Performance comparison of Random Forest and Linear Regression Models

Wassim Ben Youssef and Benjamin Auzanneau

Machine Learning, Data Science, City University

Description of Problem

- Compare the performance of two regression models
- The case study: Predicting the market value of football strikers at their physical sporting peak

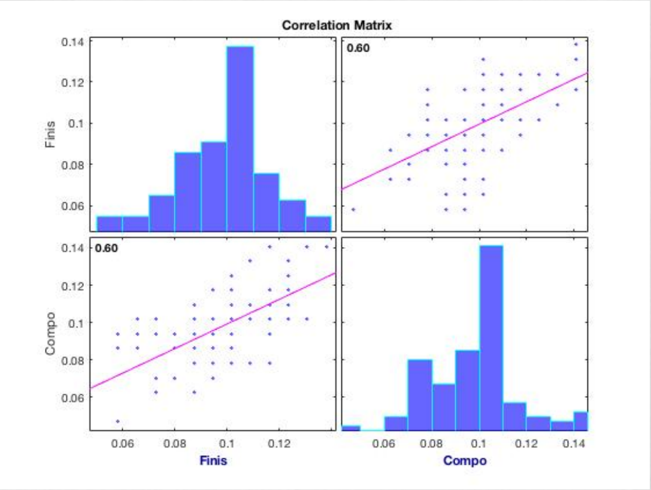
Dataset Description

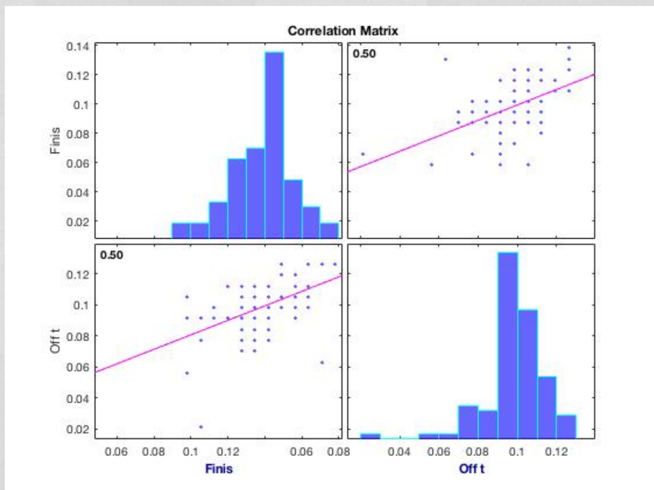
- 100 observations consisting of 7 player attributes, age, original and peak market value
- Player attributes ranging between 0 and 20 are collected from Football Manager 2011, Market values in thousand £ from transfermarkt.com
- Dataset homogenized by normalizing between 0 and 1 to attend to scale differences

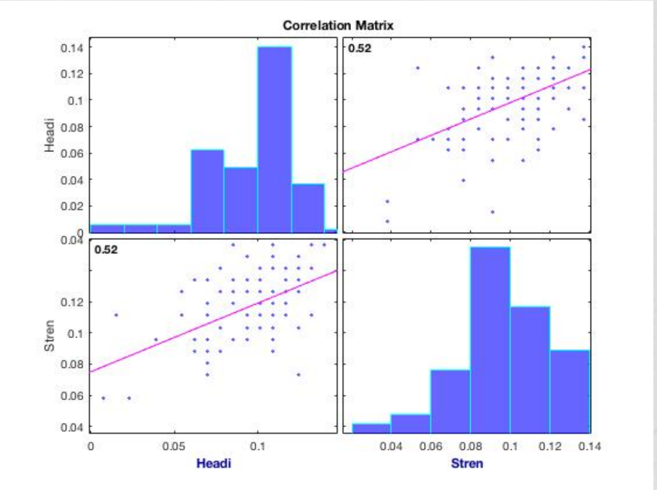
BN = Before norm. AN = After norm.	Initial age	Initial Market Value (K)	Finishing	First Touch	Heading	Composure	Off the ball	Acceleration	Strength	Peak Market value (K)
Mean BN	23.5	7277.68	13.58	13.1	12.43	12.57	14.08	14.96	12.85	10149.44
Mean AN	0.0999	0.0564	0.0987	0.0987	0.0967	0.0984	0.0988	0.0991	0.0977	0.0538
Med. BN	23.5	4375	14	13	13	13	14	15	13	3610
Med. AN	0.0999	0.0339	0.1017	0.0979	0.1012	0.1018	0.0983	0.0994	0.0988	0.0192
Std. BN	1.1237	10702.8526	2.2436	2.1719	3.2730	2.2841	2.1866	1.9844	2.8226	15965.7078
Std. AN	0.0048	0.0830	0.0163	0.0164	0.0255	0.0179	0.0153	0.0132	0.0215	0.0847
Skewness	1.3216E-13	4.6524	-0.3134	-0.3312	-1.0653	0.0982	-1.5358	-0.1315	-0.5208	3.0177
Kurtosis	1.64	30.8952	3.2129	4.4778	4.4150	3.4413	8.8588	2.8650	3.1332	14.0374

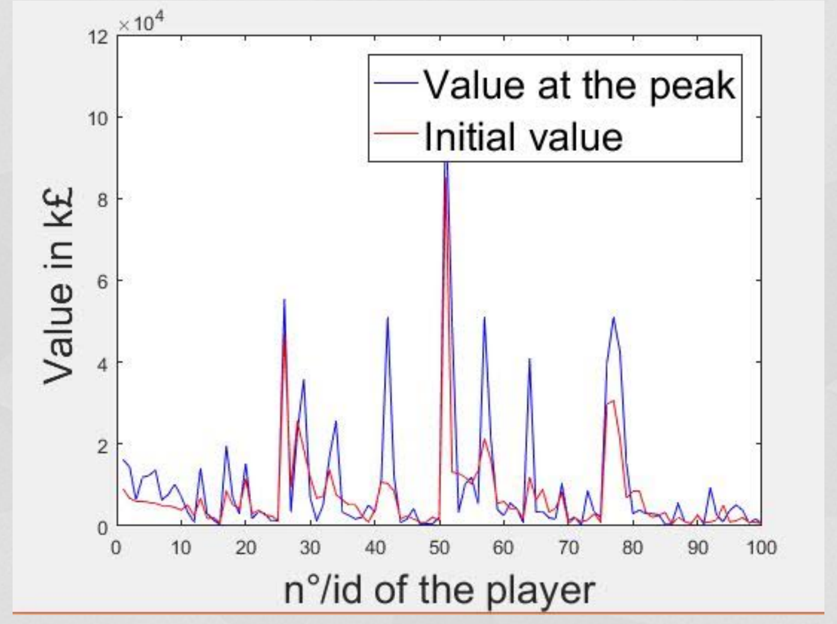
Dataset Analysis and basic statistics

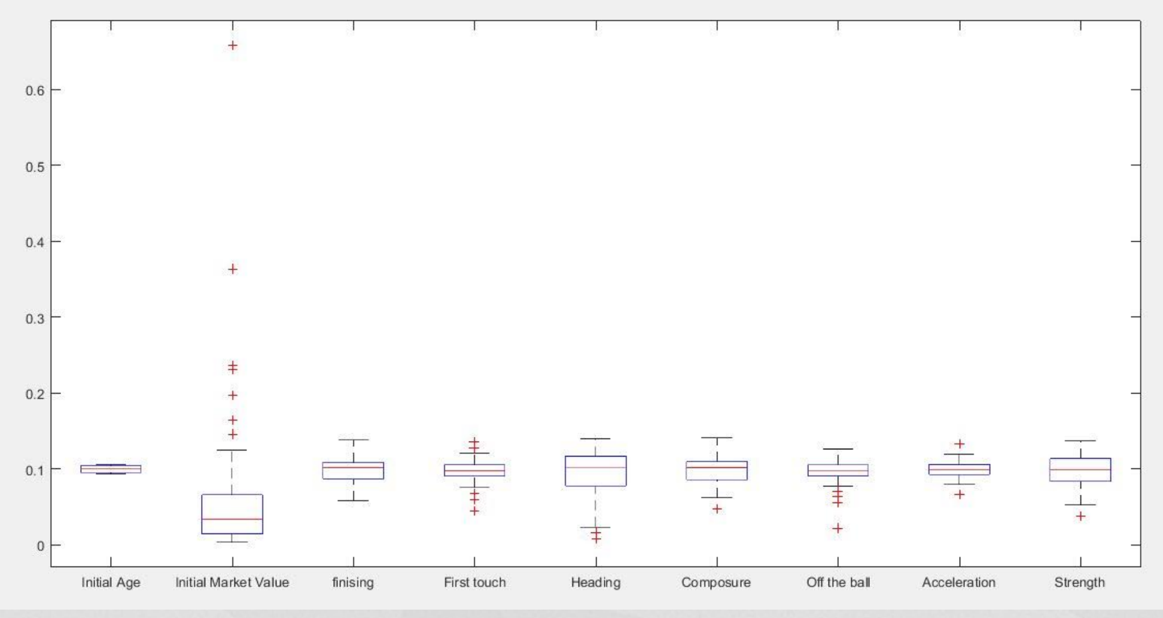
- Basic statistics pre and post normalization in table above
- Box plot post normalization (bottom/right)
  - Freak Value for initial market value is Lionel Messi, best player in the world
- Comparison of Initial vs. Peak market value (bottom/left)
- Finishing is correlated to composure (0.6) and Off the ball (0.5) (below left/center)
- Heading is correlated to strength (0.52) (below right)











References

Bishop, C.M., 2006. Pattern recognition. Machine Learning, 128

Breiman, L., 1996. Bagging predictors. Machine learning, 24(2), pp.123-140.

Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), pp.18-22. Nelder, J.A. and Baker, R.J., 1972. Generalized linear models. Encyclopedia of statistical sciences.

Song, L., Langfelder, P. and Horvath, S., 2013. Random generalized linear model: a highly accurate and interpretable ensemble predictor. BMC bioinformatics, 14(1), p.1.

Linear regression using Generalized Linear models

- Flexible model that can deal with other than normal error distribution
- Finds the best regression coefficients, the intercept and the random error
- Finds the maximum likelihood by least squared method (Nelder, 1972)
  - Pros
    - An understandable/interpretable model easy to use
    - Different variations of Linear regression can be used and compared
    - Takes account of the weight of each predictor
  - Cons
    - Not accurate for complex predictor/response relationships
    - Does not take into account correlations between variables
    - Limited for problems involving a high-dimensional input space (Bishop, 2006)
    - Risk of underfitting when little observations

Hypothesis Statement

We expect random forest to be much more accurate than any of the linear models for the following reasons:

- The ability to cure outliers
- The ability to deal with more complex variable/target relationships
- The random feature selection
- The inbuilt bootstrap aggregation

Random Forest using TreeBagger package

- Training of multiple decision trees
- Bootstrap replicates of the learning set used as new learning sets in each tree (Breiman, 1996)
- Random variables at each node of each tree to reduce correlation
  - Pros
    - Bagging increases accuracy of random forest considerably (Breiman, 2001)
    - Prevents over fitting
    - Reduces variance compared to individual trees
    - Deals with freak data points
    - Gives indications on variable importance
  - Cons
    - Difficult to visually interpret

Training and evaluation methodology

- Data split using Holdout function into 70 training observations and 30 test observations
- Evaluation of the models is based on Mean Squared Error
  - Generalized Linear Model
    - 4 different approaches/models
    - Compare and find the one with best accuracy/lowest MSE
  - Random forest
    - First, a standard model is trained free of hyperparameters
    - Hyperparameters are then adjusted to minimize MSE on test set

Choice of parameters and experimental results

General linear model

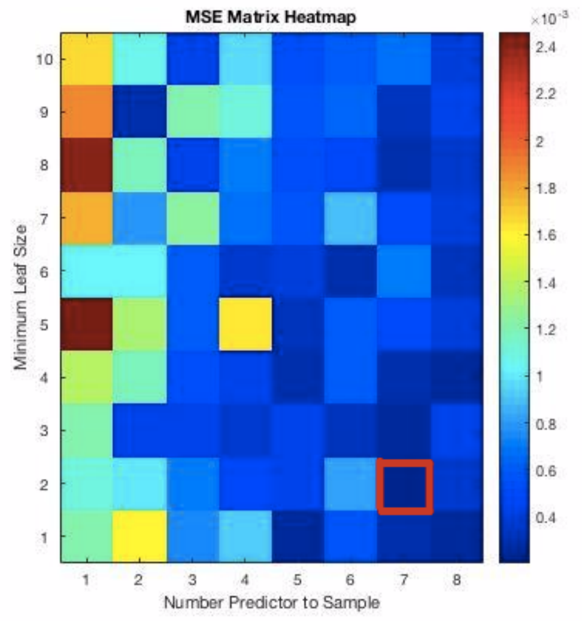
- Two hyperparameters have to be fixed:
  - "Random" refers to the distribution of the random error
  - "Link" is the function which links the response variable and the model.
- Each random error distribution corresponds to a type of regression (see table)

Model	Random	Link	Response variable	Mean Square Error
Linear regression	Normal	Identity	Continuous	7.6418 e-04
Poisson regression	Poisson	Log	Count	7.3255 e-04
Gamma model	Gamma	Reciprocal	Positive continuous	7.7856 e-04
Inverse gaussian model	Inverse gaussian	-2	Continuous	8.5535 e-04

Linear Regression		Random Forest
7.3255 e-04	Final MSE	2.0299e-04

Random Forest

- MSE with default hyperparameters: 3.8636e-04
- Number of trees was picked by comparing the predictions made by the forest and the predictions made by subsets of the forest, the Out of bag error, until performance stops improving (Liaw et Al., 2002)
- We created a loop that runs different combinations of the other two hyperparameters and picks the combination which minimizes the MSE. The red outline in the heatmap is that combination.
- Final Parameters: 5 trees, 2 MinLeafSize, 7 Predictors to sample



Analysis and critical evaluation of results

- As expected, Random forest is more performant than any of the General Linear Models for the reasons listed in the hypothesis statement.
- Random Forest
  - Results vary widely every time the model is run.
  - The model is therefore considered as high variance.
  - This is partly due to the small size of the database.
  - However it has a lower bias than the GLM models
- GLM
  - Poisson Regression is the most accurate model
  - Outliers affect the accuracy of the models.
  - The linearity of the model results in underfitting
    - This is again partly due to the small database
  - Still better at predicting random error than simple Linear Regression
- RF Hyperparameters Matrix
  - Particularly helpful at picking Leaf size and Predictors to sample from
  - However, limited to 2 parameters and unable to include number of trees or others
  - The matrix suggest using 7 random variables out of 9 to predict from.
  - This reduces the randomness and goes against the core idea of random forest which consists of randomly choosing subsets of variables to limit correlation between trees (Liaw, 2002).
  - Surprisingly, those parameters still minimize the MSE

Further work

- Mixture of GLM and Random Forest ensemble (Song, 2013)
  - Combining accuracy of Random Forest and high interpretability of GLM
- More observations and predictors would make for more accurate results
- Performing PCA to identify most relevant variables. This would make the model less complex yet more accurate