Logistic Regression VS Random Forest

Predicting home wins from half time stats with the European Football Dataset

The Goal:
Compare and contrast the performance of Random Forest and Linear Regression in a binary classification problem, to predict whether or not a home team will win a football match according to their half time match statistics.

Motivation:
- To find quantifiably significant attributes from half time match statistics which are indicative of a home team winning a soccer match.
- To determine whether machine learning techniques reveal any insight beyond our own intuitions.

The dataset:
- 'Match Statistics from top 5 European Leagues' (2012-2017), www.kaggle.com.
- 9127 observations, each with 92 attributes, pre wrangling.
- Filtered to be fit for purpose; HT stats only & removal of linearly dependant attributes.
- 8273 observations, each with 37 observations post wrangling.
- 36 numerical predictors, including calculated HT goal difference metric.
- One categorical to be predicted; HomeWin, 1 or 0.

Initial intuitions:
- We expect home team half time goal difference to have of great importance in classifying home team wins.
- We expect that certain attributes of the data to be strongly correlated, such as total shots on target and total shots.

Initial analysis of the data set including basic statistics:
- Covariance and Correlation matrices calculated (Fig. 2) where the correlated variables may be appreciated.
- Scatter matrix plot (Fig. 1) to see the data structure and histograms.
- PCA performed (Fig. 4) to see the how the features explain the variance.

Figure 1 - Scatterplot matrix of the dataset.

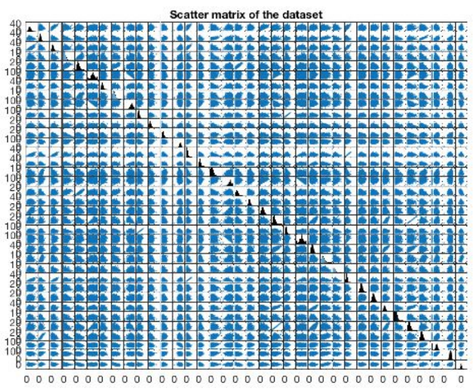
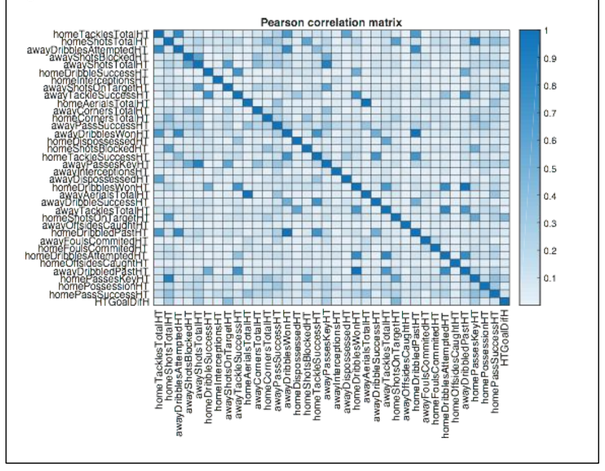


Figure 2 - Pearson's correlation matrix.



Hypothesis Statement:
- Random Forest is expected to perform better than Logistic Regression [2].
- Logistic Regression is expected to be faster than the Random Forest.
- The key feature in both models is expected to be half time goal difference.

Training and Methodology:
Holdout method used with data split as follows:
- Training set: 5791 data points to train the models (70% of the dataset).
- Validation set : 1655 data points to adjust the hyperparameters (20% of the dataset).
- Test set: 827 data points to evaluate the performance of the models (10% of the dataset).
- The considered evaluation criteria is the misclassification rate.

Hyperparameter RF selection:
Optimal number of trees and predictors for RF were obtained through a grid-search.
- Fig. 5 and 6 show the decrease in error with the increase of the number of trees and predictors. With more than approximately 50 trees and 5 predictors, the error begins to plateau.
- Fig. 7 show the increase in training time with the increase of the number of trees and predictors.
- Thus, the final model is trained with 60 trees and 6 predictors: a tradeoff between least error and minimum time.

Hyperparameter LR selection:
No hyperparameters are needed for Logistic Regression. The only decision to make with this model is a choice of link function:
- Link function: Logit function is used in this case (representing the odds between the probabilities of home time winning or not) [3].

Figure 3 - Scatterplot of Range-Normalized Mean vs Range-Normalized Standard Deviation for each Attribute.

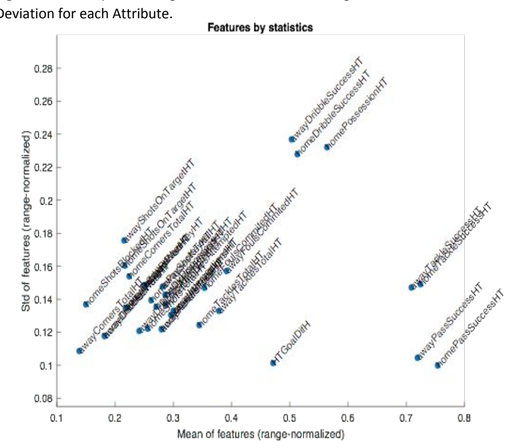


Figure 4 - Principle Component Analysis Projection.

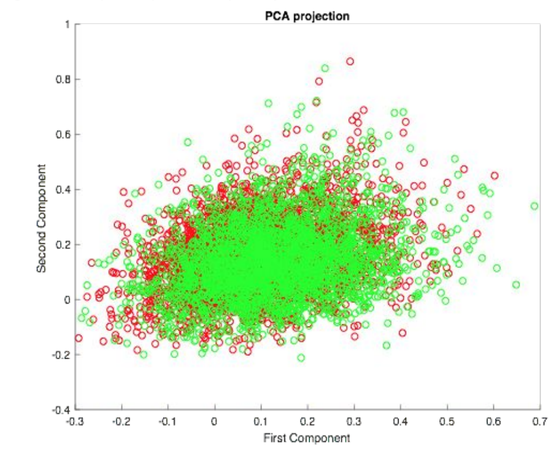


Figure 5 - 3D plot of Trees Sampled vs Predictors Sampled vs Classification Error.

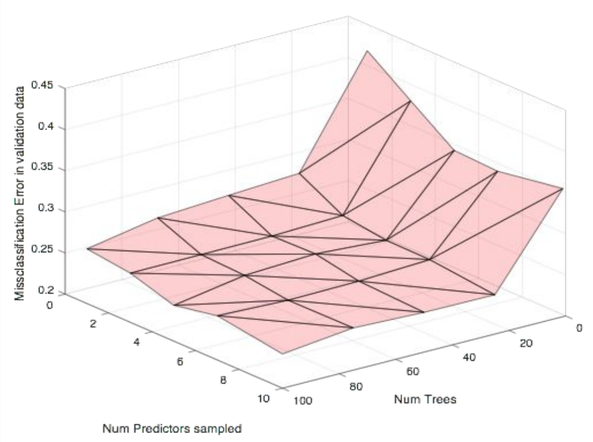
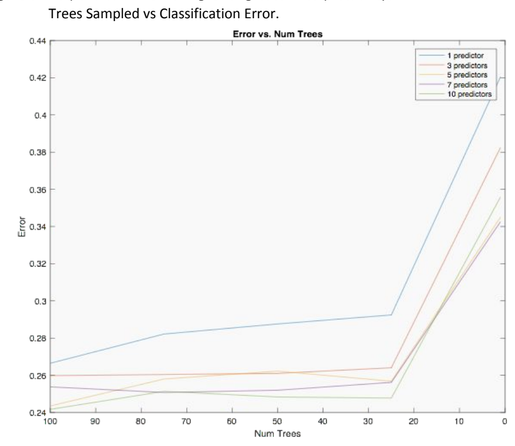


Figure 6 - Perpendicular view of Fig. 5 for greater interpretability; Trees Sampled vs Classification Error.



How do the methods line-up against each other?

Logistic Regression

1. Regression model with a categorical dependent variable.
2. Used in classification problems despite being a regression.
3. Generalized for k classes, k=2 in our case, home win; yes or no.
4. Independent variables are weighted and linearly combined to obtain a score [5]
5. The score is mapped into a probability using a link Function [5]
6. In our case, the logit function is used as the link, where the score represents the natural logarithm of the odds (ratio of both probabilities) [6]

Pros:
- Logistic regression models are fast to train.
- The model offers certain degree of interpretability.
- There are no parameters to tune in the model.

Cons:
- Unable to capture non-linearities in the data.
- Performance is poor when data is noisy.

Random Forest

1. Ensemble method which makes use of decision tree learning.
2. Randomly assigns an attribute to each node [1]
3. Randomly splits the data (usually into two) at each node [1]
4. Each tree uses a randomly selected subset of the training data, with Replacement [1]
5. Classifies each observation in the training data to a category by a majority vote [4]

Pros:
- Outperforms other techniques [2]
- Able to handle non-linearities in the data
- Unlikely to overfit data due to randomness of each tree [1]
- Ranks features according to their importance
- Allows for meaningful analysis of datasets with higher dimension than observations.

Cons:
- Computationally expensive for data sets with a large number of observations

Figure 7 - Evolution of the Training Time with the Increase in Complexity of the Forest.

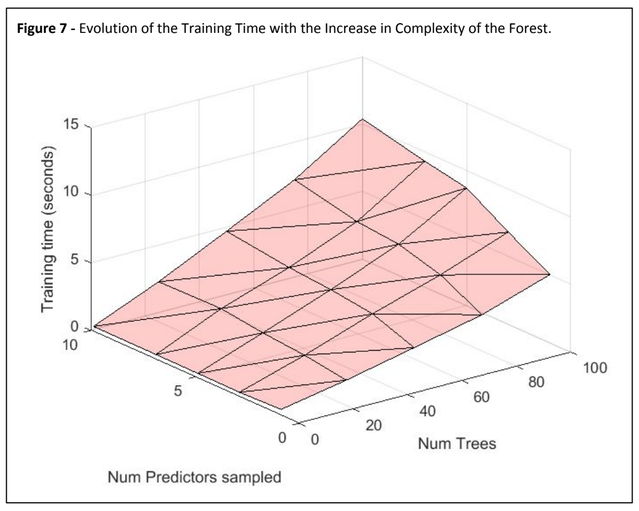
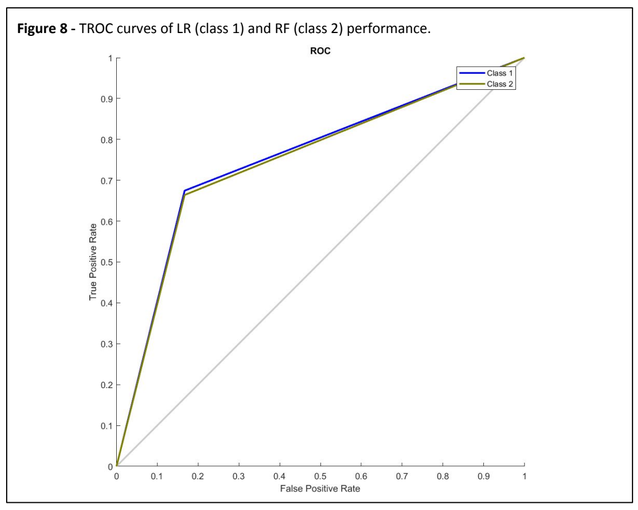


Figure 8 - TROC curves of LR (class 1) and RF (class 2) performance.



Analysis and Critical Evaluation:
Almost the same performance is achieved (misclassification error, precision, recall) using Random Forest and Logistic Regression in this dataset. However, it is interesting to remark that Random Forest takes almost twice as long to train as Logistic Regression. Moreover, this problem increases with increasing complexity of the Forest (ie. number of trees and predictors). This issue becomes crucial if the classification system is applied to a problem on a bigger scale.

Each classifier is basing the decision on the following features:
- Logistic Regression weights POSSESSION and AWAY DRIBBLE SUCCESS five orders of magnitude more than the rest of the features.
- Random Forest is strongly influenced by GOAL DIFFERENCE (4 times more importance than the rest of the features).

As expected, for Random Forest the goal difference is the driving factor to decide whether the home team will win or not. Surprisingly, Logistic Regression is taking into account other features to make its decision. The results of this project clearly demonstrate the advantages of using and comparing different methods when performing data analysis; not only performance matters but obtaining a certain degree of interpretation of the important feature of your models is key as well.

Conclusions which can be extracted from the models:
- Logistic Regression is a high bias/low variance model: the performance on training, validation and test sets is almost the same.
- Random Forest is a powerful low bias/low variance model: it is able to learn by heart the training set while not overfitting.

The Results:

Logistic Regression		Random Forest
23.64%	Training Error	0.00%
25.02%	Validation Error	24.95%
23.94%	Test Error	24.43%
83.30%	Test Recall	83.30%
75.30%	Test Precision	74.70%
1.914 s	Time	4.357 s

Future work and lessons learned:
A good overall conclusion is that although the power of complex techniques such as Random Forest stands out, they do not always outperform simpler techniques such as Logistic Regression. Thus, our two most important lessons learned are:
- Performance always depends on the context, problem and approach.
- Interpretability of models is key: in which way the techniques are managing the data offers a powerful insight in understanding the problem and the solution.

Some possible ways to continue this project are:
- Use other techniques to understand how they manage the data.
- Quantify the impact of other Random Forest hyperparameters such as maximum number of childs in a leaf node or depth of the tree.

Citations:

[1] Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.
[2] Caruana, R. and Niculescu-Mizil, A. (2006) 'An empirical comparison of supervised learning algorithms', *Proceedings of the 23th International Conference on Machine Learning*, pp. 161–168. doi: 10.1145/1143844.1143865.
[3] Crammer, J. S. (2002) 'The origins of Logistic Regression', *Tinbergen Institute discussion paper*.
[4] Liaw, a and Wiener, M. (2002) 'Classification and Regression by randomForest', *R news*, 2(December), pp. 18–22. doi: 10.1177/154405910408300516.
[5] Peng, C.-Y. J., Lee, K. L. and Ingersoll, G. M. (2002) 'An Introduction to Logistic Regression Analysis and Reporting', *The Journal of Educational Research*, 96(1), pp. 3–14. doi: 10.1080/00220670209598786.
[6] Ruiz-Gazeb, A. and Villa, N. (2007) 'Storms prediction: Logistic regression vs random forest for unbalanced data', *Case Studies in Business, Industry, and Government Statistics*, 1(2), pp. 91–101. doi: hal.archives-ouvertes.fr:hal-00270176.

CITY UNIVERSITY OF LONDON EST 1894