Prediction of NBA rookies' career length

Preparer: Alexey Pankratov

Introduction

This project aims to use the regularised regression models such as lasso regression, ridge regression, and elastic nets to identify the variables that predict whether a basketball player will play more than five seasons in the NBA.

The dataset is represented by an extract from the database on NBA rookies and consists of 600 players drafted from 1980 to 2011. The dataset contains the following variables:

Variable name	Variable description
Year_drafted	The year the player was drafted
GP	Games Played (out of 82)
MIN	Minutes per game (out of 48)
PTS	Points per game
FG_made	Field goals made (per game)
FGA	Field goal attempts (per game)
FG_percent	Field goal percentage
TP_made	Three points made (per game)
TPA	Three point attempts (per game)
TP_percent	Three point percentage
FT_made	Free throws made (per game)
FTA	Free throws attempts (per game)
FT_percent	Free throws percentage
OREB	Offensive rebounds (per game)
DREB	Defensive rebounds (per game)
REB	Total rebounds (per game)
AST	Assists (per game)
STL	Steals (per game)
BLK	Blocks (per game)
TOV	Turnovers (per game)
Yrs	Career length (in years)
Target	Is career length more than 5 years? (1 – yes, 0 – no)

All predictor variables are continuous except for the variable "Year_drafted" which is an ordered categorical variable. Response variable "Target" takes one out of two values:

- 1, if a rookie's career length (variable "Yrs") is more than five seasons/years, and
- 0, if otherwise.

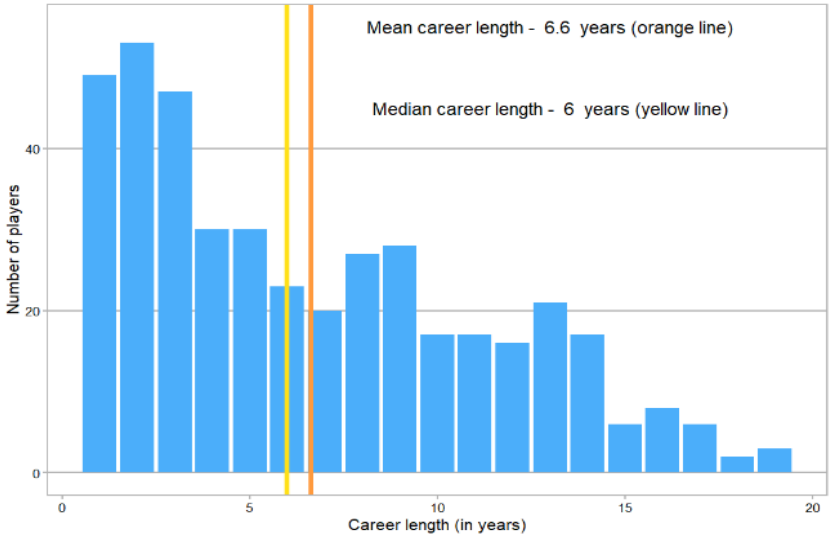
It should be noted that as the response variable "Target" is derived directly from the variable "Yrs" I will not consider variable "Yrs" in the models.

In this project, I will separate the dataset into training and test in proportions of 70/30 percent, respectively. I will then perform exploratory analysis on the training data set. Build the models mentioned above and decide which model better predicts if a rookie has a career for more than five years or not and check which predictor variables are more important in each model.

Exploratory analysis

As mentioned above, I first split the data into training (70% of all data) and test data (30% of all data). The total number of observations in the datasets is 600, so the training set contains 420 observations, and the test set contains 180 observations. I already partially explored the variables mentioning that most variables are continuous, whereas others are ordered categorical. The below exploratory analysis is done on the training set.

Figure 1 - Distribution of career length



The training set is relatively equally split between rookies whose careers lasted more than 5 seasons (211 observations) and those whose careers were no longer than 5 seasons (209 observations).

Figure 1 shows that the median career length of rookies is 6 years, whereas the mean career duration is 6.6 years. Mean value exceeding the median is due to some players had carriers of up to 19 years.

As presented in Figure 2, rookies were drafted between 1980 and 2011. There are more rookies who have been drafted before the mid-1990s than during the period from 1996 to 2011.

Figure 2 - Distribution of drafting years

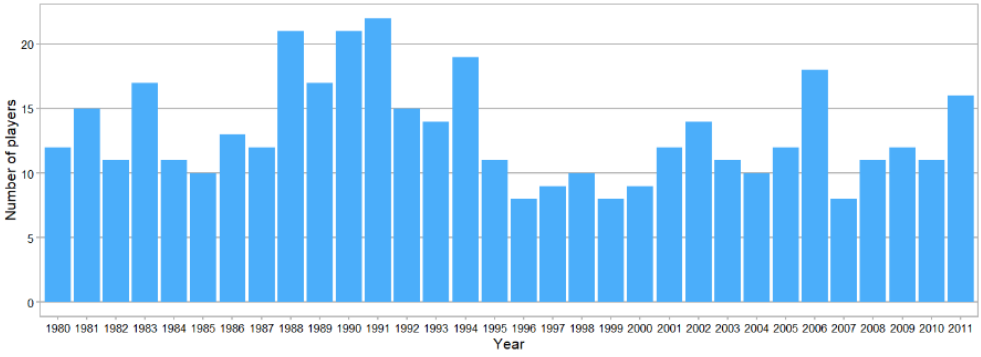


Figure 3 shows a matrix of Pearson correlation coefficients among pairs of continuous variables of the dataset. The coefficients take value in the range from -1 to 1 and measure the strongness of a linear relationship between two variables. Correlation coefficients of 1 and -1 mean perfect positive and negative correlation, respectively, and a correlation coefficient of 0 means no linear correlation between pair of variables.

I can spot that no single variable has a strong correlation with Career length (YRS) as all correlation coefficients are below 0.44. Some groups of other variables show very strong positive correlations with each other:

- Minutes per game (MIN), Points per game (PTS) and Field goals made (FG_made), Field-goals attempts (FGA), Free-throws attempted and made (FTA and FT_made), and the number of turnovers per game (TOV_ has a correlation with each other between 0.81 and 0.99.
- The number of rebounds, both offensive and defensive (variables OREB, DREB, REB), has a very strong correlation with each other (between 0.86 and 0.98) but also, they strongly correlate with almost all variables mentioned in the previous point with the exception of TOV and have correlation coefficients ranging between 0.61 and 0.80 with these variables.
- The number of blocks (BLK) strongly correlates with rebounds REB. OREB, DREB (between 0.66 and 0.69). Assists and steals per game (AST and STL) strongly correlate with each other (0.73) and relatively strongly correlate with minutes and points per page (MIN and PTS) as well as with field-goal attempts and made (FGA and FG_made) - correlation coefficients are between 0.55 and 0.75.

Figure 3 - Correlation matrix

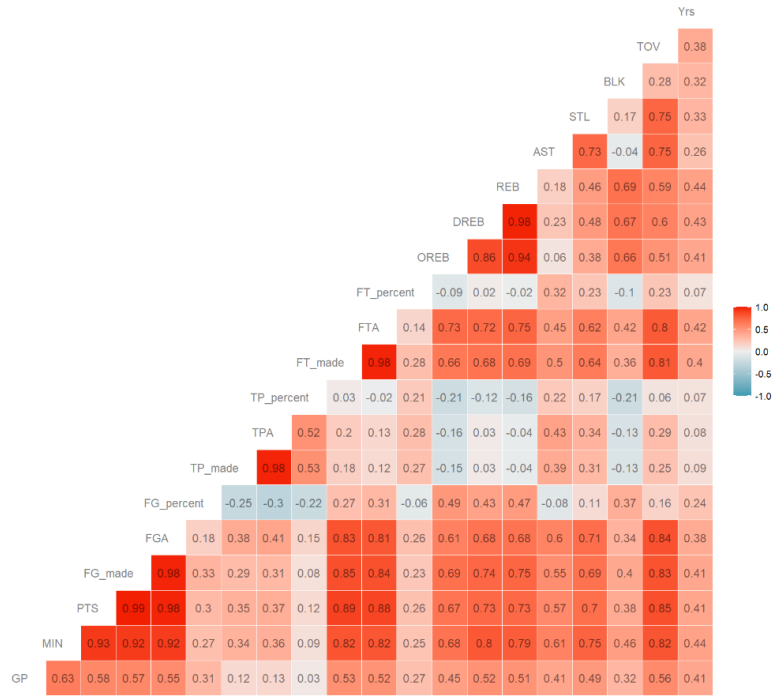


Figure 4 - Probability distribution of predictor variables (part 1)

Orange and yellow horizontal lines on the above plot pertain to mean and median values, respectively.

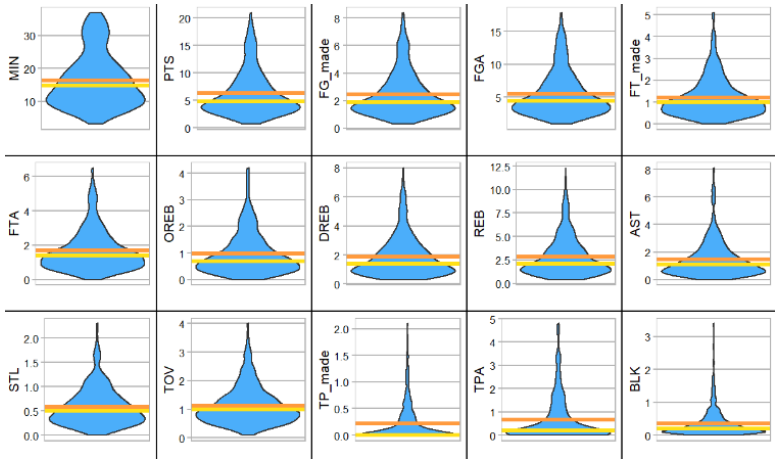
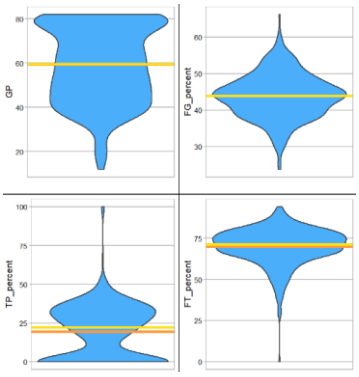


Figure 5 - Probability distribution of predictor variables (part 2)

Orange and yellow horizontal lines on the above plot pertain to mean and median values, respectively.



It should be noted that there is no strong negative correlation among the variables in the dataset. The most noticeable negative correlation is between field goals percent (FG_percent) and three points attempts (TPA, correlation -0.30), three points made (TP_made, correlation -0.25), and three points percent (TP_percent, correlation -0.22).

Figure 4 shows the probability distribution of most of the continuous predictor variables in the data except for Points per game (variable GP) and percent of different types of strikes (variables FG_percent, TP_percent, and FT_percent) that are presented later in this section of the report. Violin plots in the figure highlight the similarity of shapes of the probability distribution. Wide bottoms of the plots show that the values for these variables are distributed more towards lower values. At the same time, small lines that go towards the top of the plots reflect the observations with higher values, i.e., "top performers" who played more minutes per game, made more shot attempts and strikes, etc. These observations drive the mean of the variables above its median. This pattern is similar for all variables in Figure 4, but it is less pronounced for MIN and much more pronounced for TP_made, TPA, and BLK. Considering the strong correlation among many of these variables already presented in Figure 3, this may indicate that "top performers" observations may have high values in many variables at the same time.

Figure 5 shows that the probability distribution of the remaining four continuous variables shows the pattern that is different from the one described above:

- The probability distribution of free throws percent (FT_percent) and games played (GP) is skewed towards higher values,
- probability distribution of field goals percent (FG_percent) is symmetrical, and
- probability distribution of three-point percent (TP_percent) resembles an hourglass where there are more lower and higher values but fewer values in the middle of the range.

Model building

After exploring the data, I will build different models for classification and compare its performance and explain which predictors were selected and/or have more importance to the outcome than others. All the models explained in the project below add bias in parameter estimates but may result in lower variance and overall better performance than "unbiased" models such as generalized linear regression.

Before explaining the models, it should be noted that all continuous variables were centered (mean deducted) and scaled (divided by standard deviation) to avoid data leakage. The test dataset was scaled and centered using means and standard deviations derived from the training dataset.

Lasso regression

I will start with Lasso regression. The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator. Lasso regression performs so-called L1 regularization, which adds a penalty equal to the absolute value of the magnitude of parameter estimates (i.e., β coefficients). The penalty element is derived as follows: $\lambda \sum_{j=1}^{p} |\beta_j|$. Here $\lambda > 0$ is a hyperparameter known as the regularization parameter (also known as a tuning parameter). As the target variable is binary, the prediction of its outcome is essentially a classification question, so lasso regression adds a penalty element to the logistics regression cost function. The key in Lasso regression is to find the optimal value of regularization parameter λ that would result in the best-performing model.

I have first fitted the model and checked how the number of parameters decreases with an increase in the value of λ parameter. As the response variable is binary, I used the binomial family in this model and all other models presented in this report. Figure 6 shows that the size of the model is decreasing for a larger value of λ. This is expected since in Lasso regression the coefficients' values can be equal to 0.

Figure 6 - Fitted lasso regression against the log-lambda sequence

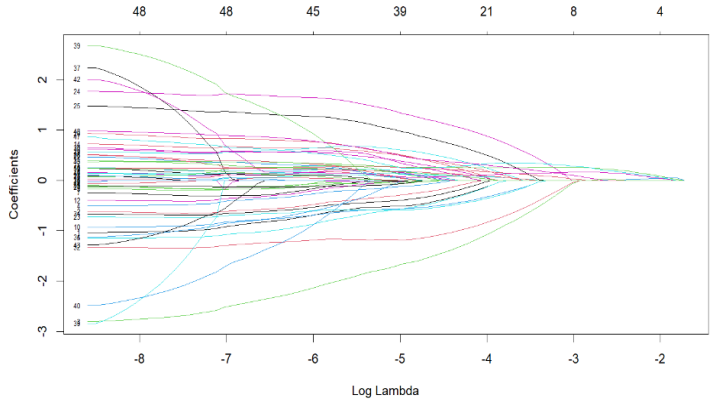
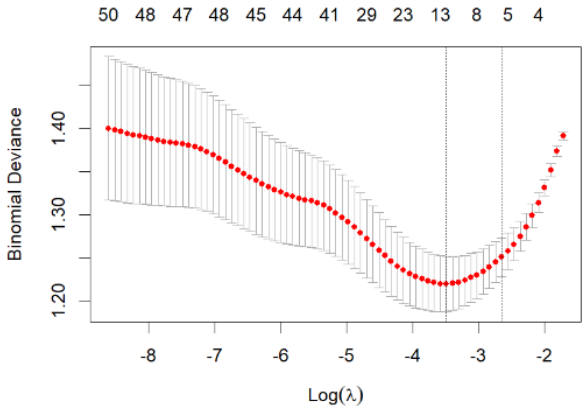


Figure 7 - Lasso regression. Binomial difference against log-λ



This regression gives multiple different λ parameters. In order to pick the model with optimal λ parameter, I performed a k-fold cross-validation and produced a graph of the binomial deviance (or binomial log-likelihood) as a function of the values of the logarithm of λ that is shown in Figure 7.

You can see that on the right-hand side of Figure 7, the binomial deviance is very high, and then at some point, the binomial deviance levels off. This indicates that the full model is doing a good job.

In Figure 7, you can also see two vertical lines. The one is at the minimum binomial deviance corresponding to the log-λ value of -3.487. The other vertical line is within one standard error of the minimum binomial deviance (i.e., the so-called "one-standard-error" rule) corresponding to the log-λ value of -2.650. The second line is a slightly more restricted model (since it refers to a larger value for λ). I will focus on the λ value within one standard error of the one with the minimum binomial deviance.

Table 1 - Lasso regression. Parameter estimates

Variable	Parameter estimate (β) value
GP	0.236
REB	0.189
FT_made	0.158
TOV	0.031
DREB	0.018

The cost function of Lasso regression is minimized only when some β coefficients are set to zero. Accordingly, Table 1 shows the following β remained in the model and considered the most important to decide whether the rookie has a career of more than 5 seasons or not. The most important predictors are GP, REB, and FT_made as they have the largest absolute β values. As all coefficients have positive value, an increase in these variables increases the odds of a rookie having a career that lasts more than 5 seasons.

Ridge regression

Ridge regression performs L2 regularization that adds $\lambda \sum_{j=1}^{p} \beta_j^2$ penalty element to the logistics regression cost function. Unlike lasso regression, in ridge regression, there is no need for the parameter β to be set to 0 for the cost function to be minimized. Variable selection tries to choose a subset of the original variables that is in some way optimal, either in terms of an outcome of interest or some more general measure. Although ridge regression does not perform variable selection, it performs grouped selection. Grouped selection refers to targeting groups of variables that are associated with each other (and deciding whether they should be included or not in the model). Ridge regression automatically includes whole groups into the model if one variable amongst them is selected.

Figure 9 - Fitted ridge regression against the log-lambda sequence

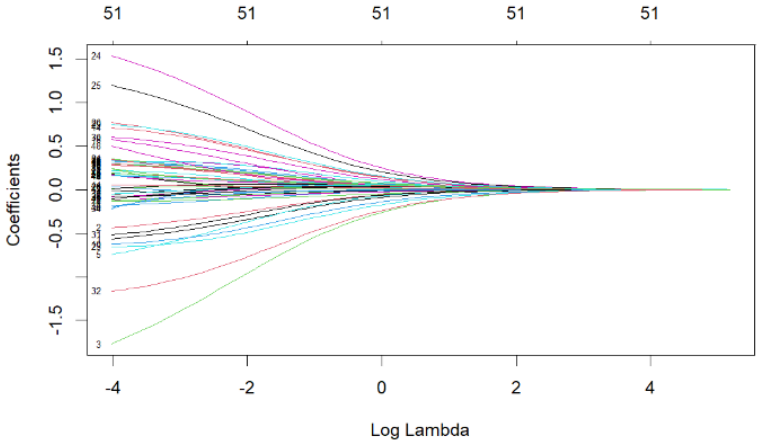


Figure 8 - Ridge regression. Binomial difference against log-λ

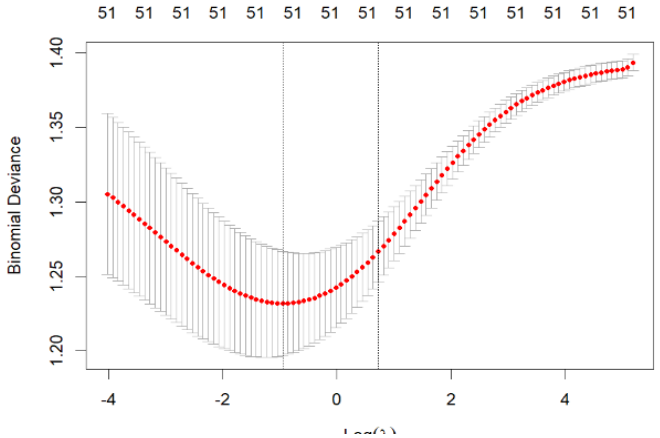


Table 2 - Ridge regression. Parameter estimates for continuous variables

Variable	Parameter estimate value (β)
GP	0.049
REB	0.042
DREB	0.042
FT_made	0.039
OREB	0.039
FTA	0.038
MIN	0.038
TOV	0.038
FG_percent	0.035
PTS	0.035
AST	0.033
FG_made	0.033
BLK	0.030
FGA	0.029
STL	0.028
FT_percent	0.013
TP_percent	0.012
TP_made	0.007
TPA	0.002

As for lasso regression, I first fit the model with different λ parameters and check how the parameter estimates change with the increase in λ. Figure 9 shows that with the increase in log-λ value, parameters estimates values approach zero. But unlike lasso regression, these parameters are not equal to zero. The same approach as for lasso regression is used in estimating optimal λ parameter. Figure 8 shows that log-λ that minimizes the binomial deviance is -0.952, and the log- λ within one standard error from it, which I will use in building the final ridge model, is equal to 0.722. β coefficients for continuous variables are presented in *Table 2*.

Same as for the Lasso model, all continuous variables have a positive coefficient, which means that an increase in the value of these variables will increase the odds of having a career of more than 5 seasons. The most important variables in the model are Games Played (GP), all rebounds related variables (OREB, DREB, REB), and free throws attempted and made (FT_made, FTA), which is close to the results of lasso regression. However, many other important variables have coefficients slightly lower than the ones mentioned above. The least important variables are all three points related variables (TPA, TP_made, and TP_percent) and FT_percent as their β coefficient is much lower than for other variables.

Lasso regression penalized categorical variable "Year_drafted" to zero. This is not the case for ridge regression that estimated separate coefficients for each year.

2003 and 2004 years have the highest positive impact on the Target variable as they have the largest coefficients of 0.134 and 0.114, respectively. Whereas 1982, 2011, and 2002 have the largest negative effect on the Target variable, their coefficients are -0.141, -0.132, and -0.102, respectively. Therefore being drafted in 1982 decreases the odds of getting the career more than 5 seasons the most compared to being drafted in other years.

Elastic nets

Elastic nets combine the strengths of lasso and ridge regression. The elastic net penalty can be seen as a compromise between lasso and ridge regression and incorporates penalties from both $L1$ and $L2$ regularization. The penalty function for elastic nets is the following: $\lambda(\alpha \sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2)$. Specifically, elastic nets use the sum of $L1$ and squared $L2$ norms as the penalty function. Elastic nets introduce another tuning parameter α that can take values from 0 to 1. The value of α equal to 0 corresponds to ridge regression, the value of α equal to 1 corresponds to lasso regression, and the values in between correspond to a combination of both.

I estimated the optimal tuning parameters α and λ using the caret package in R. Caret package runs multiple models with different tuning parameters values and derives the values that maximize the selected goodness-of-fit metric. As the response variable is binary, the appropriate goodness-of-fit metric supported by caret package is accuracy rate and area under the ROC curve (AUC). Please note that the description of both metrics is intentionally omitted in this section as both metrics are discussed in the next section, where I compare the performance of 3 models. I used the AUC metric for optimal tuning parameter values estimation and got α of 0.4 and λ of approximately 0.067.

Table 4 - Elastic nets. Parameter estimates for continuous variables

Variable	Parameter estimate value (β)
GP	0,226
AST	0,173
REB	0,122
DREB	0,117
FT_made	0,114
FG_percent	0,113
OREB	0,075
TOV	0,050
BLK	0,041
TP_percent	0,033
FTA	0,032

Table 3 - Elastic nets. Parameter estimates for categorical variable "Year_drafted"

Variable	Parameter estimate value (β)
Year_drafted1982	- 0,610
Year_drafted2003	0,521
Year_drafted2011	-0,486
Year_drafted2004	0,189
Year_drafted2002	-0,158
Year_drafted1989	-0,133
Year_drafted2008	0,037
Year_drafted1993	0,019

Non-zero parameter estimates of continuous predictor variables are presented in Table 4 and for each value of a categorical variable "Year_drafted" (also non-zero) in Table 3. As for lasso and ridge, game points, rebounds, and free throws made are among the top most important predictors. Also, as in ridge regression, being drafted in 1982, 2011, or 2002 has the most negative impact on the odds of making a career more than 5 years, while being drafted in 2003 or 2004 has the largest positive effect on such odds. At the same time, variable AST is the second most important continuous predictor variable in the elastic nets model, whereas it is not presented in lasso regression and has far lower importance in ridge regression.

Model comparison and conclusion

After building the models, it is time to compare the performance of each model based on the test data.

One way to assess the predictive power of a model is to look at the receiver operating characteristic (ROC) curve of each model. ROC curve is a measure of classifier performance. Using the proportion of positive data points that are correctly predicted as positive (true positive rate) and the proportion of negative data points that are incorrectly predicted as positive (false positive rate), I generated graphs that show the trade-off between the rate at which the model predicts the response correctly versus predicting it incorrectly. On the horizontal axis of the ROC curve, I have the false positive rate, and on the vertical axis, the true positive rate. The area under the ROC curve, known as AUC, is used as a measure of a diagnostic test's discriminatory power. An AUC value of 0.5 indicates that the predictive model is not different than a random guess. So, the better the model, the close its ROC curve to the top left corner of the chart and closer its AUC to 1. Figure 10 shows the comparison of the ROC curves for different models to help us choose between them.

Figure 10 – Comparison of ROC curves of lasso regression, ridge regression and elastic nets

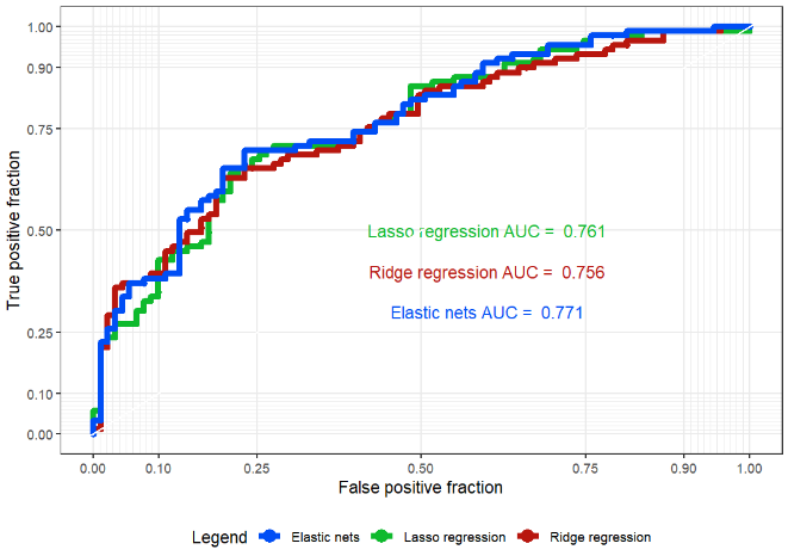


Figure 10 shows that ridge regression has the lowest performance as the curve lies slightly below the curves of other models, and this model has the lowest AUC value. Elastic nets mode has the largest AUC value that indicates a better fit than other models. However, the performance of all classification models is very similar as I can see that the curves lie very close to each other, and the difference in AUC values is small.

Therefore, before concluding on which model to select, I will also look at other performance measurements of classifiers such as accuracy, sensitivity, and specificity.

The correct classification rate is also known as the **accuracy** in the binary setting, is calculated as total cases correctly classified by the total number of cases.

Sensitivity is also known as the true positive rate. This is the proportion of all individuals who were correctly predicted as positive out of the number of true positives. In the case of the dataset in question, the rate will show the number of rookies with a career of more than 5 seasons correctly classified as such divided by the total number of rookies with a career of more than 5 seasons.

Specificity is also known as the true negative rate. This is the proportion of all individuals who were correctly predicted as negative out of the number of true negatives. In the case of the dataset in question, the rate will show the number of rookies with a career length of 5 or less seasons correctly classified as such divided by the total number of rookies with a career length of 5 or less seasons.

Table 5 shows performance measurements calculated for each model. The highest and lowest values are highlighted in green and red, respectively.

Table 5 - Accuracy, sensitivity and specificity values

	Lasso regression	Ridge regression	Elastic nets
Accuracy	0.711	0.689	0.700
Sensitivity	0.708	0.674	0.708
Specificity	0.714	0.703	0.692

As you can see, all three metrics are the highest for Lasso regression. Ridge regression performs worse in accuracy and sensitivity. Elastic nets, although having the highest sensitivity rate, have the lowest specificity. The important consideration in the model selection is the model complexity. Out of the three models discussed in the project, the lasso model is the least complex one as it includes only five predictor variables. Considering the simplicity of the lasso model together with the higher accuracy, specificity, and sensitivity and taking into account that the AUC of lasso regression is only marginally lower than that of elastic nets (by 0.01), lasso regression is considered the best model for the prediction of whether rookies have a career more than 5 years or not.

Table 6 – AUC, accuracy, sensitivity, and specificity of lasso regression model on test and training data

	Training data	Test data
AUC	0.735	0.761
Accuracy	0.662	0.711
Sensitivity	0.597	0.708
Specificity	0.727	0.714

And finally, in order to check the model for overfit, I will compare how the AUC, accuracy, sensitivity, and specificity of the model run on the test data changed compared with the model run on the training data. A decrease in the above-mentioned metrics in test data may indicate that the model has 'learned too much' form training data, including the noise in the training data, to the extent that it negatively impacts the model's performance on new data, i.e., test data. Table 6 shows no decrease in these metrics, implying that the model is not overfit. However, I see the opposite case – a significant increase in sensitivity rate that is also mainly responsible for an increase in AUC value and accuracy rate. This is also not a good indicator, and such an increase may arise from the fact that the model was built (420 observations) and tested (180 observations) on a relatively small sample. Small sample makes the models more sensitive to which observations go into training and test sets. The models would have performed better had they been trained on the dataset with more observations.