Dhruv Modi

Econometrics of Big Data

Professor Payne

27 November 2023

NBA Rookie Classification

Dataset: The dataset I have chosen is to indicate whether an NBA rookie was active after 5 years. The dataset has 19 predictors which include games played, minutes, points per game, information on their field goals, three-pointers, free throws, rebounds, assists, steals, blocks, and turnovers.

Preliminary Classification Analysis:

- The three models I used for my preliminary classification analysis were Logistic Regression, Linear Discriminant Analysis, and K Nearest Neighbors

Logistic regression:

	True	False
True	74	71
False	48	206

- Error: 0.298
- Goodness of fit(Accuracy): 0.702

- Coefficients: array([0.04056581, -0.07495366, 0.17588046, -0.36259039, 0.11093667,
  0.0291062, 0.88888584, -0.19853491, -0.00316335, -0.0118813,
  0.03733453, 0.01342029, 0.86947068, -0.35066541, 0.25003203,
  0.48865444, -0.28379245, 0.33131076, -0.65682773])
- Standard error:

```
Out[12]: gp          0.570026
         min         0.272790
         pts         0.144554
         fgm         0.055694
         fga         0.118037
         fg          0.201125
         3p_made     0.012342
         3pa         0.033916
         3p          0.510389
         ftm         0.033101
         fta         0.044175
         ft          0.340302
         oreb        0.025644
         dreb        0.045370
         reb         0.068500
         ast         0.046990
         stl         0.013126
         blk         0.014500
         tov         0.023783
```

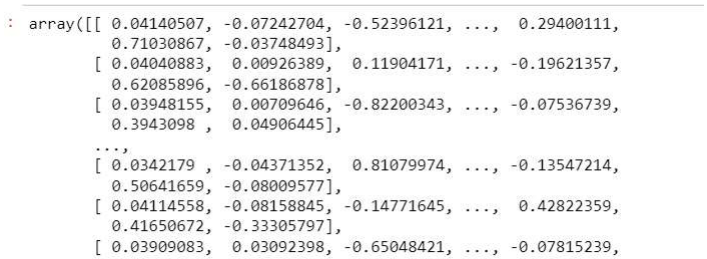
- Average Accuracy of 5 fold cross validation: 0.713

LDA:

	True	False
True	70	75
False	48	206

- Error: 0.308
- Goodness of fit(Accuracy): 0.692
- Coefficients: array([0.04643369, -0.06393741, 0.02007266, -0.76808942, 0.44159072,

0.05793012, 2.61270167, -0.83608748, -0.00546223, -0.09425374,0.18649098, 0.01664003, 1.36902285, 0.37018522, -0.47257847, 0.39240727, -0.10428514, 0.29336842, -0.66182984])

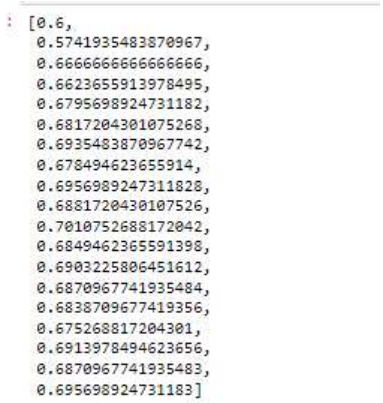


- Standard Error using Bootstrap:
- Average Accuracy of 5 fold cross validation: 0.716

KNN:

	True	False
True	74	71
False	67	187

- Error: 0.346
- Goodness of fit(Accuracy): 0.692
- Coefficients: N/A, since KNN is a nonparametric model, we do not assume a form



- Average Accuracy of 5 fold cross validation(K = 1-20):

Initial Analysis:

- When looking solely at the confusion matrices produced after training the data and running it on the test set, the best model was logistic regression. It had a lower error and higher accuracy than both of the other models.
- After running 5-fold cross-validation, the best model was LDA since it had a higher accuracy than logistic regression or any of the KNN models. LDA was only slightly better than logistic regression having only a 0.03 better accuracy.
- The best KNN model was when K=11, but accuracy was lower than Logistic Regression or LDA 5 fold CV accuracy.

Random Forest:

	True	False
True	80	59
False	55	205

- Error: 0.303
- Goodness of fit(Accuracy): 0.697
- Average Accuracy of 5 fold cross validation: 0.686

	Feature	Coefficient
0	gp	0.102259
1	min	0.071346
2	pts	0.073979
3	fgm	0.057952
4	fga	0.056218
5	fg	0.078558
6	3p_made	0.020191
7	3pa	0.032141
8	3p	0.040613
9	ftm	0.057116
10	fta	0.046469
11	ft	0.065106
12	oreb	0.043461
13	dreb	0.047075
14	reb	0.057757
15	ast	0.043236
16	stl	0.033498
17	blk	0.031619
18	tov	0.041407

- Feature Importance:
    - The most important features are games played, minutes played, and points. The least important are turnovers, blocks, and steals.
- Standard Error from Bootstrapping with Importance:

```
array([0.01235294, 0.00311001, 0.00636199, 0.00589213, 0.00245713,
       0.0086069 , 0.00293248, 0.00345454, 0.00462219, 0.0056793 ,
       0.00537389, 0.0042667 , 0.00706565, 0.00269194, 0.00689715,
       0.00178933, 0.00260127, 0.00309091, 0.0018385 ])
```

Lasso Regression:

	True	False
True	83	76
False	53	187

- Error: 0.39

- Goodness of fit(Accuracy): 0.61
- Average Accuracy of 5 fold cross validation: 0.697

	Feature	Coefficient
0	gp	0.031365
1	min	-0.044527
2	pts	0.628446
3	fgm	0.000000
4	fga	-0.499695
5	fg	-0.023647
6	3p_made	0.000000
7	3pa	0.000000
8	3p	0.000863
9	ftm	0.000000
10	fta	-0.221124
11	ft	-0.009575
12	oreb	0.548419
13	dreb	0.000000
14	reb	0.000000
15	ast	0.163005
16	stl	-0.057991
17	blk	0.180345
18	tov	0.000000

- Coefficients:
- As we can see fgm, 3ptmade, 3pa, ftm, dreb, reb, and tov were all shrunk to zero showing that according to Lasso those are not significant predictors and are not necessary to include in our model. In other words, they do not have a major effect on whether a rookie is in the NBA in 5 years.
- Standard Error with Bootstrapping:

```
array([0.0001176 , 0.00148482, 0.01461774, 0.03227086, 0.01049383,
       0.00120824, 0.03872357, 0.0126029 , 0.00012753, 0.01748201,
       0.00962478, 0.00043743, 0.01624713, 0.01643634, 0.01718153,
       0.00384349, 0.00908502, 0.01096421, 0.0046833 ])
```

PCA:

	True	False

True	71	68
False	54	206

- Error: 0.306
- Goodness of fit(Accuracy): 0.694
- Coefficients: array([-0.04764612, -0.0185592 , 0.00525772])
- Average Accuracy of 5 fold cross validation: 0.697

Final Analysis: During our initial analysis the best model was LDA when running 5-fold cross-validation. After running three more models that used different methods and provided different insights, LDA was still the best model. It seemed that when we made our models more complicated it did have a worse effect on predicting the data correctly. Even though LDA had the best average after 5-fold cross-validation, each model was between 0.67 to 0.71 accuracy so you could not go wrong on which one you choose. As I had mentioned before, each model provided different insights. For example, Lasso regression told us which features were not significant by shrinking them to zero, and Random Forest allowed us to analyze the importance of features. Furthermore, PCA provided us with a much more compact model, while keeping a very similar accuracy to the other models. An interesting insight was that though Random Forest said turnover, blocks, and steals were the least important, Lasso only shrunk turnovers to zero which highlights the importance of considering multiple models when analyzing data. Each model has its strengths and limitations and one needs to interpret the results within the context of the dataset characteristics. Along with this, I was able to obtain the Standard Error of a few of the models

using Bootstrapping. For example, I obtained the Standard Error of the potential Random Forest Coefficients using the importance function in Python. The Standard Error of the coefficients shows us the amount of variability in the coefficient estimate across different samples of the data. In other words, it gives us the precision of the coefficient estimates. One would need to look at every predictor and every Standard Error to understand which model was best for which predictor, but from a high-level view Random Forest had the lowest Standard Error for most of the coefficients. This can suggest that Random Forest is catching consistent patterns in the data and could potentially be the most robust model. To conclude, LDA was the model with the highest accuracy, but each model provided different insight into the dataset and I would say there is no clear recommendation for which model fits the data the best.

- The numbers provided in the code may be a little different than this as I ran the code a few times so it could have changed slightly.

References

Bootstrap Sampling in Python | DigitalOcean. (n.d.). Www.digitalocean.com. Retrieved December 7, 2023, from https://www.digitalocean.com/community/tutorials/bootstrap-sampling-in-python

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning*. Springer Nature.

NBA Rookie Longevity Project. (n.d.). Kaggle.com. Retrieved December 7, 2023, from https://www.kaggle.com/code/mamadoudiallo/nba-rookie-longevity-project

sklearn.discriminant_analysis.LinearDiscriminantAnalysis — scikit-learn 0.24.1 documentation. (n.d.). Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html

sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis. (n.d.). Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis.html