ML1819 Research Assignment 1

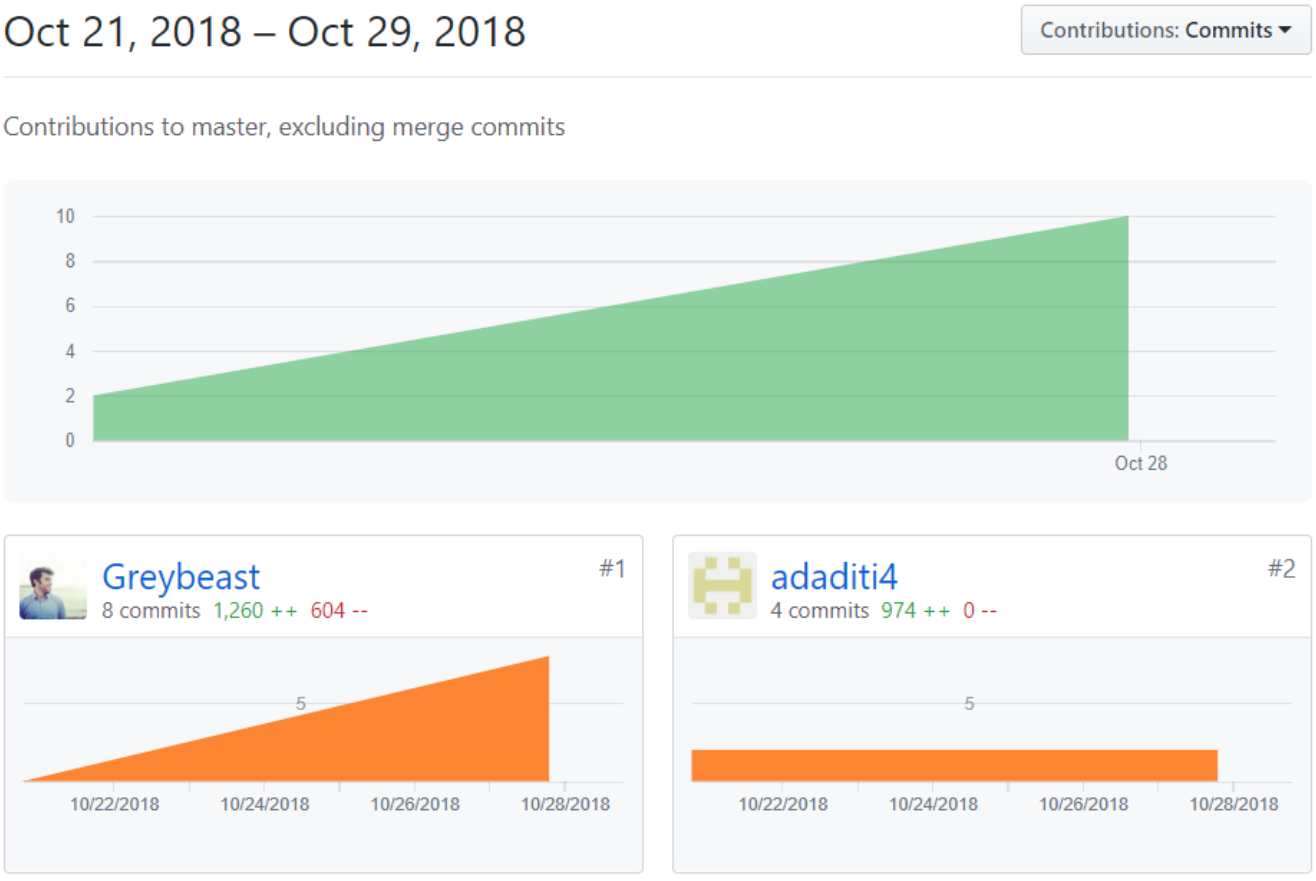
Team: 27
Task ID: 203

We would like to submit our Research assignment 1 on "***Impact of Individual Algorithms on Ensemble model".*** All the work submitted is original and all the team members contributed equally in developing code and writing report.
Total word count of the report is 998.

All the work relevant to this research is placed at below GitHub repository:
https://github.com/adaditi4/ML1819-Task-203-Team-27

URL for the source code repository activity of our team:
https://github.com/adaditi4/ML1819-Task-203-Team-27/graphs/contributors

Commit activity of project and individuals is as follows:



Impact of Individual Algorithms on Ensemble Model

Aditi Dubey
Trinity College Dublin
Dublin, Ireland
dubeya@tcd.ie

Rohit Kumar
Trinity College Dublin
Dublin, Ireland
kumarro@tcd.ie

Jian Li
Trinity College Dublin
Dublin, Ireland
lij12@tcd.ie

I. INTRODUCTION

In the corporate world, ensemble models are utilized in order to increase the accuracy to avoid losses & increase the profitability in monetary terms. However, the development of these ensemble models which are developed to achieve the most optimized result is purely based on trial & error. The algorithms used in these ensemble models are selected on basis of the data & business problem at hand. There exists no relationship at present which can emphasize the impact of each algorithm on the overall model / other models. An equation or outcome which can identify the relation between multiple algorithms and the impact that they can have on the final result based on datasets can ease the development by reducing the span of training time taken by these models. In this paper, we intend to identify the relation between multiple algorithms in individual ensemble models along with the contribution of each algorithm to the overall accuracy.

II. RELATED WORK

Oguto Jo and Jacques W. have carried out a research to identify the order of importance of algorithms, based on the accuracy achieved as an outcome. However, there exists no research about the impact of individual algorithms over the ensemble models. We have tried to utilize the conclusions from the research by Jo and Jacques to identify the order of inclusion of models in our ensemble models.

III. METHODOLOGY

To test the relation between the algorithms and their impact, two scenarios were considered i.e. Regression and Classification problems. These problems were solved using multiple algorithms and then those models were ensembled using mean of their summations. The details of the entire methodology are mentioned as follows:

A. Datasets used

Multiple open-source datasets (12, overall 24) for each category of problems were used. These datasets were taken from online platform "Kaggle". These datasets were pre-processed to achieve the required format for model input. The rows containing NA values were dropped because this had no impact on our research problem. Also, the categorical values were hot encoded to convert them into numerical columns. To generalize the code, a few columns containing text and date were removed from all datasets. This process was done manually.

B. Algorithms and Parameters

The algorithms used are:

1. Linear/Logistics Regression
2. RandomForest
3. Gradient Boosting Method
4. Xtreme Gradient Boosting
5. Light Gradient Boosting Method

All the regressor/classification models from the above-mentioned packages are used to predict the values of dependent variables from the test dataset. This test dataset was created from every dataset on which these 5 models were trained. The predictions of individual models were calculated and ensembled in a given order. The order for calculating ensemble accuracy is as follows:

1. Linear/Logistic Regression
2. Lin/Log + RandomForest
3. Lin/Log + RandomForest + Gradient Boosting
4. Lin/Log + RandomForest + Gradient Boosting + Xtreme Gradient Boosting
5. Lin/Log + RandomForest + Gradient Boosting + Xtreme Gradient Boosting + Light GBM

This sequence of models added was selected on basis of the research carried out by Oguto [1] and Jacques W. [2]

C. Evaluation

The impact of each additional model was identified by calculating the accuracy of new ensemble model with N+1 models against the accuracy of ensemble model with N models. For e.g. to identify the impact of GBM, the percentage increase in accuracy was calculated between Accuracy_Linear_RF_GBM and Accuracy_Linear_RF.

D. Observational plots

To visualize and understand the impact of individual models, the minimum impact for first model is extrapolated to 100. The additional impact of every model over the ensemble is then multiplied to the previous accuracy. For e.g. assuming the accuracy from linear regression is 100, the accuracy of ensemble models with Random Forest is 144.25 as visible in the figure 4.4.

The initial 6 rows (out of 12 datasets) are showcased in fig. 4.1 (Regression problems) and fig. 4.2 (Classification problems), where each row corresponds to one dataset. In the displayed table, the values of models (individual and ensemble) are showcased. As shown in fig. 4.1 and fig. 4.2 there is a continuous increase in accuracy of N+1th model with respect to ensemble of Nth model for every row (Datasets) for classification and regression problems.

Fig. 4.1 Accuracy of Classification model for 6 datasets for various ensemble models

Accuracy_linear_ model	Accuracy_lin_rf_ ensemble	Accuracy_lin_rf_ gbm_ensemble	Accuracy_lin_rf_gbm _xgb_ensemble	Accuracy_lin_rf_gbm _xgb_lgbm_ensemble
98.08%	98.86%	98.83%	98.77%	98.89%
78.36%	85.87%	87.22%	87.67%	87.72%
75.95%	83.66%	85.07%	85.54%	85.94%
24.83%	35.71%	37.87%	38.57%	39.49%
51.30%	54.00%	54.22%	54.19%	54.66%
81.97%	88.07%	87.23%	86.38%	87.14%

Fig. 4.2 Accuracy of Regression model for 6 datasets for various ensemble models

Accuracy_logistic_ model	Accuracy_log_rf_ ensemble	Accuracy_log_rf_ gbm_ensemble	Accuracy_log_rf_gbm _xgb_ensemble	Accuracy_log_rf_gbm _xgb_lgbm_ensemble
92.90%	93.02%	93.78%	93.61%	93.78%
77.30%	78.73%	87.94%	88.31%	91.01%
97.40%	97.40%	100.00%	100.00%	100.00%
98.68%	98.51%	99.16%	99.11%	99.18%
84.19%	84.08%	83.98%	84.03%	84.03%
84.59%	84.53%	84.11%	84.17%	84.17%

IV. RESULTS

From the developed models over multiple datasets, some clear trends are visible. In figure 4.3 and figure 4.4, the percentage increase has been showcased. This increase is calculated using the following technique:
(Accuracy_[N+1] – Accuracy_[N]) / Accuracy_[N]

The above stated formula gives the percentage contribution of algorithm at [N+1]th position to the overall accuracy.

For classification problems, when Logistic regression is the first model to be considered, RandomForest gives an impact of 0.39% whereas Gradient Boosting Algorithm contributes 7.32% in the increment of overall accuracy of the ensemble model. The impact of Xtreme Gradient Boosting is supressed by GBM, however, LightGBM has a contribution of 0.89%.

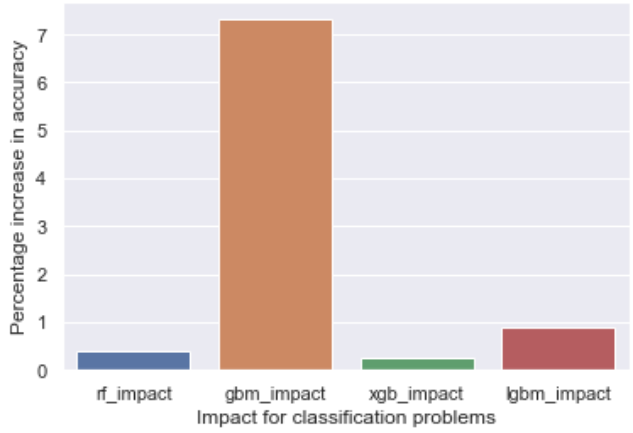


Fig. 4.3

For regression problems, the majority impact visible is because of RandomForest i.e. 44.25% and GBM shows an impact of 6.78%. The individual impact of LightGBM and XGB is under 1.5% each.

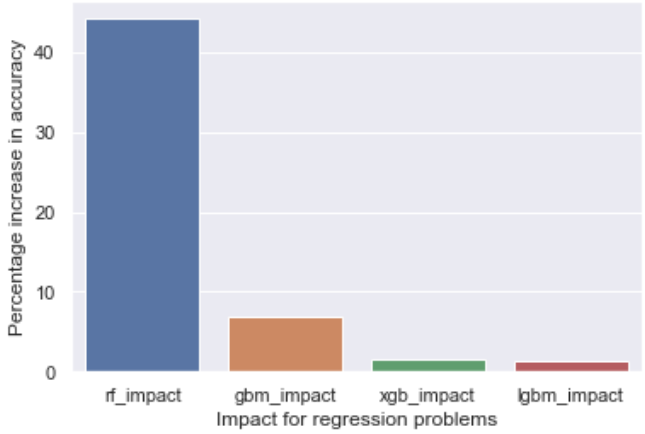


Fig. 4.4

Fig.4.5 and Fig.4.6 shows the overall impact of ensemble models considering Linear model / Logistic model for classification and regression problems respectively. In order to calculate the overall **consolidated impact of algorithms,** the mean of the percentage increase of accuracy is considered which is visible in Fig 4.3 and Fig 4.4. The calculation for this procedure is as follows:

- Accuracy of [N]th model * mean of [N+1]st model accuracy.

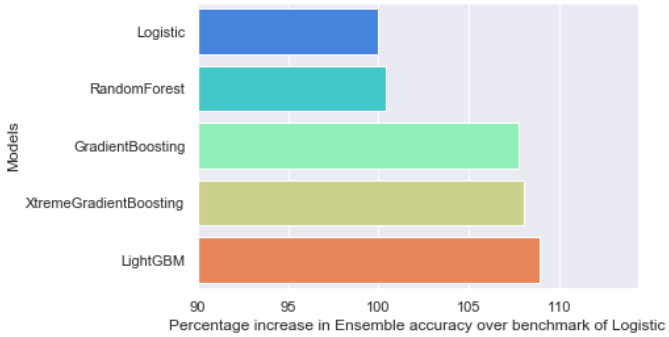


Fig. 4.5

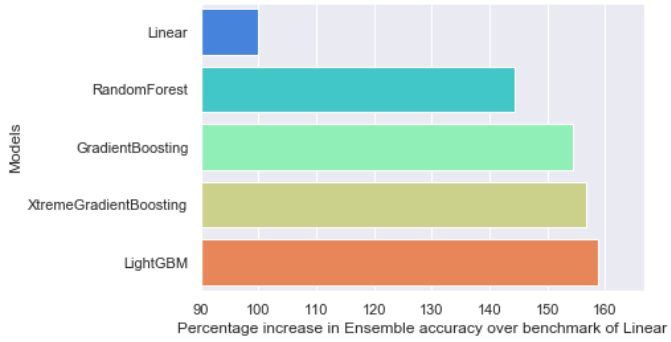


Fig. 4.6

As visible in graphs, the maximum increase in classification and regression models is brought by Gradient Boosting Method and RandomForest respectively.

V. LIMITATIONS & OUTLOOK

The total number of datasets used to understand the impact of individual models is not sufficient enough to derive relations between the stated algorithms. Other variables like dataset size, variance of dataset, standard deviation of variables, etc. can be utilized to find the relation between the algorithms. The entire exercise has to be repeated on multiple datasets varying in various aspects, in order to come to a concrete conclusion. The impact was calculated basis basic models; the overall number of ensemble models can be decreased and relation between high end algorithms can be identified as well.

VI. REFERENCES

[1] A comparison of random forests, boosting and support vector machines for genomic selection

[2] Comparison of 14 different families of classification algorithms on 115 binary datasets