INTRODUCTION

The models described in this report were created with the aim to analyse data about National Basketball Association (NBA), with the aim to predict whether a basketball player will play more than 5 seasons.

Basketball players during their first year into NBA are known as rookies. College basketball and NBA are completely different and during this period of time, players try to prove themselves to their opponents and to their teammates. The first contract of an NBA player can last up to 3 years and the second one can last up to 4 years. For this reason, on average, players sign two contracts with one or more teams during their career which, on average, can last about 5 years. Therefore, it would be interesting to predict if an NBA player's career will be longer than 5 years or not.

The data set is composed of 22 variables reporting statistics of a sample of 600 players during their rookie year. *Target* has two values 0 for players' career less than 5 years and 1 for players' career longer than 5 years, and it represents the response variable. The explanatory variables are related to the performance of each player during different games, form the number of points made per game, to the number of steals made. For a thorough description of all the variables in the dataset refer to table 1.

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	1 if Yrs>5 and 0 otherwise

Table 1 *Variables of the Dataset.*

After a brief exploratory analysis which provides information about the general features of the dataset, there will be an analysis to predict if an NBA players' career will last less or more than 5 years. Different models have been created using different kind of statistical and machine learning techniques, from logistic regression with automated selected features, to regularised regression models using Ridge, LASSO and

Elastic Net techniques. Finally, there will be discussion and conclusion about the best model predicting the length of NBA players' career.

EXPLORATORY ANALYSIS

Table 2 reports descriptive statistics of the variables. It is possible to notice for example that they all are positive and that the ranges are different. There are not missing values.

Variable	*Min*	*1st Quartile*	*Median*	*Mean*	*3rd Quartile*	*Max*
Year_drafted	1980	1988	1994	1995	2004	2011
GP	12.00	46.75	62.00	59.96	76.00	82.00
MIN	3.10	10.10	14.85	16.37	20.80	40.90
PTS	0.700	3.375	5.000	6.233	8.100	22.900
FG_made	0.300	1.300	2.000	2.419	3.100	9.000
FGA	0.800	3.100	4.500	5.447	6.825	19.700
FG_percent	23.80	39.70	43.80	43.87	47.52	66.20
TP_made	0.0000	0.0000	0.1000	0.2195	0.3000	2.1000
TPA	0.0000	0.0000	0.2500	0.6985	0.9000	6.5000
TP_percent	0.00	0.00	21.40	19.02	31.95	100.00
FT_made	0.000	0.600	0.900	1.179	1.500	5.400
FTA	0.000	0.800	1.400	1.661	2.100	8.500
FT_percent	0.00	64.38	70.70	69.84	77.20	94.10
OREB	0.0000	0.4000	0.7000	0.9303	1.3000	4.2000
DREB	0.200	0.900	1.500	1.808	2.400	8.800
REB	0.300	1.300	2.100	2.737	3.700	12.300
AST	0.000	0.500	1.000	1.432	1.900	9.300
STL	0.000	0.3000	0.5000	0.5798	0.8000	2.500
BLK	0.000	0.100	0.200	0.335	0.400	2.600
TOV	0.100	0.675	0.900	1.112	1.400	4.200
Yrs	1.000	3.000	5.500	6.702	10.000	19.000
Target	0.0	0.0	0.5	0.5	1.0	1.0

Table 2 *Descriptive statistics of all variables.*

Figure 1 shows the correlation matrix between the numeric variables. It is possible to notice that some of them have very high correlation and this can cause problems with multicollinearity. Multicollinearity is when two or more explanatory variables, in a multiple regression model, are highly linearly correlated. This can produce less reliable results from statistical inference.

Figure 2 shows the box plots of all variables vs the response variable *Target*. Some of them such as *GP, MIN, PTS, FG_made, FGA, FT_made, FTA, OREB, DREB, REB, STL, TOV, and Yrs* seem to have more influence in determining the length of the players' career than others. For these variables, the median (represented by the line in the middle of the boxes) increases going from 0 to 1 of the response variable. That means that higher values of these variables tend to be related to the value 1 of *Target* and therefore to a longer career of the NBA players.

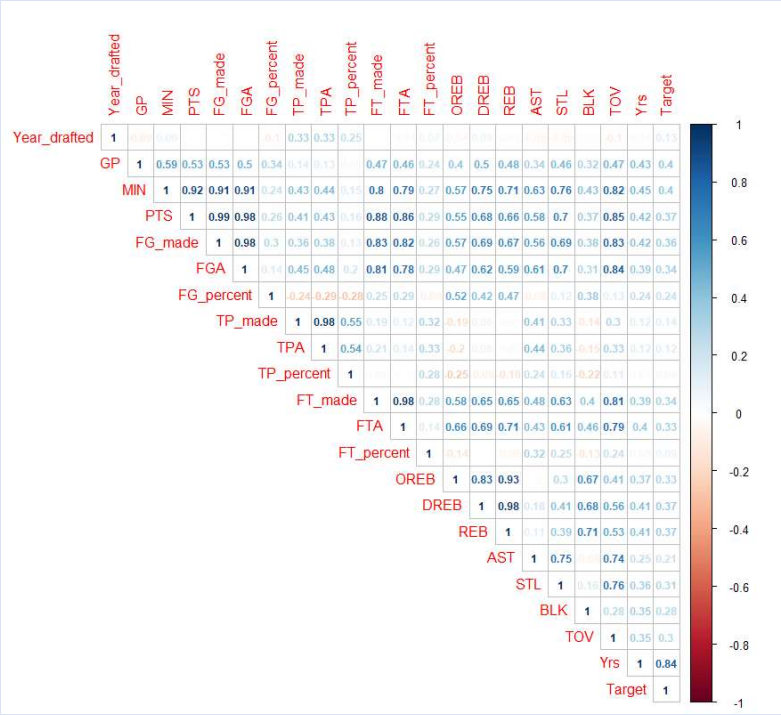

Figure 1 *Correlation matrix.*

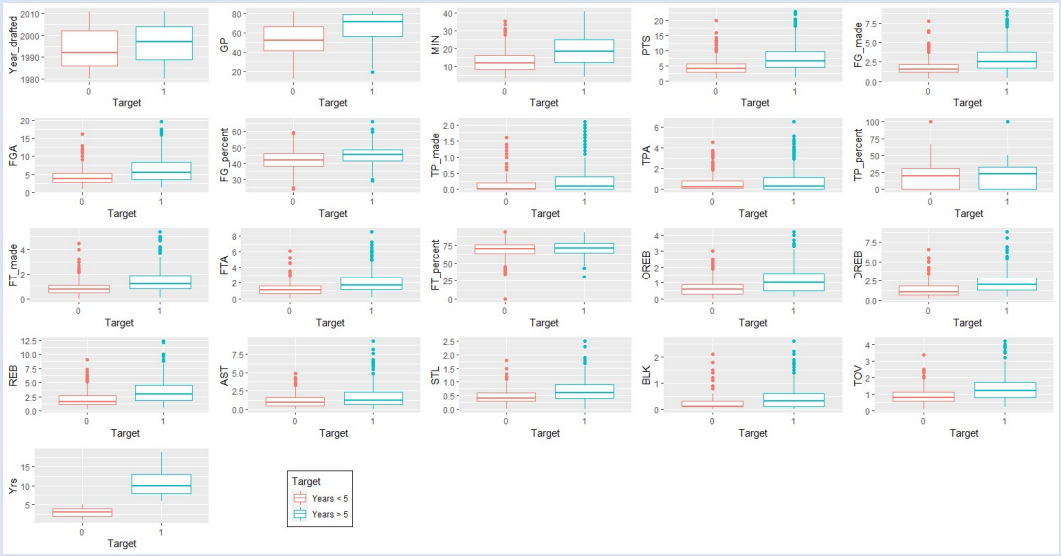

Figure 2 *Box plots.*

Same conclusions can be deducted by Figure 3 and 4. It is possible to notice how for the same above-mentioned variables, the distributions of the observations expressed as histograms for the first and points for the second, are slightly higher for the blue (value 1 for *Target*) rather than the pink (value 0 for *Target*).

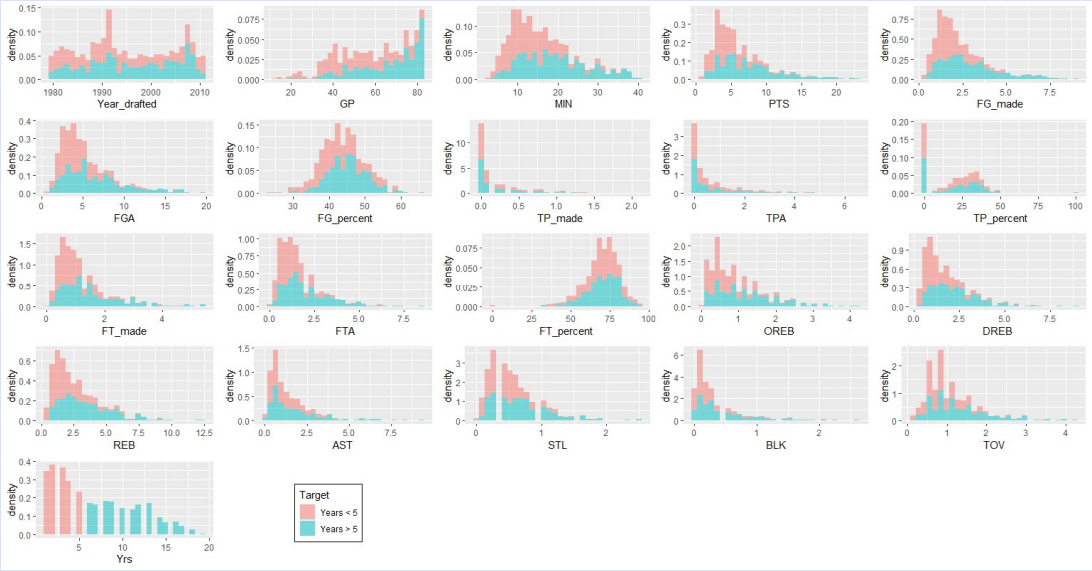

Figure 3 *Histograms.*

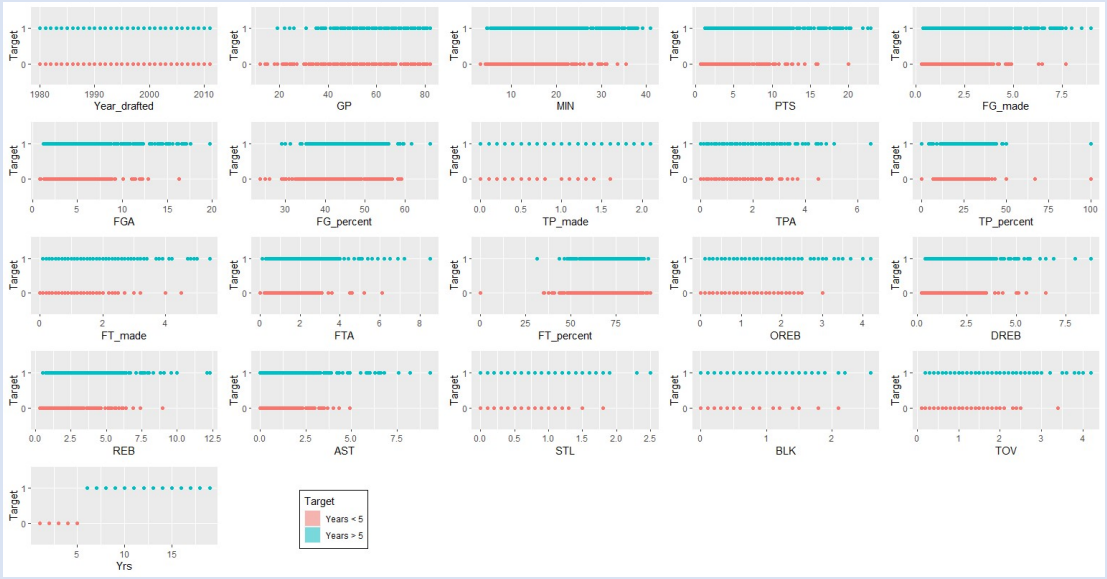

Figure 3 *Points.*

ANALYSIS

To perform the analysis the dataset has been divided into train, validation and test set. The first set has been used to make exploratory analysis illustrated in the previous section, and to train the models. The second set has been used to validate the models making predictions, and evaluate their performances in terms of accuracy (the rate of correct classified observations) and errors (Root Mean Squared Error or RMSE which is the standard deviation of the difference between observed values and predicted). In this way it is possible to choose the best model that has the highest accuracy and the lowest error. Ultimately, the third set will be used to test the final chosen model and thus determine how well it will predict the length of an NBA player's career.

The first model was created using logistic regression, a technique used to model the probability of a certain class or event to exist. All the variables in the dataset have been considered, however, the results obtained are odd, presenting behaviour not acceptable. This is mainly due to the presence of the variable *Yrs*. The model is primarily led by this variable because *Yrs* represents the number of years an NBA player's career will last and it is, for obvious reasons, strongly correlated to the response variable *Target*. Furthermore, if the intention of the analysis is to predict how long an NBA player's career will last, this information will not be available for new players. For all this reasons the variable *Yrs* will be dropped from the model, and the second model created using again logistic regression technique presents much better results. Therefore, predictions have been made using the validation set, and accuracy and RMSE have been calculated and they are reported in Table 3 along with the value of R-squared. R-squared is a value that represents how well the variables in the model explain the variance of the response. In this case the model explains only the 24% of the observations.

Accuracy	*RMSE*	*R-squared*
0.04	1.65	0.24

Table 3 *Values of accuracy, RMSE and R-squared for the second model.*

The third and fourth models created are other two logistic regressions, but this time two different techniques that perform automated features selection have been applied. These two techniques select only those variables that better explain *Target*. Accuracy and RMSE have been calculated after carrying out predictions and they are reported in Table 4. It is clear that these two models perform worse than the previous one as accuracy decreased from 0.04 to 0.02 for model 3 and to 0.03 for model 4. RMSE decreased for model 3 but this model explains less variance of the response than model 2, as R-squared passes from 0.24 to 0.22. On the other hand, model 4 explain slightly more variance but the error is higher.

	Accuracy	*RMSE*	*R-squared*
Model 3	0.02	1.50	0.22
Model 4	0.03	2.04	0.25

Table 4 *Values of accuracy, RMSE and R-squared for the third and fourth models.*

The next models have been created using regularised techniques: Ridge Regression, LASSO Regression and Elastic Net Regression. Before that, the dataset has been modified because different range of the variables can create problems with these techniques. Predictors have therefore been standardised by subtracting the mean and dividing by the standard deviation. These techniques can easily deal with multicollinearity and they introduce penalty for those models that are overly complex. The penalty introduced by Ridge regression works in a way that some predictors could be removed. LASSO regression instead, maintain all the predictors but some of them will be forced to be close to 0. The strength of the penalty is controlled by a parameter λ. The value for λ to use is determined iteratively checking for which value of λ the best model is created.

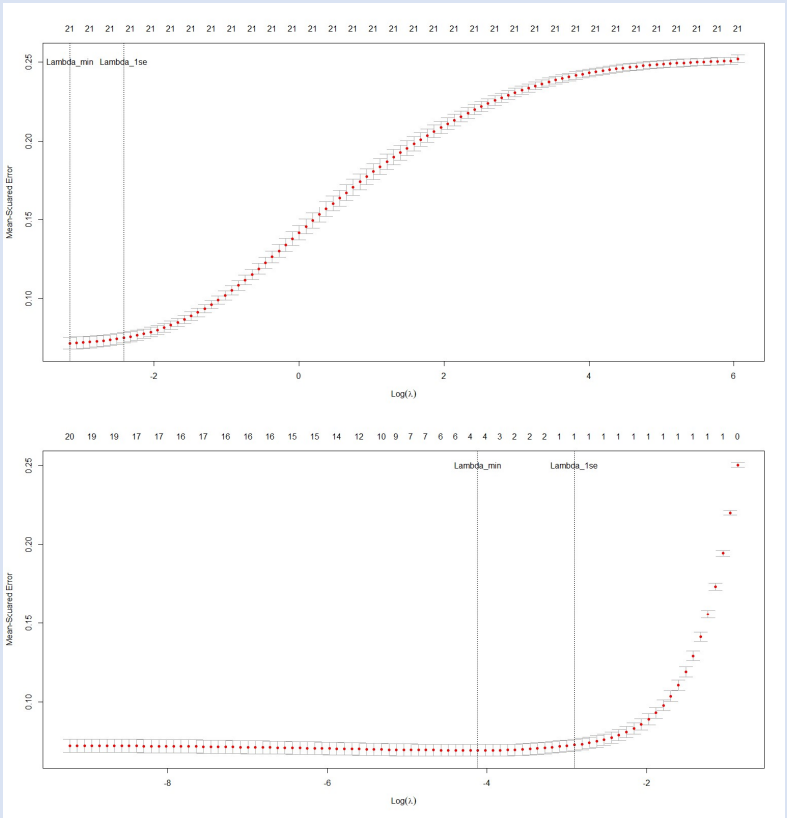

Figure 4 *TOP: best λ and 1st standard error λ for Ridge Regression model. BOTTOM: best λ and 1st standard error λ for LASSO Regression model.*

The top plot of Figure 4 refers to Ridge Regression and it is possible to observe that for values of log(λ) lower than -2, the mean squared error is minimized. Same thing in the bottom plot for LASSO regression. However, for the models created with these two techniques, the value for λ considered is the one indicated by the line Lambda_1se in the plots. This is the value of λ within one standard error of the minimum MSE, which allows slightly more restricted model and it performs almost as well as the minimum. Using the validation set, predictions have been made and the accuracy and RMSE values have been calculated and they are showed in Table 5.

	Accuracy	*RMSE*	*R-squared*
Ridge	0.95	0.22	0.82
LASSO	0.95	0.22	0.82

Table 5 *Values of accuracy, RMSE and R-squared for Ridge and LASSO models.*

Elastic Net Regression is a technique that incorporates the advantages of the penalties of the Ridge and LASSO regression combined. In this technique not only for λ has to be determined its best value, but also for α. α can be considered a parameter that determines how much Ridge Regression or LASSO regression there will be in the model.

Accuracy	*RMSE*	*R-squared*
1.00	0.00	1.00

Table 6 *Values of accuracy, RMSE and R-squared for the Elastic Net model.*

The values reported in Table 6 are too perfect and this is probably due to the presence or the variable *Yrs.* Therefore, these last three techniques are applied once again but this time without taking into consideration *Yrs*. The results for these three models are reported in Table 7.

	Accuracy	*RMSE*	*R-squared*
Ridge	0.72	0.53	0.19
LASSO	0.73	0.52	0.21
Elastic Net	0.75	1.11	0.25

Table 7 *Values of accuracy, RMSE and R-squared for the Ridge, LASSO and Elastic Net models.*

Comparing the three models it is clear that the best performing is Elastic Net model. Figure 5 shows the value of AUC (Area Under the Curve) for the 6 models created. In these plots a diagonal line would indicate that a model is correctly classifying 50% of the observations. Lines above this diagonal line indicate that the model preforms better than the randomly guessing. Figure 5 shows that all the models created for this analysis, have a quite high performance, above 75%. However, the first 3 models which have the variable *Yrs* in the dataset, have a too perfect behaviour. Instead, the last 3 have more reliable results.

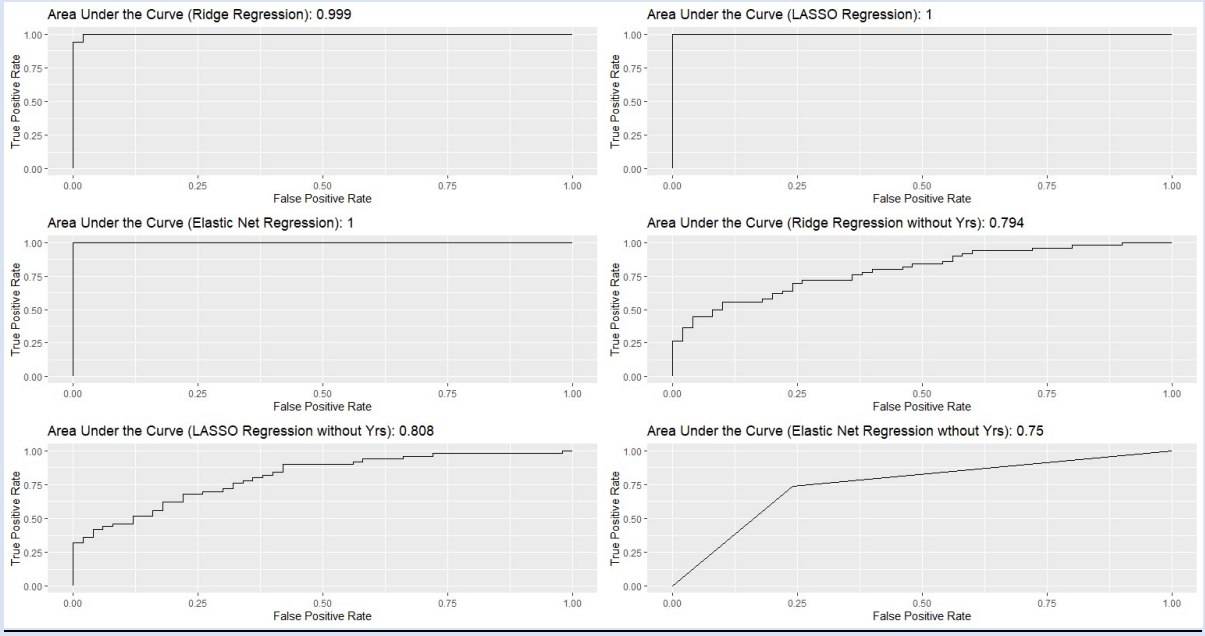

Figure 5 *Values of Area Under the Curve for all 6 models above described.*

CONCLUSIONS

The best model performing in predicting if an NBA player's career will last more or less than 5 years, is Elastic Net. The most meaningful variables for this model, are: *Year_drafted, GP, FGA, FG_percent, TP_made, TP_percent, FT_made, FTA, FT_percent, OREB, DREB, REB, AST, STL, BLK,* and, *TOV.* Other models with higher accuracy and AUC will net be taken into consideration because almost certainly they have problems of overfitting (models that perform well when trained but purely when predicting using new observations, like those described in page 6). Therefore, this model has been applied to make predictions using the test set, and in this way, evaluate its performance in terms of accuracy, RMSE, R-squared plus two other quantities: sensitivity and specificity. They are all reported in Table 8.

Accuracy	*RMSE*	*R-squared*	*Sensitivity*	*Specificity*
0.75	1.07	0.24	0.82	0.67

Table 8 *Values of accuracy, RMSE, R-squared, Sensitivity (correct classification rate of Target class 1) and Specificity (correct classification rate of Target class 0) for the final chosen Elastic Net models.*

The model correctly classifies 75% of observations. However, it has a higher misclassification error for players with length of career less than 5 years, as indicated by the value of specificity. Sensitivity indicates that among all of the players with career longer than 5 years, 82% are correctly classified and 18% are misclassified. Specificity indicates that among all of the players with career less than 5 years, 67% are correctly classified and 33% are misclassified.

Figure 6 shows that the AUC for the final model is still around 75%. There are no problems with overfitting and multicollinearity as expected even though the whole dataset is composed of very highly linearly correlated variables.

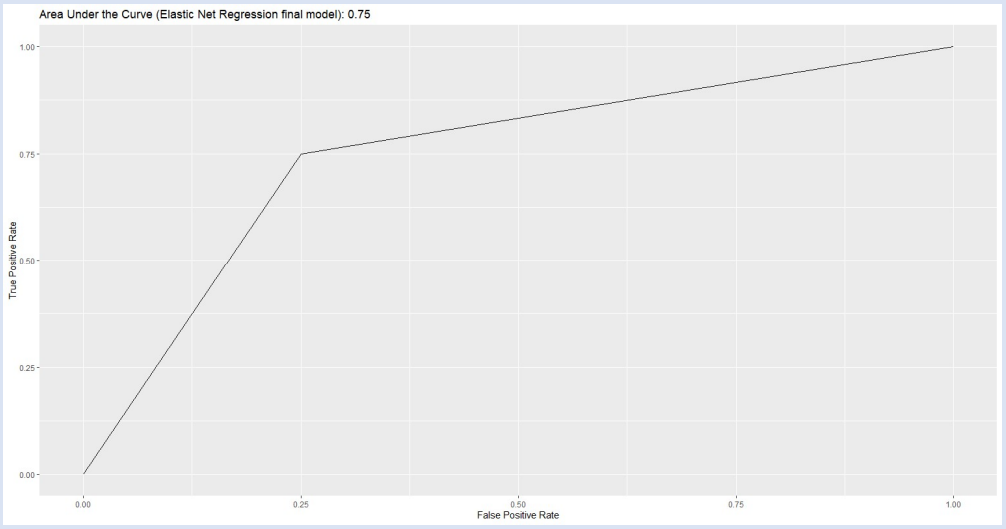

Figure 6 *Values of Area Under the Curve for the final model.*