Predictive Model

The data for the Kobe Bryant Kaggle completion is a time series having twenty years of historical shot records. It only makes sense to train the model on data recorded prior to the shot we are predicting and because of Kobe’s skills could be changing over time we use a rolling window of one year (365 days). This window is used for a traning dataset before every day we are making a prediction. We move the window every day we are making predictions.

To create the “rolling window” we use first extract all unique dates that have shot\_made\_flag equal to null (dates on which we have to predict) and we store them into a dataset called “dates”. Since the date type in SAS is just a number, we create the rolling window to be in between the current row from “dates” dataset and the same value minus 365. In this case we are training and predicting 1457 times in a loop.

We use logistic regression for making prediction because we are predicting a binary feature: “Shot Made” or “Shot Not Made” represented by 1 or 0. Because of the data in the rolling window is only fraction of the whole dataset, the categorical features will not include all of the categories and we found that using only the continuous features is giving us better score. Specifically, we are using minutes\_remaining, period, seconds\_remaining, and shot\_distance as features for the model.

The output from the logistic regression is in a form of logits: a probability in between 0 and 1 representing how likely is “Shot Made”. Normally, we would transform the output logits to be 0 if probability is less than .5 and 1 if the probability is greater than .5. However, the Kaggle competition requires a submission file with shot IDs and probability, so we don’t make any transformations of the logits. After each train/predict iteration, we collect the output of the logistic regression and finally we export the fields needed to a CSV file. We achieved a Kaggle score of 0.77881.

1. PROC IMPORT DATAFILE='/home/iangelov0/project3/data.csv' replace
2. DBMS=CSV
3. OUT=data;
4. GETNAMES=YES;
6. proc sort data=data out=data;
7. by game\_date;
9. data data;
10. set data;
11. date=input(game\_date,yymmddd10.);
13. proc sql;
14. create table dates as
15. select distinct date
16. from data where data.shot\_made\_flag is null;
18. data result;
19. set data(obs=0);
20. logits = 0;
21. logits\_backup = 0;
23. %MACRO LOOP\_DATA;
24. /\* 1457 \*/
25. %DO i = 1 %TO 1457;
26. DATA current\_date;
27. SET dates(obs=&i firstobs=&i);
29. proc sql;
30. create table rolling\_data as
31. select \* from data a
32. inner join current\_date b
33. on  (a.date **>** b.date-365 and a.date **<** **b.date** and a.shot\_made\_flag is not null)
34. or (a.date = b.date);
36. proc logistic data=rolling\_data noprint;
37. model shot\_made\_flag = minutes\_remaining period seconds\_remaining shot\_distance;
38. output out=predicted\_backup(where=(shot\_made\_flag=.)) p=logits;
40. data predicted;
41. merge predicted predicted\_backup;
43. proc delete data=rolling\_data;
44. proc append base=result data=predicted force;
45. %END;
46. %MEND LOOP\_DATA;
47. %LOOP\_DATA;
49. data result;
50. set result;
51. shot\_made\_flag=logits;
52. keep shot\_id shot\_made\_flag;
54. proc export data=result dbms=csv
55. outfile='/home/iangelov0/project3/submition.csv'
56. replace;
57. run;