	(0.354)	(0.230)	(0.274)
Constant	2.095***	−3.115***	2.306***	0.857**
	(0.255)	(0.600)	(0.277)	(0.335)
Akaike Inf. Crit.	1,837.345	1,837.345	1,837.345	1,837.345

Note: *p<0.1; **p<0.05; ***p<0.01

Next, we refit our multinomial logistic regression models, this time incorporating a lasso penalty to see if the coefficients of these less important variables will shrink towards zero. We use cross-validation to obtain the optimal value of lambda (lambda.min). We also apply the 1-standard error rule to identify the values of lambda within one standard error of lambda.min which gives us the simplest possible models (lambda.1se). Our values of lambda.min are 0.002 and 0.013 for Era 1 and Era 2, respectively. Our values of lambda.1se are 0.002 and 0.016 for Era 1 and Era 2, respectively.

Figure A.1: Plot of Cross-Validation Lambdas for MLR with Lasso for Era 1 (left) and Era 2 (right)

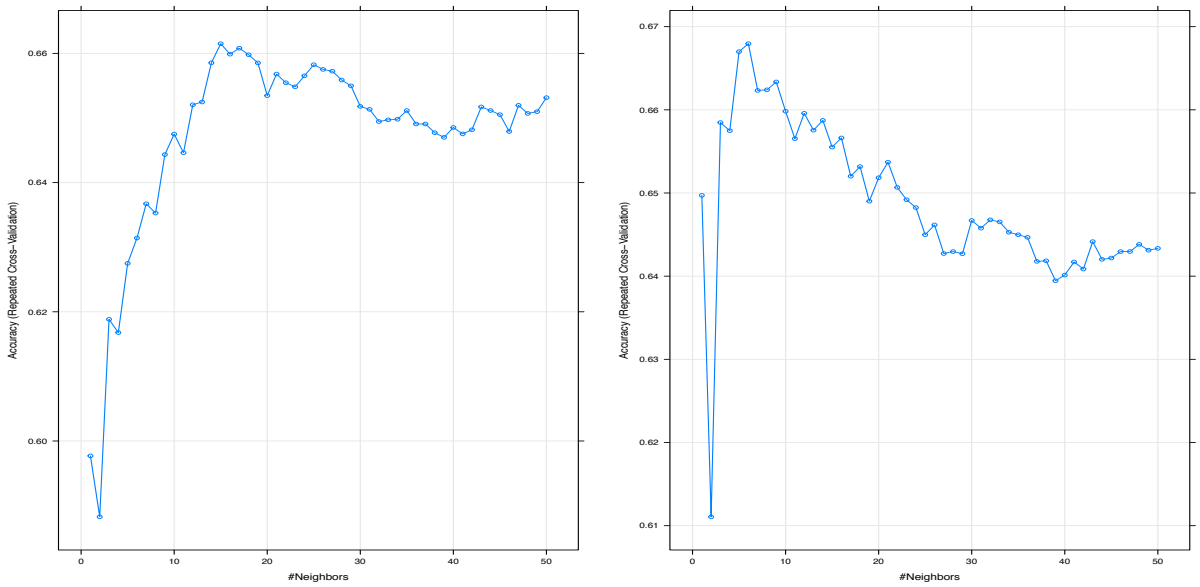


We then refit our models using lambda.min and see that all of our predictors are still included. We also try refitting our model using the lambda.1se and see that Minutes Played (MP) is no longer considered for our Era 1 model, but that all variables are still included in our Era 2 model. Looking at the misclassification error rates for our lambda.min and lambda.1se models, we notice that neither of these models improves substantially on the performance of the original MLR models for Era 1 and Era 2. This is reflected in the fact that our lambda estimates are very small. As a result, we proceed using only the original MLR models.

K-Nearest Neighbors

We apply 10-fold cross validation, repeated 10 times to select the optimal number for K within the range 1 to 50. Cross-validation yields optimal values of K = 15 and K = 6 for Era 1 and Era 2, respectively (k.min). We then consider selecting the optimal K by applying the 1-standard error rule. This approach yields K = 27 for Era 1 and K = 6 for Era 2 (k.1se).

Figure A.2: Cross-Validation Results for Optimal K of Era 1 (left) and Era 2 (right)



We then proceed to train KNN models on our Era 1 and Era 2 datasets using both k.min and k.1se. In the case of Era 2, K = 6 using both cross-validation approaches and so we do not need to make a choice. For Era 1, on the other hand, using k.1se results in a 2% increase in the misclassification error rate compared to k.min. As a result, we proceed using k.min. Of course, by choosing a larger value for K, we run the risk of oversimplifying the distribution and being more computationally expensive. With over 1,000 training observations, though, we do not think either of these issues are causes for concern.

Classification Tree

We fit classification trees on our Era 1 and Era 2 datasets. Looking at the results (Listing:A.2 and A.3), we see that the Era 1 tree contains 8 terminal nodes and the Era 2 tree contains 11 terminal nodes. We also see that Total Rebounds (TRB), Assists (AST), Blocks (BLK), and Three-Point Shot Attempts (X3PA) are the only variables considered for both the Era 1 and Era 2 trees.

Listing A.2: Summary of Initial Classification Trees of Era 1

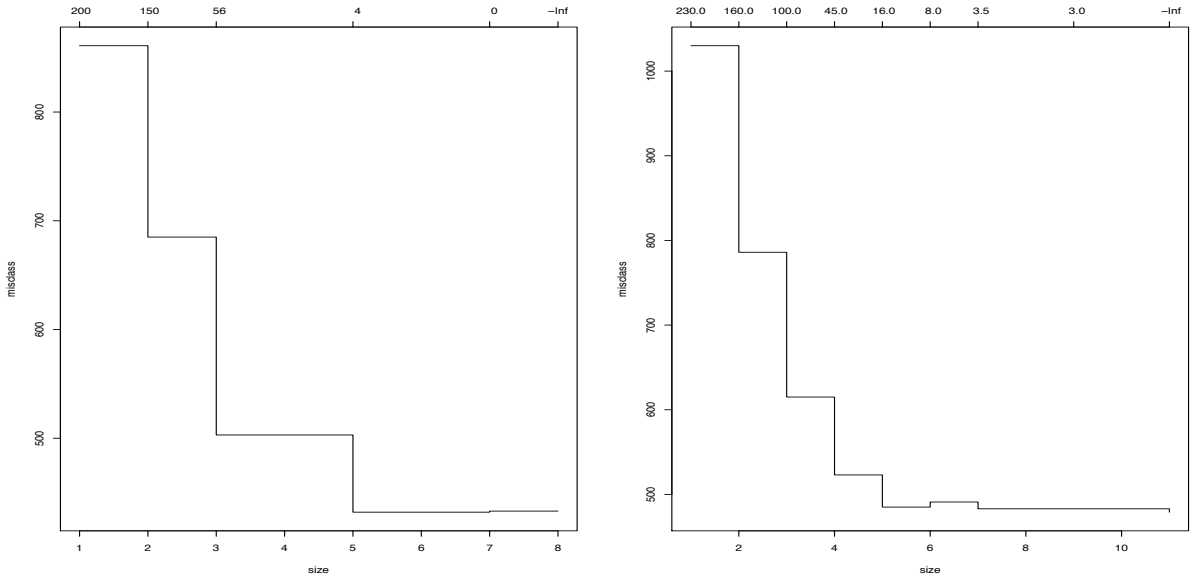
```
Classification tree:
tree(formula = Pos ~ ., data = era1_train)
Variables actually used in tree construction:
[1] "TRB"  "AST"  "BLK"  "X3PA"
Number of terminal nodes:  8
Residual mean deviance:  1.662 = 1825 / 1098
Misclassification error rate: 0.3599 = 398 / 1106
```

Listing A.3: Summary of Initial Classification Trees of Era 2

```
Classification tree:
tree(formula = Pos ~ ., data = era2_train)
Variables actually used in tree construction:
[1] "TRB"  "AST"  "X3PA" "BLK"
Number of terminal nodes:  11
Residual mean deviance:  1.533 = 1944 / 1268
Misclassification error rate: 0.3339 = 427 / 1279
```

Next, we prune the trees to see if misclassification error rate improves. We perform cross-validation to determine that optimal tree complexity is reached at 5 terminal nodes for Era 1 and at 11 terminal nodes for Era 2.(Figure:A.3)

Figure A.3: Cross-Validation Results for Optimal Tree Complexity of Era 1 (left) and Era 2 (right)



We then refit our classification trees based on this new information. For Era 2, the initial tree and the pruned tree have the same number of terminal nodes. However, the pruned tree has a slightly higher misclassification error rate. In the case of Era 1, the pruned tree has a misclassification error rate that is about 3.5% higher than that of the original tree. Because our initial trees have higher accuracies and are relatively small (and therefore easily interpretable), we use them going forward.

Support Vector Classifier (SVC)

We fit SVCs on our Era 1 and Era 2 datasets using the one-versus-one classification approach. We apply 10-fold cross-validation to find the optimal tuning parameter cost within the range 0.1, 1, 10, 100, 1000. This yields optimal parameters cost = 1,000 for Era 1 and cost = 10 for Era 2.

Listing A.4: Cross-Validation Results for Optimal Parameter Cost of Era 1

```
Parameter tuning of    s v m   :

- sampling method: 10-fold cross validation

- best parameters:
 cost
 1000

- best performance: 0.3227355

- Detailed performance results:
   cost     error dispersion
1 1e-01 0.3263636 0.02249381
2 1e+00 0.3263718 0.02378727
3 1e+01 0.3272645 0.02047354
4 1e+02 0.3245455 0.02061949
5 1e+03 0.3227355 0.01997285
```

Listing A.5: Cross-Validation Results for Optimal Parameter Cost of Era 2

```
Parameter tuning of    s v m   :

- sampling method: 10-fold cross validation

- best parameters:
 cost
   10

- best performance: 0.2970595

- Detailed performance results:
   cost     error dispersion
1 1e-01 0.3080217 0.03547532
2 1e+00 0.2978408 0.04240264
3 1e+01 0.2970595 0.03985912
4 1e+02 0.3009658 0.03867292
5 1e+03 0.3017470 0.03935543
```

Support Vector Machine (SVM) with Radial Kernel We fit SVMs with a radial kernel on our Era 1 and Era 2 datasets using the one-versus-one classification approach. We apply 10-fold cross-validation to obtain the optimal tuning parameters cost from the range {0.1, 1, 10, 100, 1000} and gamma from the range {0.001, 0.01, 0.1, 1, 2, 3}. We find that cost = 1 and gamma = 0.1 for both Era 1 and Era 2.

Our SVM with a radial kernel performed slightly better than the SVC for both Era 1 and Era 2. As a result, we implement the algorithm using the one-versus-all classification approach to see if our accuracy improves even more. This approach requires the construction of five different models, one for each class in Pos (i.e. C vs. All, PF vs. All, etc.). For each model, we must recode our Pos variable. When considering C vs. All, for example, we recode C as 1 and all other positions (PF, PG, SF, and SG) as 0. For each model, we apply 10-fold cross-validation to obtain optimal tuning parameters cost and gamma from the ranges listed above. We then use the optimal parameters to calculate predicted probabilities for each class. Finally, we assign the predicted class for each observation as the maximum from all five individual model probabilities.

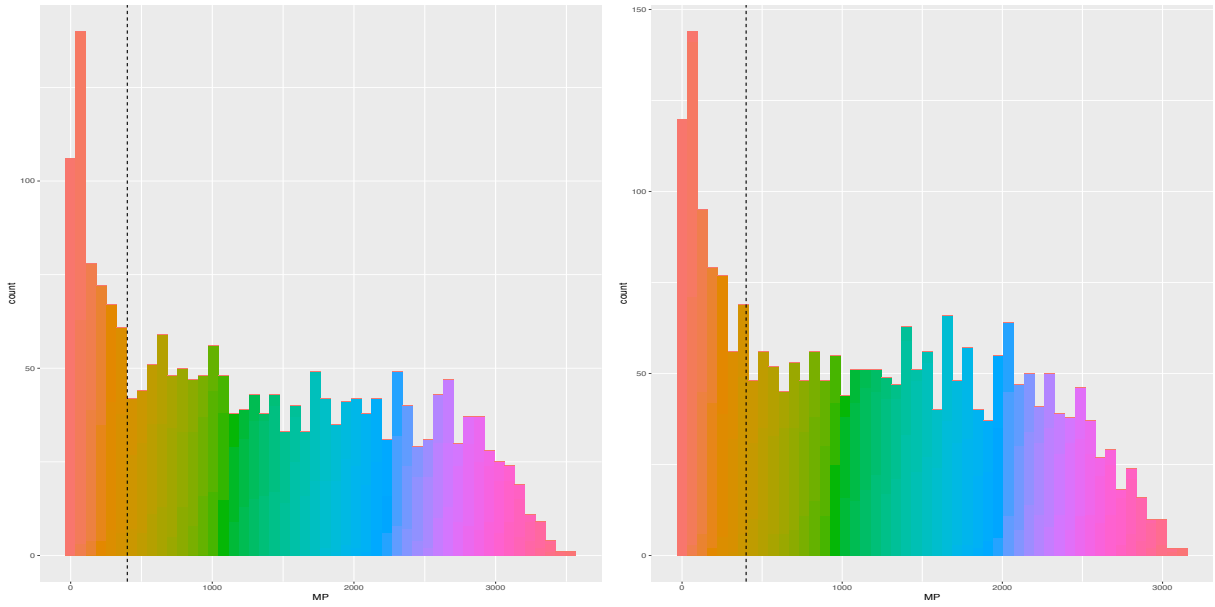
B Additional Tables and Figures

Table B.1: Description of Variables

Name	Description
Rk	Rank
Pos	Position
Age	Age of player at the start of February 1st of that season
Tm	Team
G	Games
GS	Games started
MP*	Minutes played
FG	Field goals per 36 minutes
FGA	Field goal attempts per 36 minutes
FG.	Field goal percentage
X3P*	3-point field goals per 36 minutes
X3PA* *	3-point field goal attempts per 36 minutes
X3P.*	3-point field goal percentage
X2P*	2-point field goals per 36 minutes
X2PA* *	2-point field goal attempts per 36 minutes
X2P.*	2-point field goal percentage
FT*	Free throws per 36 minutes
FTA* *	Free throw attempts per 36 minutes
FT.*	Free throw percentage
ORB*	Offensive rebounds per 36 minutes
DRB*	Defensive rebounds per 36 minutes
TRB* *	Total rebounds per 36 minutes
AST* *	Assists per 36 minutes
STL* *	Steals per 36 minutes
BLK* *	Blocks per 36 minutes
TOV* *	Turnovers per 36 minutes
PF* *	Personal Fouls per 36 minutes
PTS*	Points per 36 minutes

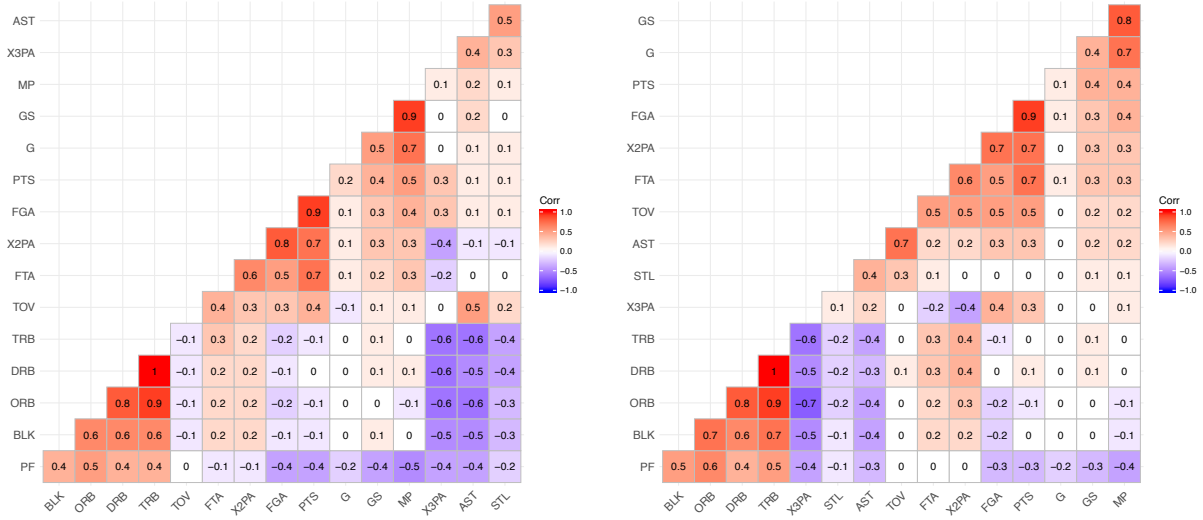
Note: Descriptions were taken from Basketball-Reference. Variables included in the final training and testing datasets are denoted by (*). When reading in the data, R changes variables with % to . (i.e. 3P% $\rightarrow$ 3P.). All variables starting with numbers now start with X (i.e. 3PA $\rightarrow$ X3PA).

Figure B.1: Histogram of Minutes Played for Era 1 (left) and Era 2 (right)



Note: We exclude all player-seasons with MP less than or equal to 400 (represented by the dashed line).

Figure B.2: Correlation Matrix Heat Map for Era 1 (left) and Era 2 (right)



Note: We deem a correlation too high if it exceeds 0.7. From this, we are able to remove the following variables from both datasets: ORB, DRB, G, GS, PTS, and FGA. We also remove these variables from the 2018 dataset. After this step, we are left with 10 predictors.

Table B.2: Balanced Datasets Era 1 (left), Era 2 (middle) and 2018 (right)

Postion	Freq	Percentage
C	295	18.65908
PF	344	21.75838
PG	324	20.49336
SF	311	19.67109
SG	307	19.41809

Postion	Freq	Percentage
C	360	19.69365
PF	381	20.84245
PG	369	20.18600
SF	327	17.88840
SG	391	21.38950

Postion	Freq	Percentage
C	73	19.51872
PF	69	18.44920
PG	84	22.45989
SF	62	16.57754
SG	86	22.99465

Note: A balanced dataset requires that the different levels of our response variable, Pos, are evenly represented in our datasets. While shooting guards (SG) are slightly underrepresented in our Era 1 and 2018 datasets (17.9% and 16.6%, respectively), our data is more or less balanced overall.

Table B.3: Summary of Results for Era 1 Dataset

Method	MER	Accuracy	Precision	Recall	F1	Kappa
1v1SVM (Radial Kernel)	0.265	0.735	0.735	0.729	0.727	0.668
SVC	0.272	0.728	0.727	0.722	0.720	0.660
RF	0.276	0.724	0.722	0.721	0.718	0.655
KNN	0.293	0.707	0.705	0.703	0.701	0.634
MLR	0.284	0.716	0.711	0.709	0.709	0.644
Bagging	0.320	0.680	0.683	0.678	0.676	0.600
RDA	0.291	0.709	0.711	0.706	0.704	0.637
Double Bagging	0.295	0.705	0.701	0.701	0.700	0.631
1vAll SVM (Radial Kernel)	0.297	0.703	0.696	0.697	0.694	0.628
LDA	0.309	0.691	0.701	0.687	0.686	0.613
QDA	0.320	0.680	0.684	0.677	0.665	0.600
Boosting	0.320	0.680	0.683	0.678	0.676	0.600
MDA	0.326	0.674	0.672	0.669	0.669	0.592
NB	0.337	0.663	0.677	0.664	0.642	0.580
Tree	0.356	0.644	0.660	0.639	0.637	0.555

Note: MER stands for misclassification error rate.

Table B.4: Summary of Results for Era 2 Dataset

Method	MER	Accuracy	Precision	Recall	F1	Kappa
RF	0.279	0.721	0.715	0.717	0.715	0.651
Bagging	0.302	0.698	0.692	0.694	0.693	0.622
1v1SVM (Radial Kernel)	0.313	0.687	0.678	0.681	0.678	0.608
SVC	0.317	0.683	0.675	0.679	0.675	0.603
MLR	0.321	0.679	0.671	0.675	0.671	0.599
1vAll SVM (Radial Kernel)	0.322	0.678	0.666	0.674	0.667	0.597
MDA	0.324	0.676	0.670	0.671	0.670	0.594
RDA	0.326	0.674	0.666	0.669	0.666	0.592
Double Bagging	0.332	0.668	0.662	0.665	0.663	0.585
Boosting	0.333	0.667	0.662	0.661	0.661	0.582
KNN	0.344	0.656	0.649	0.653	0.649	0.570
QDA	0.337	0.663	0.661	0.658	0.658	0.578
LDA	0.339	0.661	0.654	0.655	0.652	0.576
NB	0.364	0.636	0.629	0.632	0.626	0.544
Tree	0.379	0.621	0.643	0.609	0.611	0.523

Note: MER stands for misclassification error rate.

Table B.5: Variable Importance Measures for Era 1 (left) and Era 2 (right) MLR Models

Variables	Overall
TRB	16.290039
BLK	10.115129
STL	7.206230
AST	4.973704
PF	4.287331
X3PA	3.580499

Variables	Overall
TRB	19.945733
BLK	9.001371
STL	5.368123
AST	5.189474
PF	1.571665
X3PA	1.377068

Table B.6: Relative Risk Ratios for Era 1 MLR Models

Pos	(Intercept)	MP	X3PA	X2PA	FTA	TRB	AST	STL	BLK	TOV	PF
PF	12.286	0.878	2.655	1.007	1.269	0.886	1.095	2.363	0.302	0.766	0.686
PG	0.013	0.433	2.101	1.287	0.756	0.0003	33.481	14.334	0.011	0.974	0.209
SF	28.246	0.699	2.895	1.323	1.266	0.097	1.319	4.931	0.173	0.768	0.359
SG	3.216	0.563	2.222	1.452	1.480	0.003	2.989	8.070	0.071	0.834	0.267

Table B.7: Relative Risk Ratios for Era 2 MLR Models

Pos	(Intercept)	MP	X3PA	X2PA	FTA	TRB	AST	STL	BLK	TOV	PF
PF	8.128	1.176	1.610	0.952	0.864	0.435	0.650	1.475	0.326	0.854	0.766
PG	0.044	0.600	1.186	0.929	1.128	0.0001	41.823	6.264	0.027	0.588	0.869
SF	10.032	1.121	1.165	0.513	1.484	0.023	0.672	4.160	0.155	0.957	0.551
SG	2.355	0.774	1.782	1.095	1.071	0.001	1.875	5.579	0.090	1.012	0.566

Table B.8: Confusion Matrices for Era 1 and Era 2 MLR Models

Pos	C	PF	PG	SF	SG
C	60	22	0	4	0
PF	26	69	0	12	0
PG	0	0	96	1	11
SF	3	8	0	53	14
SG	0	0	13	21	62

Pos	C	PF	PG	SF	SG
C	81	33	1	2	0
PF	21	61	1	16	1
PG	0	2	102	2	10
SF	0	19	2	48	22
SG	0	1	15	28	81