**Detail Report of Analytics**

**Dr Milan A Joshi**

Following Steps are taken in training- testing - predicting models.

|  |  |
| --- | --- |
| 1. | Had a well look at data ( All variables except team name are numeric) |
| 2. | Response variable is coded as 0 ,1 ( It's a classification problem) |
| 3. | Converted response variable to factor for analysis purpose. |
| 4. | The variables RS and RA are of very large magnitude as compared to other numeric variable to avoid the dominance of these two variables pre-processing (scaling and cantering) of RS and RA is performed to match up with other variables. Package caret is used. |
| 5. | Checked for Missing Values , No missing Values found, plotted pattern for missing values blue colour indicates that no missing values. VIM package is used. |
| 6. | Bar Graphs shows Proportion of 0 and 1s there are 80.2 zeros and 19.8 ones. package scales and ggplot2 is used. |
| 7. | Done Exploratory analysis on Numeric variables ( Correlation plots, density Plots)  It is found that all the variables are highly correlated but being a classification problem there is no worry most of the classification algorithms are not sensitive to correlations  From the density plot it has observed that BA,RA,RS has more contribution to 1 and OBP,SLG has more contribution to 0, RA is less correlated with other variables.  We use Performance Analytics library to test distribution and Correlations. |
| 8. | Once all Exploratory analysis is done we turn to data splitting using caret library.  We split the dataset in 70-30 train test. Then we convert response variable to factor in train and test data. |
| 9 | We use the 10 fold cross validation in tuning parameters of the model using caret library. |
| 10 | First we train gbm ( gradient boosting) model parameters to tune n.trees  interaction.depth and shrinkage , we settled down at the final values used for the  model were n.trees = 100, interaction.depth = 2 and shrinkage = 0.1**.** after 10 fold cross validation each 3 times. the model was chosen with best parameters and used it to test on test set found the accuracy on test set to be 0.8566, model is predicted on test set |
| 11. | Next we train RandomForest parameters to tune mtry , we perform 10 fold cross validation 3 times , we settled down at mtry = 3 giving accuracy on test set 0.8566  We used best model to predict on test set |
| 12. | Next we run extraTress parameters to tune number of random cuts and mtry we settled down at 2 each and got the accuracy on test set to be 0.8721. |
| 13 | Next we run logistic regression we got the accuracy 0.8837 best out of all models got  residual deviance 290.24 and AIC = 302.24 (minimum) using 10 fold cross validation  the best model is used for prediction on test set as well on unknown dataset |
| 14 | Next we run SVM with Gaussian kernel , parameters to tune and c  After 10 fold cross validation we arrive at best sigma = 0.3252305 and C = 0.5,with accuracy on test set 0.876.  We used it to predict on test set. |
| 15 | Next we compare all model . Plot the accuracy using dot plot , box plot correlation plot. and also plotted the difference between the performance metrics. figures are shown in  Pdf file attached herewith. |
| 16 | Finally we run Simulated Annealing technique to get best variables contributing got the following results RA (100%), SLG (100%), RS (90%), BA (74%), OBP (42%)  % for importance,  and we also plotted important variable for each model separately |
| 17 | Finally we predicted logit model on unknown dataset as logit model gives the best  accuracy on test set. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | k | Sensitivity | Specificity |
| gbm | 0.8566 | 0.5157 | 0.9324 | 0.5490 |
| RandomForest | 0.8566 | 0.53 | 0.9179 | 0.6078 |
| extraTree | 0.8721 | 0.5813 | 0.9324 | 0.6275 |
| logit | 0.8837 | 0.6334 | 0.92 | 0.7059 |
| svm | 0.876 | 0.5908 | 0.93 | 0.6275 |