HEADHUNTER - NBA

Daniel M. M. da Serra

Universidade Federal Fluminense

Objective

- Create a predictor of athletes who have a promising future in the NBA, based on their freshman year statistics.
- It is considered an athlete with a promising future, the one who stays at least five years in the league.

Methodology

- Database with 1340 observations and the variables are the freshman year statistics of the athlete.
- The result variable **(TARGET_5Yrs)** indicates whether the player stayed in the league for 5 years or more.

Methodology

- Two more databases were used to complement the analyses:
  - One of then with the same observations as the principal database (the one with the result variable) + the draft year of the player.
  - The other had a lot of variables, but it was used only the player height and college.

Methodology

- The only variable with *N/A* was **X3P.** that indicates the percentage of the converted 3 point baskets. As the database also offered the number of baskets of 3 attempted and the amount converted, it was possible to fill in these *N/A*.
- The data also presents information about players drafted from 2013. So, regardless of their statistics, it was impossible, in 2016, for any of those to be five years or more in the NBA.

Methodology

- There are two variables left with *missing values*:
  - **player_height** = the height of the player in centimeters
  - **college** = the college attended by the player
- The **kNN** function from the **VIM** package solve these *missing values*.

Methodology

- To create the model, the variables **Name**, **Year Drafted**, **REB** and all the ones that corresponds to attempted and converted were removed.
- Result variable: **numeric** $->$ **logical**
- 'college' variable: **character** $->$ **factor**

About the database

- There were some mistakes in the data:
  - The same player with the same statistics but two different results for **TARGET_5Yrs**.
- These mistakes were solved to improve the model performance.

Pre-processing

In the pre-processing, apart from, **kNN**, were tested:

- **nearZeroVar**
  - Variances zero or near zero
- **Correlation**
  - Verify correlated columns
  - It was found one: **MIN**, but the model worked better with this information.

Model

- Database divided into TRAIN (75%) and TEST (20%)
- Proportion defined by the client
- The GRADIENT BOOSTING method was also determined by the client
- Hyper parameters evaluated:
  - Interaction depth: 1, 2, 3, 4, 5
  - Shrinkage: 0.05, 0.01, 0.1
  - Number of trees: 25, 50, 75, 100
- Total of 60 models evaluated
- Choice method: higher AUC (area under the curve)

Results

Some results:

interactionDepth <dbl>	n.trees <dbl>	shrinkage <dbl>	AUC <dbl>
NA	*NA*	*NA*	*NA*
1	25	0.05	0.7998070
1	25	0.01	0.7263854
1	25	0.10	0.8107803
1	50	0.05	0.8132616
1	50	0.01	0.7723187
1	50	0.10	0.8166805
1	75	0.05	0.8183899
1	75	0.01	0.7850014
1	75	0.10	0.8166805
1	100	0.05	0.8222222
4	75	0.01	0.7969672
4	75	0.10	0.8262476
4	100	0.05	0.8220568
4	100	0.01	0.8078853
4	100	0.10	0.8262476
5	25	0.05	0.8316515
5	25	0.01	0.7795975
5	25	0.10	0.8284808
5	50	0.05	0.8388475

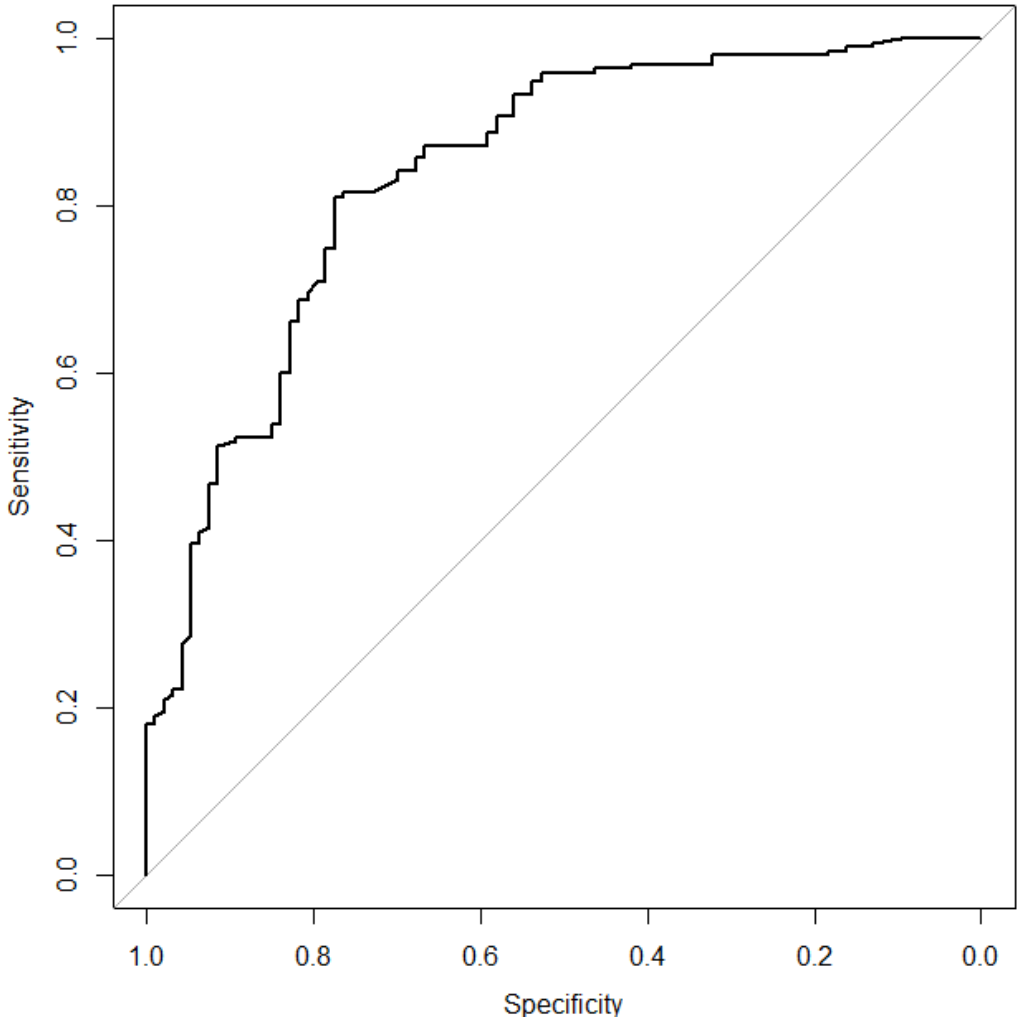
Final model

Were chosen for the final model:

- n.trees = 50
- interaction.depth = 5
- shrinkage = 0.05

Behavior in the TEST sample

The model generated an AUC = 0.8388475 in the TEST sample



Behavior in the TEST sample

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE    72   38
     TRUE     21  157

               Accuracy : 0.7951
                 95% CI : (0.7439, 0.8402)
    No Information Rate : 0.6771
    P-Value [Acc > NIR] : 5.815e-06

                  Kappa : 0.5529

 Mcnemar's Test P-Value : 0.03725

            Sensitivity : 0.7742
            Specificity : 0.8051
         Pos Pred Value : 0.6545
         Neg Pred Value : 0.8820
             Prevalence : 0.3229
         Detection Rate : 0.2500
   Detection Prevalence : 0.3819
      Balanced Accuracy : 0.7897

       'Positive' Class : FALSE
```