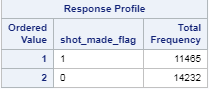
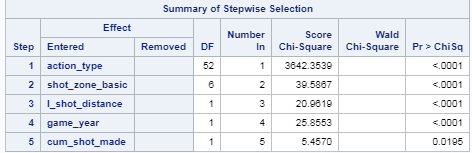
**Predictive Model**

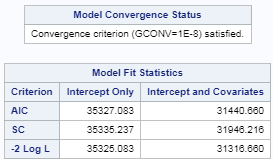
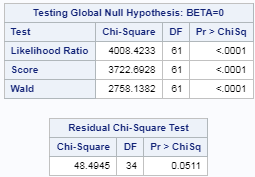
Our final model is a logistic model with probability threshold of 0.47 and above for a successful shot. The dependent variable shot\_made\_flag is binary with 1 for a shot successful and 0 for shot not successful.

Our model uses action\_type, shot\_zone\_basic, l\_shot\_distance, game\_year, cum\_shot\_made to predict shot\_made\_flag. As shown in figure 5.1 they all are statistically significate impact on shot\_made.

**Test of linearity of log(odds) with independent variables:**

We have use the Hosmer and Lemeshow test for lack-of-fit . As we can see in figure 5.3, The test fails for lack-of-fit. We proceed further.

**Model statistics:**

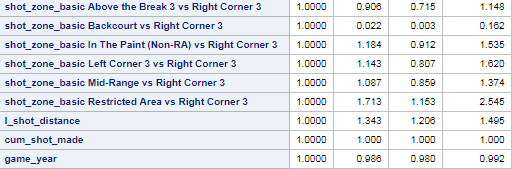
 

The model AIC of 35327 was least of all other models we created. Hence the choice of these combination of variables. The global null hypothesis tests shows that these variables of choice have statistical significance on the model. Model convergence status of statisfied indicates overall it is a good model.

|  |  |
| --- | --- |
|  | ROC curve for sensitivity/c statistics of 0.6929 shows that the model can accuratly predict success or failure at 0.6929 or 69.29 %. We can look at this accuracy when we look at the confusion matrix. |

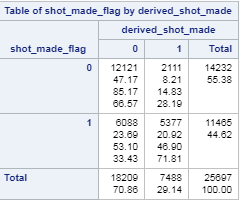
**Odds:**

Figure AMLH in appendix shows the analysis of maximum likelihood estimates. Figure oddsration in appendix shows the odds rations for these estimates. Below figure small subset of the overall table (which is in the appendix).

 Intepreting odds ratios for Kobe. From the above snapshot we can say,

* The odds of Kobe making the shot increases by 1.184 times when “in the paint” zone vs “right corner 3”.
* The odds of Kobe making a shot decreses by 0.906 when taking a shot from “backcourt” vs taking a shot from “right corner”.
* As game\_year increases, the odds of Kobe making a shot decreses by 0.98 times.

**Confusion matrix/ cross validation:**

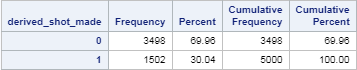


From the matrix we can calculate:

* Accuracy = (12121 + 5377)/ 25697 = 68.1 %
* sensitivity = 5377/11465 = 46.9%
* Specificity =12121 /14232 = 85.2%
* False positives = 2111/7488 = 28.2%
* False negatives 6088/18209 = 33.4%

**Prediction for test data:**

We have used the above model to run the prediction on our test dataset. Below is the summary and Kaggle score.



**Kaggle score**

