Data Mining and Machine Learning II

Project Assessment - NBA Basketball and Rookies

Robert Shepherd
2431907s

Contents

Introduction

A typical NBA basketball player has a career length of around 5 years. During the first year they are classed as a 'rookie'. This is a particularly challenging year for most rookies as they look to adapt to the intensity of the NBA. Their success (or failure) during this year plays a big part in determining how their career develops.

This analysis looks to determine what information captured during players rookie year is indicative of whether their career will last more than 5 years. To do this, a dataset has been provided of 600 players who played their rookie season between 1980 and 2011.

This dataset contains:

1. Descriptive information (e.g. year drafted and player name)
2. Rookie year player statistics (e.g. games played and points per game)
3. Career length statistics (e.g. career length in years and whether the years played was over 5)

A dictionary of variables contained within the data can be found below as well as the role they assume within the model build process[1]:

Variable	Description	Role
Name	Player name	Description
Year_drafted	The year the player was drafted	Description
GP	Games played (out of 82)	Feature
MIN	Minutes per game (out of 48)	Feature
PTS	Points per game	Feature
FG_made	Field goals made (per game)	Feature
FGA	Field goal attempts (per game)	Feature
FG_percent	Field goal percentage	Feature
TP_made	Three points made (per game)	Feature
TPA	Three point attempts (per game)	Feature
TP_percent	Three point percentage	Feature
FT_made	Free throws made (per game)	Feature
FTA	Free throws attempts (per game)	Feature
FT_percent	Free throws percentage	Feature
OREB	Offensive rebounds (per game)	Feature
DREB	Defensive rebounds (per game)	Feature
REB	Total rebounds (per game)	Feature
AST	Assists (per game)	Feature
STL	Steals (per game)	Feature
BLK	Blocks (per game)	Feature
TOV	Turnovers (per game)	Feature
Yrs	Career length (in years)	Description
Target	1 if Yrs>5 and 0 otherwise	Target

[1] Additional features are developed as part of the model build process however these are extrapolated from this information

The analysis will build a series of penalised regression models in order to:

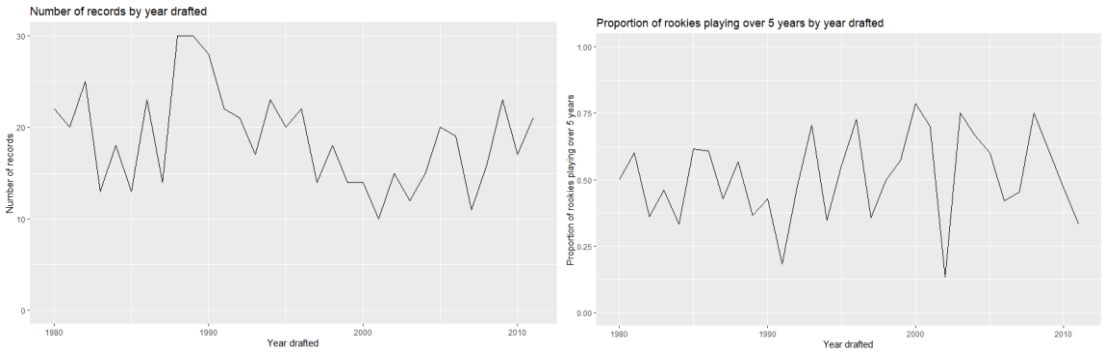
1. Identify which model type is most predictive of which players' careers will last over 5 years
2. Identify which variables common across different models are most predictive of which players' careers will last over 5 years

Exploratory Analysis

Variance over time

Before starting on exploring and engineering the features, it is worth evaluating whether the data has changed significantly over time.

The below chart shows number of records and average career length by year:



While the number of records varies by year (left chart above), this does not suggest there is any skew or missing years that would cause bias within our data.

In order to get a fair comparison over the years we would need to leave enough time for players careers to mature otherwise our data would be biased. While the maximum career declines up to 2011 it is still above 6 years. It is therefore assumed that the data records any players that are still playing as over 5 years even if their career has not ended. This assumption is validated by observing the variation in the proportion of successes which stays (relatively) consistent up to 2011 (right chart above).

Partitioning data

Two options were considered for partitioning data:

1. Creating a train/test split randomly
2. Holding out X years' worth of data as a 'test' partitioning and using the rest to train the model

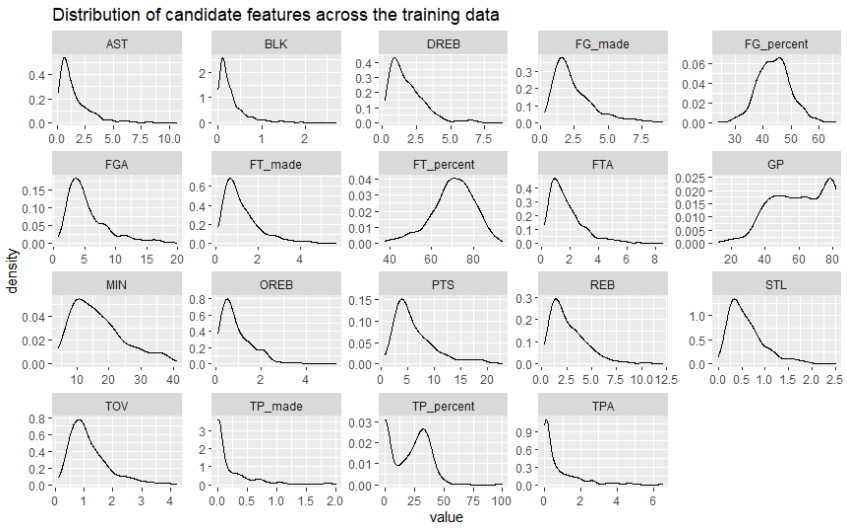
Option 2 may have been a better fit if the goal of the analysis was to estimate the career length of future rookies and if the sample of data was large enough that cutting off the latest year would give robust samples. However, as the dataset is small and we are interested in explaining the impact of variables to date, the data has been partitioned randomly as below:

- Training (70%): 420 rows of which 207 had a career length > 5 years (49% of rows)
- Test (30%): 180 rows of which 93 had a career length > 5 years (52% of rows)

Understanding features and additional feature engineering

Checking a summary of the data showed that there were no missing values.

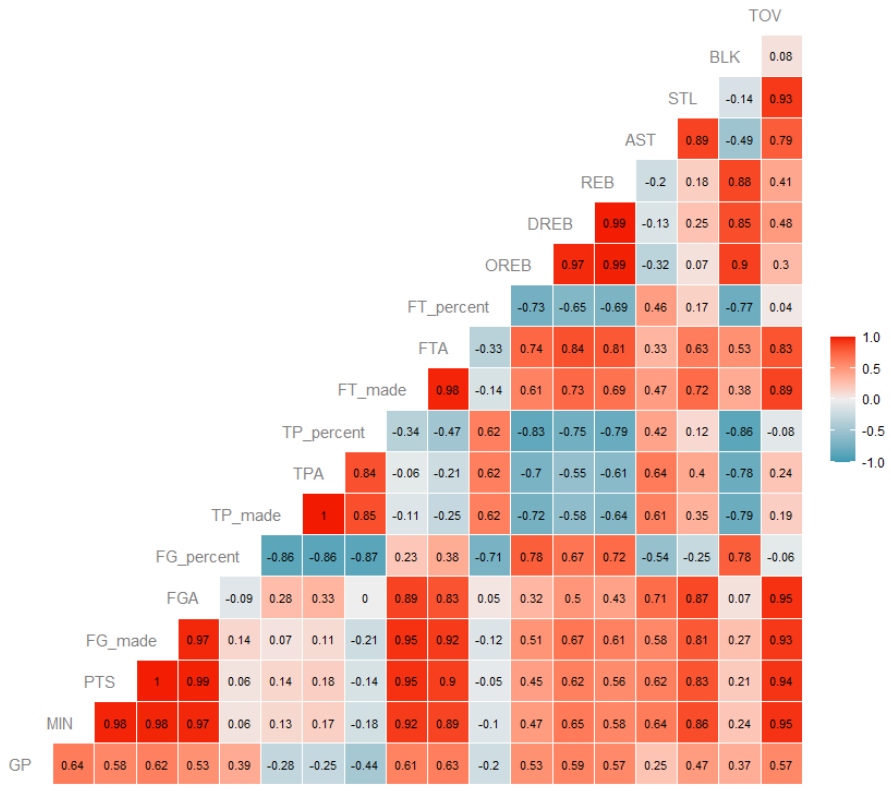
To understand how the variables are distributed, the below chart shows an estimation of density by variable across the training data:



From these distributions it can be observed that:

1. AST, BLK, DREB, FG_made, FGA, FT_made, FTA, MIN, OREB, PTS, REB, STL, TOV have a significant right skew
2. TP_made and TPA appear to be zero inflated and have a strong right skew
3. GP has a significant left skew
4. FT_percent and FG_percent are approximately normally distributed

To understand if there is a risk of multicollinearity, the below chart shows the pairwise correlations across the training features:



This plot shows that there is a high amount of correlation between many variables which suggests penalised regression methods will be a good fit.

The amount of correlation is intuitive as many of the variables will correlate with the number of minutes played. To remove as much correlation as possible before introducing penalised methods, the below treatments have been applied before modelling:

Feature	Description	Treatment	Transformation applied
GP	Games played (out of 82)	No treatment	None
MIN	Minutes per game (out of 48)	No treatment	None
PTS	Points per game	Transformed to PTS_minute	PTS / MIN
FG_made	Field goals made (per game)	Removed	None
FGA	Field goal attempts (per game)	FGA_minute	FGA / MIN
FG_percent	Field goal percentage	No treatment	None
TP_made	Three points made (per game)	Removed	None
TPA	Three point attempts (per game)	Transformed to TPA_minute	TPA / MIN
TP_percent	Three point percentage	No treatment	None
FT_made	Free throws made (per game)	Removed	None
FTA	Free throws attempts (per game)	Transformed to FTA_minute	FTA / MIN
FT_percent	Free throws percentage	No treatment	None
OREB	Offensive rebounds (per game)	Removed	None
DREB	Defensive rebounds (per game)	Removed	None
REB	Total rebounds (per game)	Transformed to REB_minute	REB / MIN
AST	Assists (per game)	Transformed to AST_minute	AST / MIN
STL	Steals (per game)	Transformed to STL_minute	STL / MIN
BLK	Blocks (per game)	Transformed to BLK_minute	BLK / MIN
TOV	Turnovers (per game)	Transformed to TOV_minute	TOV / MIN

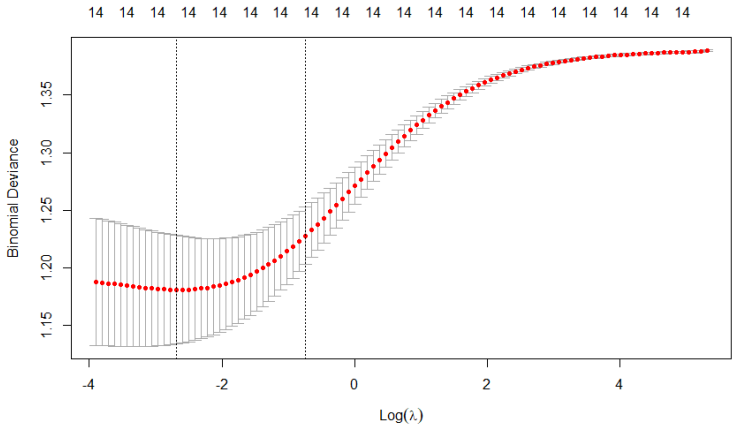
'FG_made', 'TP_made' and 'FT_made' were removed as the information within the fields are contained within other variables (i.e. attempts * percent). 'OREB' and 'DREB' were remove as they had a very high correlation with 'REB' (which would act as the total).

After transformation, variables were scaled so that each feature had a mean 0 and standard deviation of 1.

Results

Ridge regression

To find the optimum lambda parameter, cross validation is used to assess the deviance over different values of log lambda while fitting a Ridge regression:



The two dotted lines represent two values of interest for the Ridge regression lambda parameter:

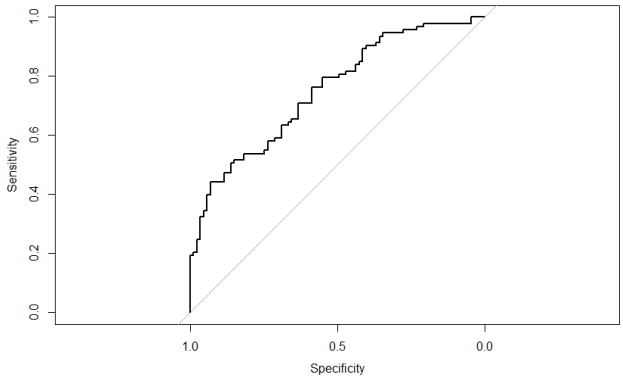
1. The log lambda where the binomial deviance is lowest (lambda equal to 0.068)
2. The log lambda within one standard error of the one with minimum binomial deviance (lambda equal to 0.479)

A Ridge regression model was subsequently fit using the second value (within one standard error). The most influential coefficients were:

1. Minutes played: 0.219
2. Games played: 0.179
3. Field goal percentage: 0.111

This returned the below performance:

```
Confusion Matrix and Statistics

predictions.ridge.class  0  1
                      0 84 66
                      1  3 27

               Accuracy : 0.6167
                 95% CI : (0.5414, 0.688)
    No Information Rate : 0.5167
    P-Value [Acc > NIR] : 0.004367

                  Kappa : 0.25

 Mcnemar's Test P-Value : 8.398e-14

            Sensitivity : 0.9655
            Specificity : 0.2903
         Pos Pred Value : 0.5600
         Neg Pred Value : 0.9000
             Prevalence : 0.4833
         Detection Rate : 0.4667
   Detection Prevalence : 0.8333
      Balanced Accuracy : 0.6279

       'Positive' Class : 0
```

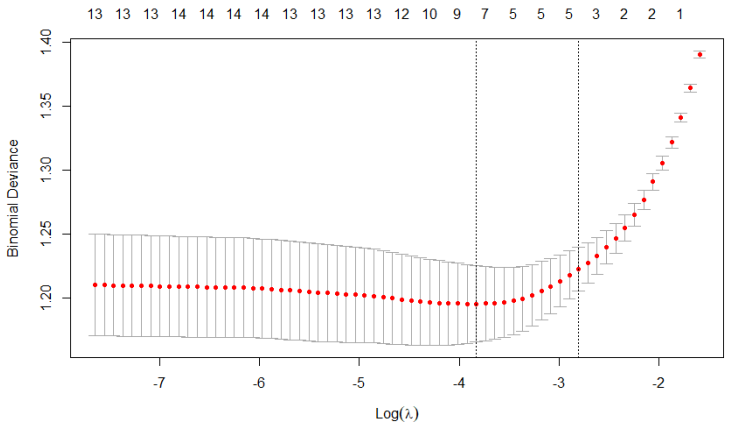


The right chart shows the ROC curve with an AUC of 0.750, quite strong model performance (compared to a null model of 0.5).

The table on the left shows the confusion matrix and some descriptive statistics on performance. This shows that the model is strong at identifying players whose career does not last more than 5 years (sensitivity of 0.97) but is quite poor at identifying those that do (specificity of 0.29). Overall, the model is 62% accurate at estimate which rookies' careers will last more than 5 years.

Lasso regression

To find the optimum lambda parameter, cross validation is used to assess the deviance over different values of log lambda while fitting a Lasso regression:



The two dotted lines represent two values of interest for the Lasso regression lambda parameter:

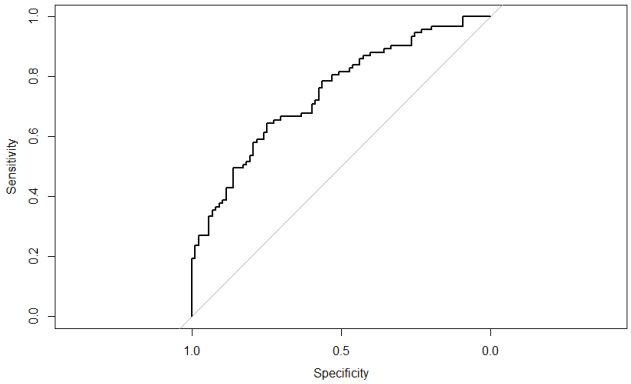
1. The log lambda where the binomial deviance is lowest (lambda equal to 0.021)
2. The log lambda within one standard error of the one with minimum binomial deviance (lambda equal to 0.060)

A Lasso regression model was subsequently fit using the second value (within one standard error). The most influential coefficients were:

1. Minutes played: 0.487
2. Games played: 0.178
3. FG_percent: 0.077

This returned the below performance:

```
Confusion Matrix and Statistics

predictions.lasso.class  0  1
                      0 80 60
                      1  7 33

               Accuracy : 0.6278
                 95% CI : (0.5527, 0.6985)
    No Information Rate : 0.5167
    P-Value [Acc > NIR] : 0.001722

                  Kappa : 0.2691

 Mcnemar's Test P-Value : 2.114e-10

            Sensitivity : 0.9195
            Specificity : 0.3548
         Pos Pred Value : 0.5714
         Neg Pred Value : 0.8250
             Prevalence : 0.4833
         Detection Rate : 0.4444
   Detection Prevalence : 0.7778
      Balanced Accuracy : 0.6372

       'Positive' Class : 0
```



The right chart shows the ROC curve with an AUC of 0.745, slightly below that of the Ridge regression.

The table on the left shows the confusion matrix and some descriptive statistics on performance. Compared to the Ridge regression, the overall accuracy is slightly higher (0.628). It is generally better at estimating which rookies' careers will last more than 5 years (specificity of 0.36) but less accurate at estimating those that do not (sensitivity of 0.92).

Elastic net

The final model built was an Elastic net model. To get the optimal values for the alpha and lambda hyperparameters, six-fold cross validation was used (to get 100 values per fold) with 100 different values tested for each parameter.

The cross validation estimated the below as optimal parameters:

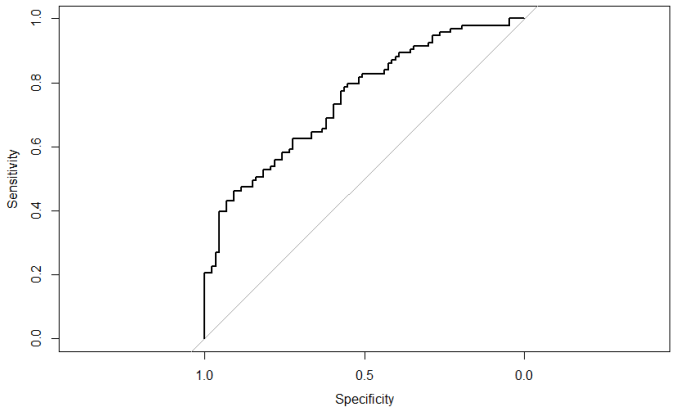
- Alpha: 0.491
- Lambda: 0.017

Fitting the Elastic net with these parameters identified the below coefficients as the most influential:

1. Minutes played: 0.572
2. Games played: 0.324
3. Rebounds per minute: 0.286

The model returned the below performance:

```
Confusion Matrix and Statistics

predictions.net.class  0  1
                    0 62 35
                    1 25 58

               Accuracy : 0.6667
                 95% CI : (0.5927, 0.735)
    No Information Rate : 0.5167
    P-Value [Acc > NIR] : 3.261e-05

                  Kappa : 0.3351

 Mcnemar's Test P-Value : 0.2453

            Sensitivity : 0.7126
            Specificity : 0.6237
         Pos Pred Value : 0.6392
         Neg Pred Value : 0.6988
             Prevalence : 0.4833
         Detection Rate : 0.3444
   Detection Prevalence : 0.5389
      Balanced Accuracy : 0.6681

       'Positive' Class : 0
```



The right chart shows the ROC curve with an AUC of 0.749, on par with below that of the Ridge regression (0.750) and above the Lasso regression (0.745).

The table on the left shows the confusion matrix and some descriptive statistics on performance. Compared to the Ridge and Lasso regression, the overall accuracy is higher (0.667). It is significantly better at estimating which rookies' careers will last more than 5 years (specificity of 0.624) but also significantly less accurate at estimating those that do not (sensitivity of 0.713).

Comparison

	Ridge regression	Lasso regression	Elastic net
Accuracy	0.617	0.628	0.667
Specificity	0.290	0.355	0.624
Sensitivity	0.966	0.920	0.713
AUC	0.750	0.745	0.749

Main findings:

1. Both the Ridge and Lasso regression performed similarly with high sensitivity but low specificity meaning it struggled to identify players whose careers lasted more than 5 years

2. The Elastic net was more accurate and was significantly better at identifying players whose careers lasted more than 5 years, but at a cost of decreased sensitivity
3. All three models had similar AUC

Discussion

Referring to the original goals of the study:

1. Identify which model type is most predictive of which players' careers will last over 5 years

   The Elastic net model is preferred as it is the most accurate and the most effective at identifying players whose careers lasted over 5 years. This would be a very useful model if the goal of the model was to identify rookies that were most likely to have long careers.

   If there is a strong risk with respect to making a wrong 'success' prediction (such as tying down players on long contracts who do not perform) then the Ridge or Lasso regression models may be more fit for purpose as they are more risk averse (i.e. they are more accurate at predicting players with shorter careers)

2. Identify which variables common across different models are most predictive of which players' careers will last over 5 years

   The two most important variables across all models were how many minutes and games were played. This can be interpreted as saying the more a rookie played during their debut season the more likely they were to have a long career.

   The models disagreed on what other variables were important but generally field goal percentage and rebounds per minute scored highly.

There were no major concerns with the data though a larger sample would likely improve model performance. The exploratory analysis showed that the data was censored correctly up to 2011 (i.e. there was a minimum of 6 years observed time after each rookie season).

The models are thought to be valid as the accuracy and AUC on unseen data was high. However, the exploratory analysis did reveal a high amount of multicollinearity and non-normally distributed data.

The models could potentially be improved in future iterations by testing the use of monotonic transformations on features to 'normalise' the distribution prior to creating the model. Additionally, more feature selection can be done prior to creating the models to reduce the variables further to avoid the effect of multicollinearity which may improve our models' ability to extrapolate on to new data.

Appending more additional features could also improve our ability to predict career length. This could include personal information, such as height, or career statistics before joining the NBA, such as college statistics (where available).

Finally, other classification models could be tested such as logistic regression, support vector machines and ensemble decision trees (though the latter may mean we lose some interpretability).