Project: NBA player 5-year career longevity prediction with logistic regression

Goal

Implement a basic classification model using logistic regression. The provided dataset contains statistics for new NBA players. The goal is to train and test a logistic regression model to predict whether a player will have a career of at least five years in the NBA.

Let's get some insights of Logistic Regression and its working before coding the prediction model.

Introduction – Understanding of Logistic Regression

In the world of data science and machine learning, classification methods play an important role in solving the real-life problems and data mining applications. One of the popular prediction algorithms is Logistic Regression which is used to solve binary classification issues with categorical or qualitative response. The fundamental concept relates to the relationship between 1 dependent variable coded as binary (0 or 1, True/False, Yes/No) and independent variables reflecting the probability of the occurrence of an event. This method uses logit function and calculates the probability of the input that belongs to a category i.e. constant output. For example: Whether a student is pass or fail, whether the patient is diabetic or not etc.

Here, with the given dataset, the prediction uses the target variable with 2 values (0 or 1) – and the probability is calculated whether a player will have a career of at least 5 years or not.

Theoretically, the implementation is achieved by calculating the odds of the event (P/1−P), where, P is the probability of event. So, P always lies between 0 and 1 and gives a S-shape curve instead of straight line as observed in linear regression.

Logistic Math Equation –

$$Z_i = ln\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta_1 x_1 + .. + \beta_n x_n$$

And the probability is given as Sigmoid function which is represented as shown in Fig.1 –

$$P_i = E(y = 1|x_i) = \frac{e^z}{1 + e^z} = \frac{e^{\alpha+\beta_i x_i}}{1 + e^{\alpha+\beta_i x_i}}$$

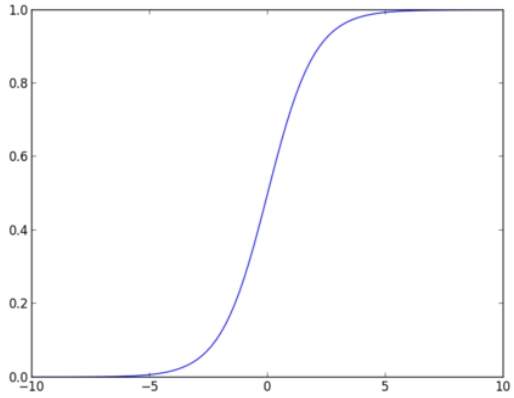

Fig.1 Sigmoid Function – S-shaped Curve

Here, in R this equation uses the glm() function by setting the family argument to "binomial".

To show this, logistic regression uses Maximum likelihood estimation (MLE).

The simplest representation is shown in Fig.2. Here, the inputs are passed in the model, and relationship is determined by the equation which results in the probabilities. These probabilities are transformed into binary values using sigmoid function. After this, the function maps the values between the range of 0 and 1 which is then transformed into either 1 or 0 based on the threshold set.

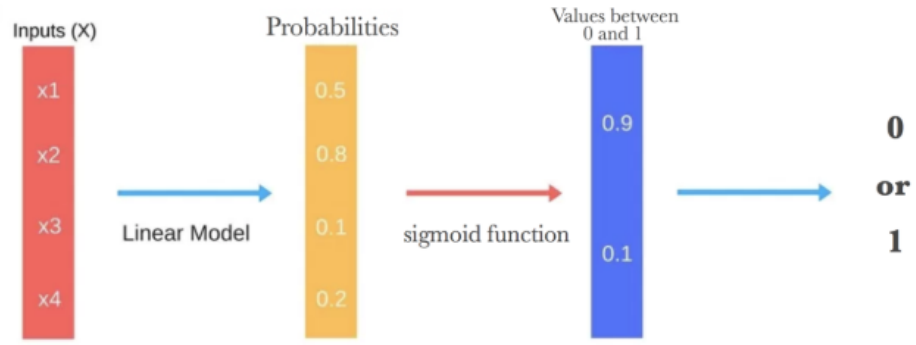

Fig.2 Steps of Logistic Regression

Analysis – Implementation using R

The Dataset – nba_logreg.csv shows the facts of the performance metrics of the NBA players and whether the player will have a career of 5 years or not. To predict the career span of the players is more than 5 years or not is

performed by applying the data analytical cycle. To model and predict the outcome, following steps are performed:

1. Load the dataset and check the structure of the dataset – It has 1340 observations and 21 variables. The response variable is "TARGET_5Yrs" with 2 values (0 and 1).
2. After loading the data, basic data exploration is performed using SmartEDA package in R.

   Result shows:

```
> ExpData(data=data_nba, type=1)
                                     Descriptions         Obs
1                            Sample size (Nrow)        1340
2                      No. of Variables (Ncol)          21
3                     No. of Numeric Variables          20
4                      No. of Factor Variables           1
5                        No. of Text Variables           0
6                     No. of Logical Variables           0
7                        No. of Date Variables           0
8    No. of Zero variance Variables (Uniform)           0
9       %. of Variables having complete cases 95.24% (20)
10 %. of Variables having <50% missing cases   4.76% (1)
11 %. of Variables having >50% missing cases       0% (0)
12 %. of Variables having >90% missing cases       0% (0)
```

3. As observed from the EDA, data has null values so the data cleaning process is performed. It is considered as an important step to check the quality, missing, corrupted data within the dataset and perform necessary actions to provide clean data., In the given dataset, 4.76% missing data is replaced by the mean of the values of the variable.
4. Once the EDA is completed, I will check the Target proportion – 0's and 1's count for the response variable (TARGET_5Yrs). The proportion of the events (0 or 1) should be same to create the unbiased results from the model. However, the observations are not equal for the given data.

```
>  table(data_nba$TARGET_5Yrs)

  0   1
509 831
```

Players having career more than 5 years are more as compared to the rest of them. So, for better models, the observations are made equal by creating the samples.

Moreover, I calculated the summary of the given dataset which displays that the mean value of the response variable is greater the 0.5 for the no. of players having at least 5 years of career.

Based on this given information, I have built a model by creating the training and test datasets and developing the model with the training dataset and predicting using the test dataset. The steps are:

The Model

5. Creating Training and test samples - (80% of the rows as training set and 20% as test set).

   The dataset is divided in such a way that both the training and test dataset have equal 0's and 1's of the response variable for the unbiased predictions.

   It can be showed as - $p \geq 0.5$, class=1 and $p < 0.5$, class=0.

   The result shows Development (Training) data denoted as nba_train has 1328 observations and validation (Test) data denoted as nba_test has 508 observations.

```
> #Count of Train anad Test Dataset
> nrow(nba_train)
[1] 1328
> nrow(nba_test)
[1] 508
> nrow(data_nba)
[1] 1340
```

6. Once the training and test dataset are ready, the process of fitting the model starts on training dataset using glm() function using family as binomial. I used logistic regression considering the required 16 features and the response variable with outcome: 1(positive label) if career length >= 5 years or 0(negative label) if career length < 5 years. Then, I calculated the summary using summary () function which provides the significant variables with the p-values on each of the coefficients. It also mentions the null deviance and residual deviance which can help to determine if the model is better or not.

   With logistic regression, calculating summary is not enough, one must perform odds ratio for the variables to predict right and what factors contributes to the success and failure rate.

The odds ratio output is shown as:

```
> exp(coef(nba_model))
(Intercept)          GP         MIN         PTS         FGM         FGA    X3P.Made        X3PA         FTM         FTA
 0.06212838  1.04309027  0.89307653  1.92467923  0.33213473  1.05899941  5.88253152  0.45303797  1.95063055  0.46377782
       OREB        DREB         REB         AST         STL         BLK         TOV
 0.77658633  0.49403992  2.40212267  1.60054401  1.10065860  1.90558681  0.51397170
```

The odds ratio greater than 1 increases the occurrence of the event. This means, **X3P. Made (3 points made), REB (Rebounds), BLK (Blocks), PTS (Points per Game), FTM (Free Throw Made), AST (Assists), FGA, GP (Games Played)** are the factors that contribute to the longevity of the career of the players.

For ex: the odds of players having career more than 5 years is 5.88 times more if a 3 point is made than a player who has not made.

The factors that seem to determine a short career are **X3PA (3 points attempt), FGM, OREB (Offensive Rebounds), DREB (Defensive Rebounds), FTA (Free Throw Attempts), TOV (Turn Overs)** as the odds ratio is very less.

```
> nba_model <- glm(TARGET_5Yrs ~ GP + MIN + PTS + FGM + FGA + X3P.Made +        X3PA    + FTM + FTA +   OREB +  DREB +  REB
 +      AST     +STL +  BLK +   TOV, family=binomial(link=logit), data = nba_train)
>
> summary(nba_model)

Call:
glm(formula = TARGET_5Yrs ~ GP + MIN + PTS + FGM + FGA + X3P.Made +
    X3PA + FTM + FTA + OREB + DREB + REB + AST + STL + BLK +
    TOV, family = binomial(link = logit), data = nba_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2026  -0.9447  -0.1356   0.9401   2.2383

Coefficients:
             Estimate Std. Error z value             Pr(>|z|)
(Intercept) -2.778552   0.252742 -10.994 < 0.0000000000000002 ***
GP           0.042188   0.004698   8.980 < 0.0000000000000002 ***
MIN         -0.113083   0.033088  -3.418             0.000632 ***
PTS          0.654759   0.847489   0.773             0.439767
FGM         -1.102215   1.700139  -0.648             0.516786
FGA          0.057325   0.146024   0.393             0.694637
X3P.Made     1.771987   1.242710   1.426             0.153896
X3PA        -0.791779   0.388957  -2.036             0.041786 *
FTM          0.668153   0.926374   0.721             0.470752
FTA         -0.768350   0.335137  -2.293             0.021868 *
OREB        -0.252847   1.303281  -0.194             0.846169
DREB        -0.705139   1.303881  -0.541             0.588645
REB          0.876353   1.297034   0.676             0.499257
AST          0.470344   0.109714   4.287            0.0000181 ***
STL          0.095909   0.303633   0.316             0.752101
BLK          0.644790   0.256381   2.515             0.011904 *
TOV         -0.665587   0.272911  -2.439             0.014734 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1841  on 1327  degrees of freedom
Residual deviance: 1518  on 1311  degrees of freedom
AIC: 1552

Number of Fisher Scoring iterations: 4
```

7. After evaluating the fitting of model and computing the odds ratio, I proceeded to see how this model predicts the new data i.e. Test data with the response variable "TARGET_5Yrs". To predict the model, I assigned the parameter "response" to predict () function. Here, the output of the probability will be based on the boundary set i.e. $p \geq 0.5$, class=1 and $p < 0.5$, class=0.

   Now, as the last step, I have assessed the predictive ability of the model by calculating the accuracy, optimal cutoff, concordance index, sensitivity, specificity, ROC curve and Confusion matrix. These all measurements help to identify whether the model was able to predict correctly or not.

8. **ROC and Confusion Matrix** –

   The ROC curve shows the true positive rate against the false positive rate. It's the adjustment between sensitivity and specificity. The AUC score for this classifier is 0.769 which seems to be good fit as it refers to the accuracy index or concordance index. It shows the success rate of the players having career longer than 5 years.

The confusion matrix shows the actual and predicted values which helps to identify the performance of the model. This shows 46+342 correct predictions and 31+88 incorrect predictions.

```
> # Confusion Matrix
> confusionMatrix(nba_test$TARGET_5Yrs, predicted, threshold = optCutOff)
   0   1
0 46  31
1 88 342
```

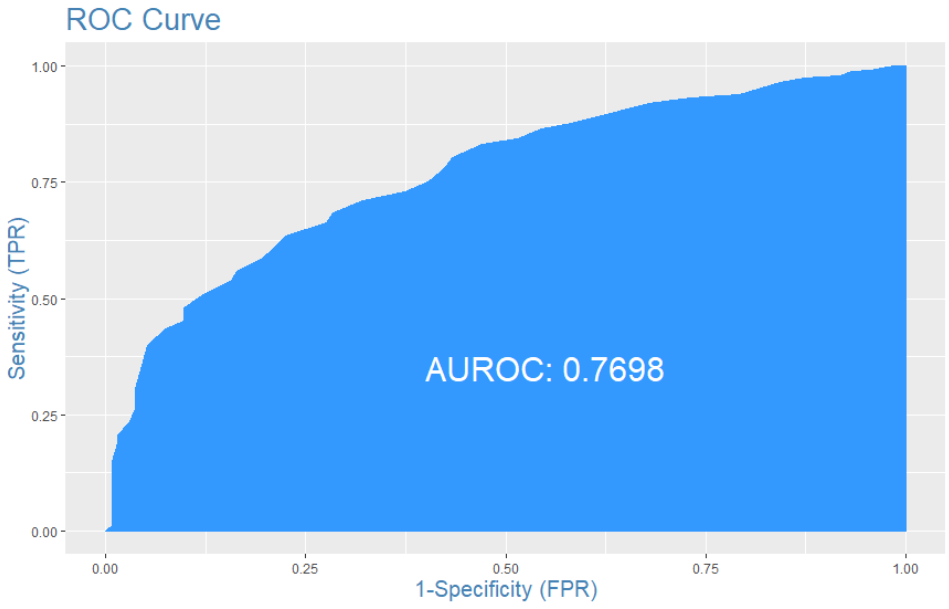

Fig.3 ROC Curve

Conclusion

After creating, analyzing the model and performing the prediction, the results conclude that the model was able to predict the career longevity of the players with **69% accuracy rate**. The model also seems to be good fit because the difference of the null and residual deviance is big enough as greater the difference, better the model. The area under curve plotted using **ROC is 0.769 (77%)** which is close to 1 which results that the model is pretty good. The confusion matrix also shows that the true positives rate is high as compared to the true negatives

For more accurate results, optimization techniques can be used to produce more accurate results and more efficient model. Overall, the model is acceptable for determining whether a player will have a career of at least five years in the NBA.

References

Prabhakaran S. (2016-17) Logistic Regression with R. R-statistics.co. Retrieved from http://r-statistics.co/Logistic-Regression-With-R.html

Fig.2 Donges, V. (2018). The Logistic Regression Algorithm. machinelearning-blog.com. Retrieved from https://machinelearning-blog.com/2018/04/23/logistic-regression-101/