APPLICATION OF RANDOM FORESTS AND LOGISTIC REGRESSION ON MIXED MARTIAL ARTS DATA

GEDIMINAS SADAUNYKAS AND VAIDA GULBINSKAITE

Descriptive Statistics

	Minimum (won)	Minimum (lost)	Maximum (won)	Maximum (lost)	Mean (won)	Mean (lost)	Std. Deviation (won)	Std. Deviation (lost)
Fighter1_Implied_Win	0.13	0.06	0.96	0.95	0.62	0.45	0.18	0.18
WinStreak_diff_Scaled	0.00	0.00	1.00	1.00	0.52	0.49	0.14	0.15
LossStreak_diff_Scaled	0.00	0.00	1.00	1.00	0.45	0.46	0.13	0.13
TotalFights_diff_Scaled	0.00	0.00	1.00	1.00	0.48	0.48	0.18	0.18
WinRatio_Diff	-1.00	-1.00	1.00	1.00	0.06	-0.02	0.25	0.23
Height_Diff_Scaled	0.00	0.00	1.00	1.00	0.51	0.51	0.17	0.17
Age_diff_Scaled	0.00	0.00	0.97	1.00	0.46	0.51	0.16	0.16

Table 1: *descriptive statistics, segmented by classes: fight won (target = 1) and fight lost (target = 0)*

DESCRIPTION AND MOTIVATION OF THE PROBLEM

- Define fight instance as feature vector of differences in past performance and natural attributes, between two fighters.
- Investigate whether 'crowd-wisdom', approximated by implied probability from bookmaker odds, could be improved using declared features.
- Compare and contrast Logistic Regression(LR), Random Forest (RF) and implied bookmaker probability, using reliability curves. (DeGroot & Fienberg, 1982)

INITIAL ANALYSIS OF THE DATA SET INCLUDING BASIC STATISTICS

- The Mixed Martial Arts data has been scraped from sherdog.com and bestfightingodds.com.
- The scraped data includes 18 variables, however only 7 of them were used as predictors to build the models, due to redundance and multicolinearity. (Table 1)
- A binary fight outcome variable (1 = won, 0 = lost) was used as a class variable for the model.
- There was a total of 8583 cases in the dataset however, one of the predictor variables (WinRatio_Diff) had 294 missing values. Cases with missing values were removed from the dataset.
- Initial data analysis show no significant differences between winners and losers, except for implied bookmaker probability. It suggests noisy, redundant variables or/and strong non-linearities.

TWO MACHINE LEARNING MODELS WITH THEIR PROS AND CONS

LOGISTIC REGRESSION

- LR assumes binary target variable (1 or 0) and calculates weights via stochastic gradient descent in cost-function. Cost function is logarithmically transformed, as a result non-convexity is eliminated.
- Weights and predictor variables are inserted into sigmoid function, odd ratio and probability of binary target variable estimated.
- Default probability threshold of 0.5 is used to discriminate between classes.

PROS

- Unlike linear regression, it does not assume a linear relationship between predictors and output, multivariate normality or homoscedasticity.
- Can show which of the assessed variables have the most influence on the outcome, easily interpretable.(Tolles & Meurer 2016)

CONS

- Does not work with highly correlated variables. This can produce unstable estimates or leave the algorithm vulnerable to overfitting. (Cessie & Houwelingen 1992)
- Does not handle non-linearities in feature space well.

RANDOM FOREST

- Ensemble learner made of multiple decision trees. Each of tree using a sub-sample of data, randomly chosen with replacement. (bootstrap)
- Each tree is grown using particular set of random features with a default number of *sqrt(n),* where *n* is total number of features.
- Sampled data, in each leave, is split maximizing the decrease in impurity. (maximizing homogeneity of classes after split)
- Each data instance is sent down the tree, the majority vote is take in order to determine the predicted class. (aggregating)

PROS

- State of the art accuracy performance on many diverse problems. (Caruana & Mizil, 2006)
- Insensitive to noisy features, as a results randomization of features and samples. (Scornet et al., 2015)
- Intrinsic feature importance measure; ranks according to the extent to which splitting on particular feature 'purifies' the data.
- Highly parallelizable, therefore applicable to large datasets.
- Intrinsic generalization error measure-out of bag error, so no need for additional cross-validation.

CONS

- Biased in favor of features with higher number of levels/categories. (Strobl et al., 2007)
- RF have been observed to overfit for some datasets with noisy classification/regression tasks. . (Scornet et al., 2015)
- Unlike decision trees, RF are still difficult for humans to interpret.

HYPOTHESIS STATEMENT

- We expect both LR and RF to improve upon implied probability estimates, derived from bookmaker odds.
- RF is expected to have higher accuracy than LR, as it is one of the most robust and precise algorithms, tested on diverse problems. (Caruana & Mizil, 2006)
- RF, LR and bookmaker probabilities are expected to produce balanced reliability plots. (Caruana & Mizil, 2005) However, brier score should be higher for RF, due to tighter, non-linear fit.

DESCRIBTION OF METHODOLOGY

- The data was prepared, so that every observation was a fight, comprised of differences in skills and attributes, scaled to interval [0,1].
- Optimization carried by searching a viable hyperparameter space on full 80% of data. RF via Bayesian optimization, LR via grid-search.
- Best LR and RF models were trained on 80% full data, and tested on 20% data.
- The classification performance evaluated based on the accuracy measures, and reliability curves derived from probability estimates.

PARAMETER OPTIMIZATION AND EXPERIMENTAL RESULTS

LOGISTIC REGRESSION

PARAMETERS

- Exhaustive grid search was used to find the best learning rate. 11 different levels tested, spanning from 0.0001 to 0.9.
- Number of iterations are optimized separately, due to strictly monotonic improvement in performance. Learning curves are used to find the point of diminished marginal improvement

MAIN RESULTS

- The best learning rate was found to be 0.0005, which is considered a small learning rate, increasing chance for the gradient to follow steepest descent to the lowest cost value.
- Optimal number of iterations was chosen to be 50, even though after 40 learning curves show no improvement, we add 10 extra iterations on final model.(figure 2)

Measure	Random Forest	Logistic Regression	Bookmaker Probability
Training Accuracy	99%	68%	67%
Testing Accuracy	83%	66%	65%
Brier Score	0.117	0.213	0.213

Table 2: *model performance measures*

RANDOM FOREST

PARAMETERS

- Bayesian optimization (Snoek 2016) is used to find the best hyperparameters. for Random Forest generalization performance – out-of-bag error, used in the objective function. The Acquisition function is expected to improve objective function. Two optimized hyperparameters are number of random features to use per tree; and minimum number of samples per leaf.
- The number of trees are optimized separately, due to their inherently monotonic improvement in performance. Learning curves are used to approximate the point of the diminished marginal improvement.

MAIN RESULTS

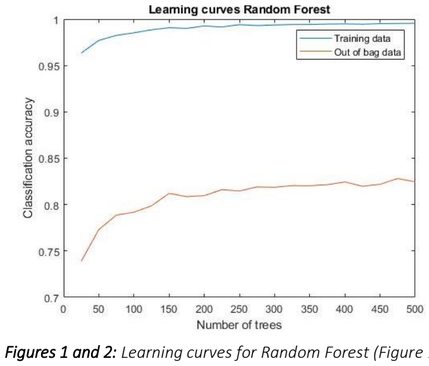
- The optimal number of maximum features per tree, was 2, (3 in some experiments. Not far from researched default – sqrt(total features).
- The optimal number of minimum samples per leaf was 1 (researched default), meaning that the trees are not pruned maximizing accuracy, at the risk of overfitting.
- The number of trees above 250 did not seem to improve accuracy, therefore it is chosen as an optimal number. (figure 1)
- It only took 5 functional evaluations to find the minimum of objective function, via Bayesian optimization. The Full exhaustive grid-search would have required full 99 evaluations. (99 pairs of chosen variables).

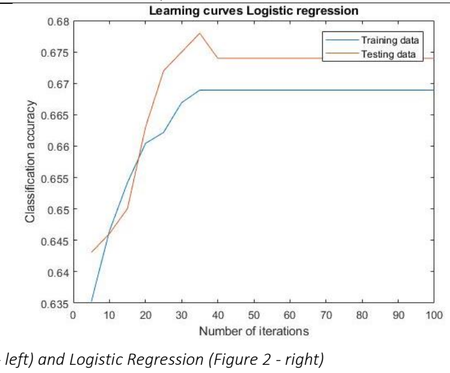
CRITICAL EVALUATION

- LR classifies data by splitting feature space with a single hyperplane, defined by the optimized parameters often called weights. RF performs binary splits in feature space, each split is perpendicular to axis of a single feature. Thus, without pruning or cross-validated optimization, RF tend to overfit the data.
- LR had a miniscule drop in performance measures between training and testing data sets, despite being unregularised. It suggests, that the difference in performance measures arise from non-linearities in feature space, rather than from high variance caused by overfitting. Also, it fails to improve upon bookmaker probability, not utilizing other variables, suggesting redundancy.
- RF overfits training data, as expected. It is no surprise, as decision trees are not pruned. (Min. leaf size = 1) Accuracy drop on testing data is significant (99% to 83%), but it still outperforms LR and improves upon bookmaker probability measures. Implied probability, as expected, is the most important feature (8.3 delta error), improving the heterogeneity of splits the most. Other features end up in range [4, 5 of delta error].
- LR calculates odds ratios, from which unbiased probability estimates are derived. However, in this instance it overestimates probabilities on the lower end, and underestimates on the higher end. It is surprising, as such kind of pull from the extremes is expected from Support Vector Machines, not from LR. (Caruana & Mizil, 2005)
- On the other hand, RF returns very balanced probability estimates, on the lower and mind range, as expected. (Caruana & Mizil, 2005) But,at the level of 0.7 in mean predicted value, there is substantial drop in fraction of positives, as it fails to estimate high probabilities effectively. Main reason, likely is insufficient and noisy data. Also, instances with very high probability of one fighter victory could be considered outliers, as skill level in the biggest mixed martial arts organization is high and reasonably similar. However, outliers does not explain why such behavior is not observed on the lower tail of reliability curves.

CONCLUSIONS AND FUTURE WORK

- Performance evaluation indicates, that RF significantly outperforms LR and crowd-intelligence measure, in fight outcome predictions.
- Bayesian optimization proved to be fast and effective technique for hyperparameter tuning in RF.
- Probability calibration techniques like Plats scaling and isotonic regression could be applied to improve brier score.
- More features could be extracted, such as striking accuracy, defense, volume etc.





Figures 1 and 2: *Learning curves for Random Forest (Figure 1 - left) and Logistic Regression (Figure 2 - right)*

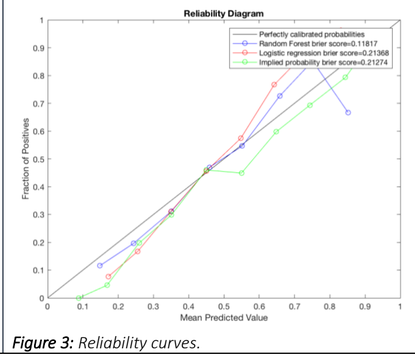
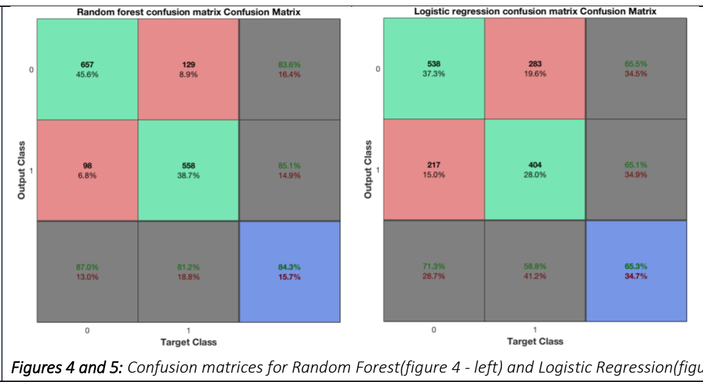


Figure 3: *Reliability curves.*



Figures 4 and 5: *Confusion matrices for Random Forest(figure 4 - left) and Logistic Regression(figure 5 - right)*

REFERENCES:

DeGroot, M., & Fienberg, S. (1982). The comparison and evaluation of forecasters. *Statistician.*, *32*, 12–22.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, *C*(1), 161–168.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*, (1999), 625–632.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms, 1–12.

Le Cessie, S., & Van Houwelingen, J. (1992). Ridge Estimators in Logistic Regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *41*(1), 191-201. doi:10.2307/2347628

Tolles,J,. Meurer,W,J,(2016) Logistic Regression Relating Patient Characteristics to Outcomes. *JAMA*.2016;316(5):533–534. doi:10.1001/jama.2016.765

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, *8*, 25. https://doi.org/10.1186/1471-2105-8-25

Scornet, E., Biau, G., & Vert, J. P. (2015). Consistency of random forests. *Annals of Statistics*, *43*(4), 1716–1741. https://doi.org/10.1214/15-AOS13