Prerna Prakash Gupta
Max Salomon

A CRITICAL COMPARISON OF LOGISTIC REGRESSION AND LOGITBOOST TREE ENSEMBLE

MOTIVATION OF THE ANALYSIS

- The goal of this critical analysis is two-fold: first, we compare classification capabilities of Logistic Regression (LR) and LogitBoost Tree Ensemble (LBTE). According to previous experimental results, LBTE should outperform LR, including in terms of Area Under Curve (AUC) **(1)**.
- Second, we use a specific dataset from Social Sciences research rather than a popular machine learning one in order to evaluate our models' performances in a realistic context, with complex and noisy data. Indeed, Social Sciences research and policy-making increasingly embrace Machine Learning **(2)(3)**.
- Practically, we assess both models' ability to accurately classify smokers and non-smokers, a relevant public policy goal.

HYPOTHESIS STATEMENT

- First, we expect LR, which tries to separate classes using a simple hyperplane, to result in low variance but potentially high bias. Meanwhile, we expect LBTE to exhibit low bias but high variance as it attempts to capture complex non-linear patterns in the data, potentially resulting in overfitting **(7)**.
- Second, we attempt to contradict the claim that boosting algorithm's number of iterations is irrelevant for classification (8).
- Third, we expect LBTE to be inappropriate for best discriminating between the two classes given their implicit assumption of equal cost for each class and the imbalanced nature of the data(8)
- Finally, we expect LBTE's best parameters to rely on numerous stumps, as often featured in practical applications (9) (10).

METHODOLOGICAL APPROACH

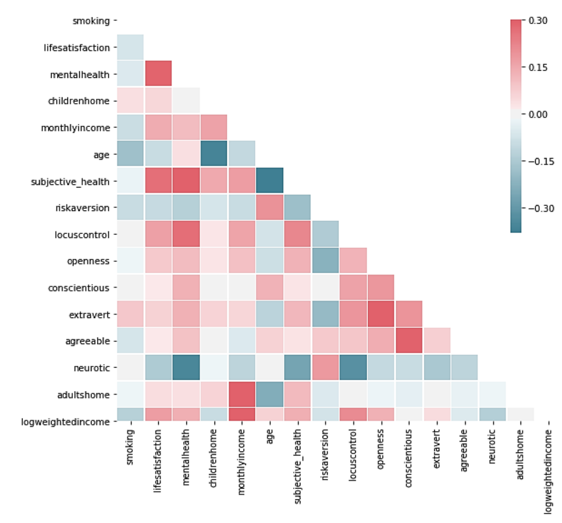
- The randomly sampled data was split into a training set and a testing set using a 80/20% stratified partitioning.
- Model tuning and selection was performed inside the training set using 5-fold cross-validation.
- Model selection was based on out-of-sample performance on AUC.
- Optimal parameters were found through a combined grid-search / random-search approach.
- AUC was chosen as the main selection criterion during cross-validation given the imbalance of the dataset. The metric does not depend on the choice of specific thresholds but on the ability of the model to accurately rank predictions. Indeed, the goal was to generate a balanced model able to accurately discriminate between smokers and non-smokers.
- Finally, models' optimal threshold for classifying as a smoker was chosen to maximize F1-Score.

DESCRIPTION OF THE DATA

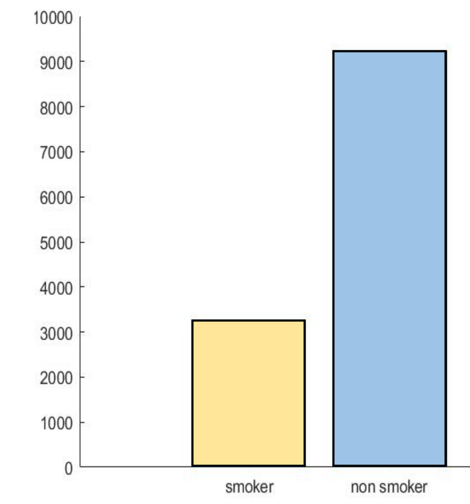
- The chosen data is a subset of the German Socio-Economic Panel, a large panel study extremely popular in Social Sciences Research **(4)**. The original dataset was randomly sampled to lower computational demands.
- The data features a random subset of 12,500 individual observations containing sociodemographic information as well as rarer psychometric variables.
- Features were selected based on their importance in past research on behavioural prediction (5, 6). 27 features were kept in the end, including 14 categorical ones that were transformed using one-hot-encoding. Continuous and discrete features were standardized and normalized as much as possible before analysis.

	smoking	age	life satisfaction	mental health	children home	adults home	log weighted income	subjective health	risk aversion	locus control	openness	conscientious	extravert	agreeable	neurotic
mean	No	52.54	7.78	50.92	0.61	2.13	7.38	3.33	5.42	3.94	4.76	5.5	4.48	5.11	3.51
	Yes	45.35	7.45	49.72	0.7	2.1	7.23	3.28	4.92	3.92	4.72	5.49	4.68	4.98	3.53
std	No	18.26	2.12	9.98	1.01	0.85	0.5	0.99	2.38	0.7	1.02	0.83	1.04	0.85	1.11
	Yes	14.8	2.17	10.66	1.06	0.88	0.51	0.98	2.48	0.75	1.05	0.88	1.06	0.9	1.15
skewness	No	0.1	-1.46	-0.84	1.76	1.32	-0.18	-0.42	0.07	-0.21	-0.27	-0.77	-0.26	-0.39	0.13
	Yes	0.29	-1.16	-0.78	1.79	1.22	-0.16	-0.41	0.2	-0.15	-0.24	-0.78	-0.37	-0.29	0.15

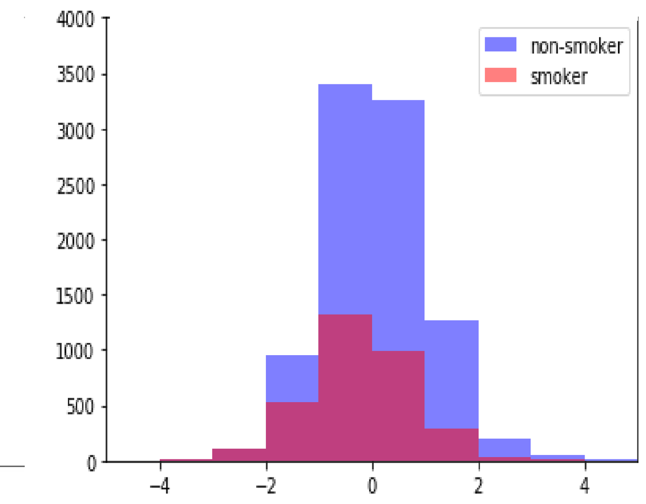
- Exploration of data revealed slightly different distributions for the two classes, especially for features like age and risk aversion. Meanwhile, categorical features relevant to parental education and occupation also offered possibility for better discrimination.
- The two classes to predict feature a slight imbalance (26% of smokers), giving an incentive to use measures like AUC and F1-Score to evaluate models rather than accuracy



Correlation matrix for numerical features



Proportion of smokers and nonsmokers



Distribution of logged weighted income for smokers and non-smokers (standardized values

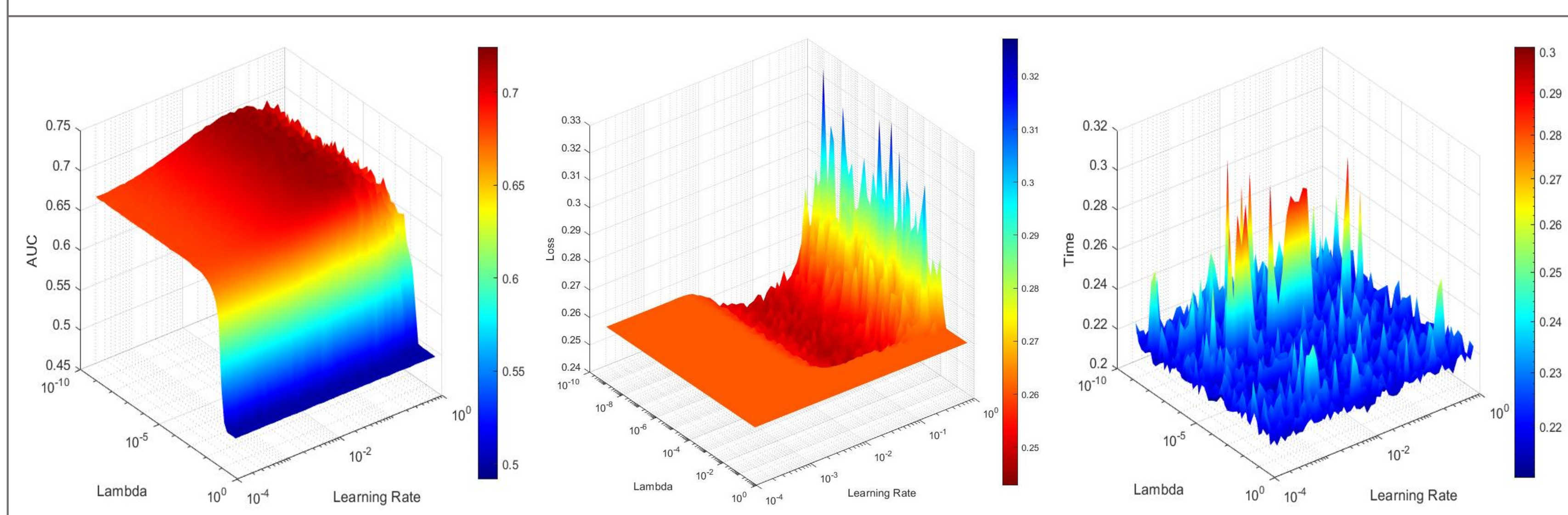
REGULARIZED LOGISTIC REGRESSION

- Probabilistic model outputting log odds originally, but which can also be used as a classifier by setting a classification threshold.
- Regression model part of the Generalized Linear Models as the log odds are modelled as a linear combination of the features.
- Given that it models linear relationships between predictors and log odds, the decision boundary is drawn by a simple hyperplane.
- Although binary by default, it can also accommodate multiple classes through multinomial logistic regressions.
- ➢ PROS
- Fast and simple to implement probabilistic model, with high interpretability, which explains its low variance and high bias. **(11)**
- Regularization methods, such as Lasso or Elastic Net make it potent when dealing with high-dimensional data as the regularization penalizes non-zero coefficients, providing an additional protection against overfitting and modelling noise .
- Can be made more computationally efficient by using optimization algorithms such as Stochastic Gradient Descent **(12)**.
- Quite robust to imbalanced data, especially if the training sample is representative of the real distribution of events.
- ➢ CONS
- Logistic Regression, while robust, tends to be less accurate than ensemble methods, especially to map complex dynamics. Although interactions can increase the sophistication of the models, they are limited in terms of complexity as seen from their simplistic decision boundaries.
- The performance of a logistic regression, as a discriminative model, depends mainly on the quality, quantity and representativeness of the training data

LOGITBOOST TREE ENSEMBLE

- Builds an ensemble of weak learners, trees
- Trees are grown sequentially: each tree is grown using information from previously grown trees.
- Boosting focuses on incrementally reducing bias, by reweighting training examples for the next ensemble model, based on a measure of error for the current ensemble of models
- The Logit Boost algorithm is an implementation of boosting where training examples are reweighted based on the logistic loss function. Logit Boost is a method of fitting an additive logistic regression to minimize the expectation of a binomial log-likelihood loss function which changes linearly with the classification error.
- ➢ PROS
- Logitboost is less sensitive to noise and outliers than initially proposed Adadboost which tries to minimize the expectation of an exponential loss function changing exponentially with the classification error, making AdaBoost more at risk of overfitting.
- ranks features according to their importance, which some other models can't do
- ➢ CONS
- In case of class imbalance, boosting algorithms may suffer from bias towards majority class, since their learning process is guided by their loss function and therefore correct classification, implicitly giving more weight to correct classification of majority class.
- It searches a less restricted space of models, allowing it to capture nonlinear patterns in the data, but making it less stable and prone to overfitting, especially with weak classifiers that are too complex

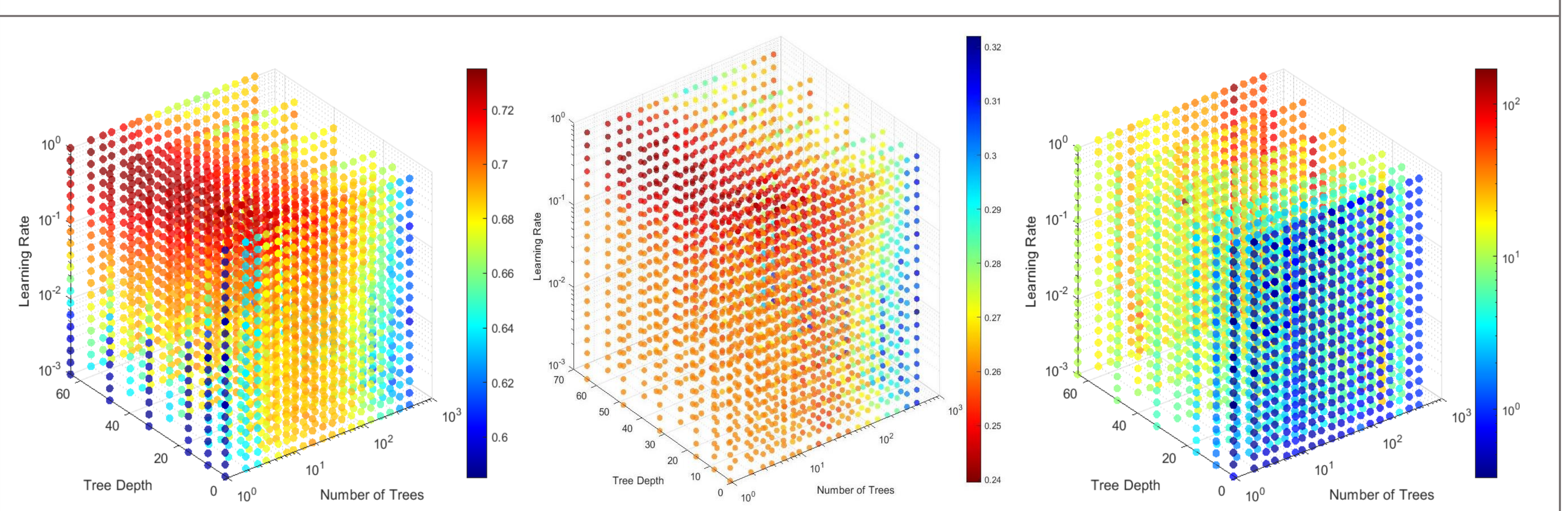
LOGISTIC REGRESSION HYPERPARAMETERS SELECTION



Lasso Logistic Regression performance in terms of AUC, Error Rate and Time, as a function of regularization parameter and learning rate of the Stochastic Gradient Descent Solver

- The different searches look for optimal values for number of trees, tree depth and learning rate for the Lasso Logistic Regression.
- Results of cross-validation accuracy and AUC suggest optimal performance for a low value of the regularization parameter and a high learning rate. This means that the best model avoids shrinking to 0 parameters and keeps a high amount of information.
- Models show little variability in terms of computational demands as indicated by the low training times, as compared to LogitBoost.

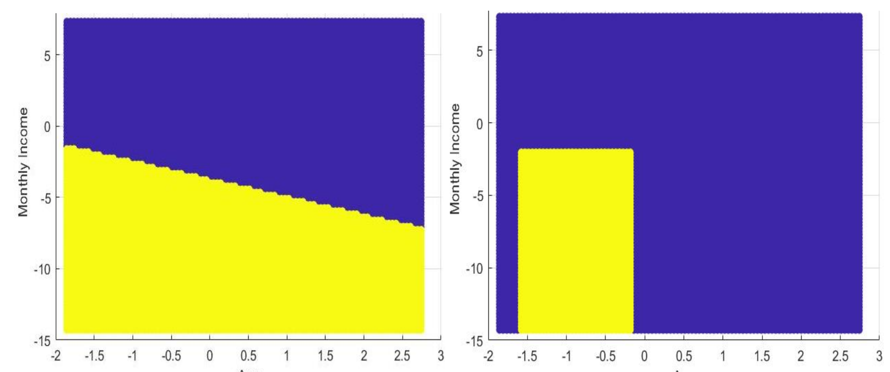
LOGITBOOST TREE ENSEMBLE HYPERPARAMETERS SELECTION



LogitBoost Tree Ensemble performance in terms of AUC (left), Loss (middle) and time (right), as a function of values for hyperparameters

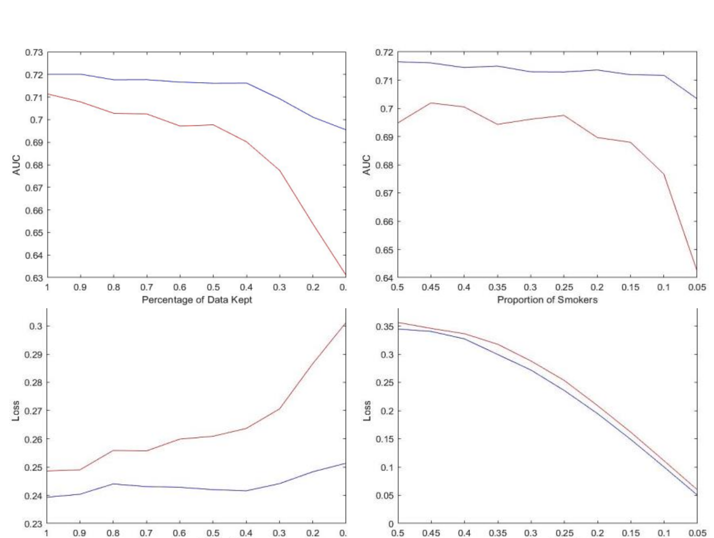
- The different searches look for optimal values for number of trees, tree depth and learning rate for LogitBoost.
- Results of cross-validation accuracy and AUC suggest optimal performance for a relatively low number of deep trees w
- While this is in line with Wyne and Buja (2007) research that boosting algorithms are insensitive to number of iterations, it is in contrast to DeBarr, D. and Wechsler, H. (2012) research which specified iterations of regression stumps and weak learners to be fit on training data to reach optimal model
- Work on binning specified that small changes in binning of data had an impact on improvement in accuracy and computation time.
- Measuring the training time shows high cost of increasing number of iterations and deeper trees.

DECISION BOUNDARIES COMPARISON



- By observing the decision boundaries of both models, we better understand why LR counts as a high bias / low variance model and why LBTE is more prone to overfitting.
- While LR tries to fit a hyperplane separating both classes, LBTE is able to fit more complex nonlinear interactions, which helps fitting complex data but also make it vulnerable to overfitting and high variance.

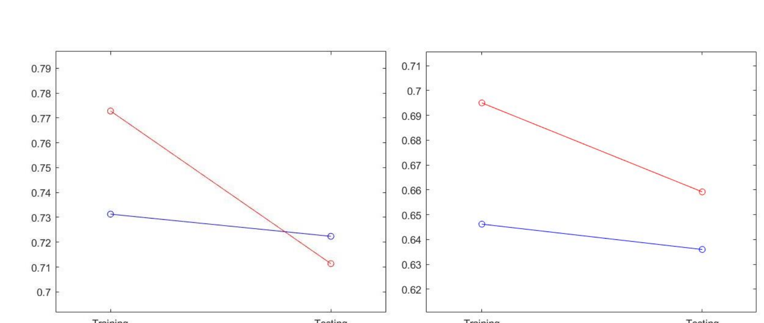
SAMPLE SIZE AND PANEL IMBALANCE



- Here, we manipulate the size of the training sample to observe changes in performances, both in terms of AUC and misclassification.
- Experiments with the percentage of data kept suggest that Logistic Regression is more robust than LogitBoost and shows less variability. As predicted, performance drops as training sample size decreases.
- Experiments with the proportion of smokers in the training set suggest that LogitBoost benefits only slightly from more balanced proportions in terms of AUC while there is hardly any change for Logistic Regression. Finally, misclassification error increases with the proportion of smokers given that both models are discriminative models, adjusting the prior probability of an observation being a smoker.

EVALUATION OF THE RESULTS

- In line with expectations, logistic regression shows high bias and LogitBoost is indicative high variance with i) misclassification error for both training and test data in logistic regression exceeding LogitBoost ensemble and ii) misclassification error for both training and test data showing lower variability and more stability in logistic regression as compared to LogitBoost
- In line with expectations, LogitBoost model shows that in case of class imbalance the hyper parameters that minimize the misclassification error, are not ones giving optimality in terms of AUC.
- When results are analysed, it is noticed that LogitBoost is slightly outperforming logistic regression in terms of accuracy for train, validation and test error. However, when we see in terms of maximum AUC, while LogitBoost vastly outperforms logistic regression during, logistic regression outperforms in terms of test error.
- This suggests that low-bias LogitBoost model is overfitting the data and likely processing minority class instances as noise (while minimizing misclassification error at each iteration).
- In our case, given a noise ridden, real-world survey data, the low-bias and high-variance LogitBoost ensemble method overfits the training data (as opposed to high-bias & low-variance, relatively stable logistic regression), giving rise to sub-optimal results.



AUC (left) and accuracy (right) for Logistic Regression (blue) and LogitBoost (red), for both final training and testing

	Accuracy			AUC		
	Train	Test	Cross-Validation	Train	Test	Cross-Validation
Logistic Regression	0.65	0.64	0.75	0.73	0.72	0.72
LogitBoost	0.7	0.66	0.76	0.77	0.71	0.74

Final results for Logistic Regression and LogitBoost during final training and testing as well as cross-validation out-of-sample evaluation

LESSONS LEARNED AND FUTURE WORK

- Performances could be increased by using different methods than undersampling the majority class when handling imbalanced data: ADASYN could be a better alternative to SMOTE as it generates more realistic synthetic cases.
- Performances of the LogitBoost could be improved far more by optimizing further the untouched hyperparameter. This would result in a longer grid-search however. Meanwhile, logistic Regression could be improved by adding interaction to better model complexity.

1. Pham, B.T. and Prakash, I., (2019). Evaluation and comparison of LogitBoost Ensemble, Fisher's Linear Discriminant Analysis, logistic regression and support vector machines methods for landslide susceptibility mapping. Geocarto International, 34(3), pp.316-333.
2. Yarkoni, T. and Westfall, J., 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), pp.1100-1122.
3. Sanders, M. and Lawrence, J. (2017). *Using Data Science in Policy*. [online] Bi.team. Available at: https://www.bi.team/publications/using-data-science-in-policy/ [Accessed 25 Nov. 2019].
4. Goebel, J., Grabka, M., Liebig, S., Kroh, M., Richter, D., Schröder, C. and Schupp, J. (2019). The German Socio-Economic Panel (SOEP). Jahrbücher für Nationalökonomie und Statistik, 239(2), pp.345-360.
5. Armin, F., Becker, A., Dohmen, T., Enke, B., Huffman, D. and Sunde, U. (2015). The Nature and Predictive Power of Preferences: Global Evidence. SSRN Electronic Journal.
6. Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J. and Wagner, G. (2011). INDIVIDUAL RISK ATTITUDES: MEASUREMENT, DETERMINANTS, AND BEHAVIORAL CONSEQUENCES. Journal of the European Economic Association, 9(3), pp.522-550.
7. Landwehr, N., Hall, M. and Frank, E. (2005). Logistic Model Trees. Machine Learning, 59(1-2), pp.161-205.
8. Buja, A., Mease, D. and Wyner, A. (2007). Comment: Boosting Algorithms: Regularization, Prediction and Model Fitting. Statistical Science, 22(4), pp.506-512
9. DeBarr, D. and Wechsler, H. (2012). Spam detection using Random Boost. Pattern Recognition Letters, 33(10), pp.1237-1244.
10. Friedman, J., Tibshirani, R. and Hastie, T. (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). The Annals of Statistics, 28(2), pp.337-407.
11. Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia medica: Biochemia medica, 24(1), 12-18.
12. Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010 (pp. 177-186). Physica-Verlag HD.E