GBR with SVR

https://github.com/liorsidi/GradientBoostingRegressorSVR

Regressors

For this assignment, we evaluated three type of gradient boost regressors:

1. Original Gradient Boosting Regressor - Scikit-learn, train trees with average in the leaves
2. Gradient boosting Regressor SVR in leaves – we generalized the scikit-learn implementation to get any tree model and not to be a specific decision tree. under BaseGradientBoosting class we added 2 new parameters to the constructor:
    a. Tree constructor: the constructor function of the new tree
    b. Extra_args: relevant args for the tree

    We also implemented a new class of decision tree regressor, DecisionTreeSVRRegressor that hold SVR model in the tree leaves. We use this tree in our new gradient boosting regressor implementation
3. Gradient boosting SVR wrapper – GradientBoostingRegressorSVRSimpleWrapper, train an original GBR with the regular decision tree but when predict it uses SVR models in the leaves in each stage. This model apply the SVR only on predict.

Experiment

We evaluate “max depth” and the “min split” tree parameters on 5 different datasets, with 5 fold cross validation, for each evaluation we calculated the MSE (mean & std), MAE (mean & std), and the training time of the model.

Regard all other gradient boost and the SVR parameters; we set the following fixed parameters:

- Gradient boost - 100 estimators with 0.01 learning rate with friedman_mse criterion
- SVR in the leaves - rbf kernel with 0.1 epsilon

(These parameters showed in general good enough baseline results compare to other parameters)

Datasets

Dataset 1 - Boston Housing

https://archive.ics.uci.edu/ml/datasets/housing

Data Set Characteristics:	Multivariate	**Number of Instances:**	506	**Area:**	N/A
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	14	**Date Donated**	1993-07-07
Associated Tasks:	Regression	**Missing Values?**	No	**Number of Web Hits:**	331146

Source:
Origin:
This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University.
Creator:
Harrison, D. and Rubinfeld, D.L.
'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.

Data Set Information:
Concerns housing values in suburbs of Boston.

Dataset 2 - Facebook metrics

https://archive.ics.uci.edu/ml/datasets/Facebook+metrics

Abstract: Facebook performance metrics of a renowned cosmetic's brand Facebook page.

Data Set Characteristics:	Multivariate	**Number of Instances:**	500	**Area:**	Business
Attribute Characteristics:	Integer	**Number of Attributes:**	19	**Date Donated**	2016-08-05
Associated Tasks:	Regression	**Missing Values?**	N/A	**Number of Web Hits:**	38115

Source:
Created by: SÃ©rgio Moro, Paulo Rita and Bernardo Vala (ISCTE-IUL) @ 2016

Data Set Information:
The data is related to posts' published during the year of 2014 on the Facebook's page of a renowned cosmetics brand.
This dataset contains 500 of the 790 rows and part of the features analyzed by Moro et al. (2016). The remaining were omitted due to confidentiality issues.

Dataset 3 - Bike Sharing Dataset

https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

Abstract: This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.

Data Set Characteristics:	Univariate	**Number of Instances:**	17389	**Area:**	Social
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	16	**Date Donated**	2013-12-20
Associated Tasks:	Regression	**Missing Values?**	N/A	**Number of Web Hits:**	131969

Source:
Hadi Fanaee-T
Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto
Original Source: http://capitalbikeshare.com/system-data
Weather Information: http://www.freemeteo.com
Holiday Schedule: http://dchr.dc.gov/page/holiday-schedule

Data Set Information:
Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

Dataset 4 - Online News Popularity

https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

Abstract: This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. The goal is to predict the number of shares in social networks (popularity).

Data Set Characteristics:	Multivariate	**Number of Instances:**	39797	**Area:**	Business
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	61	**Date Donated**	2015-05-31
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	107290

Source:
Kelwin Fernandes (kafc â€˜@â€™ inesctec.pt, kelwinfc â€™@â€™ gmail.com) - INESC TEC, Porto, Portugal/Universidade do Porto, Portugal.
Pedro Vinagre (pedro.vinagre.sousa â€™@â€™ gmail.com) - ALGORITMI Research Centre, Universidade do Minho, Portugal
Paulo Cortez - ALGORITMI Research Centre, Universidade do Minho, Portugal
Pedro Sernadela - Universidade de Aveiro

Data Set Information:
* The articles were published by Mashable (www.mashable.com) and their content as the rights to reproduce it belongs to them. Hence, this dataset does not share the original content but some statistics associated with it. The original content be publicly accessed and retrieved using the provided urls.
* Acquisition date: January 8, 2015
* The estimated relative performance values were estimated by the authors using a Random Forest classifier and a rolling windows as assessment method. See their article for more details on how the relative performance values were set.

Dataset 5 - Student Performance

https://archive.ics.uci.edu/ml/datasets/Student+Performance

Abstract: Predict student performance in secondary education (high school).

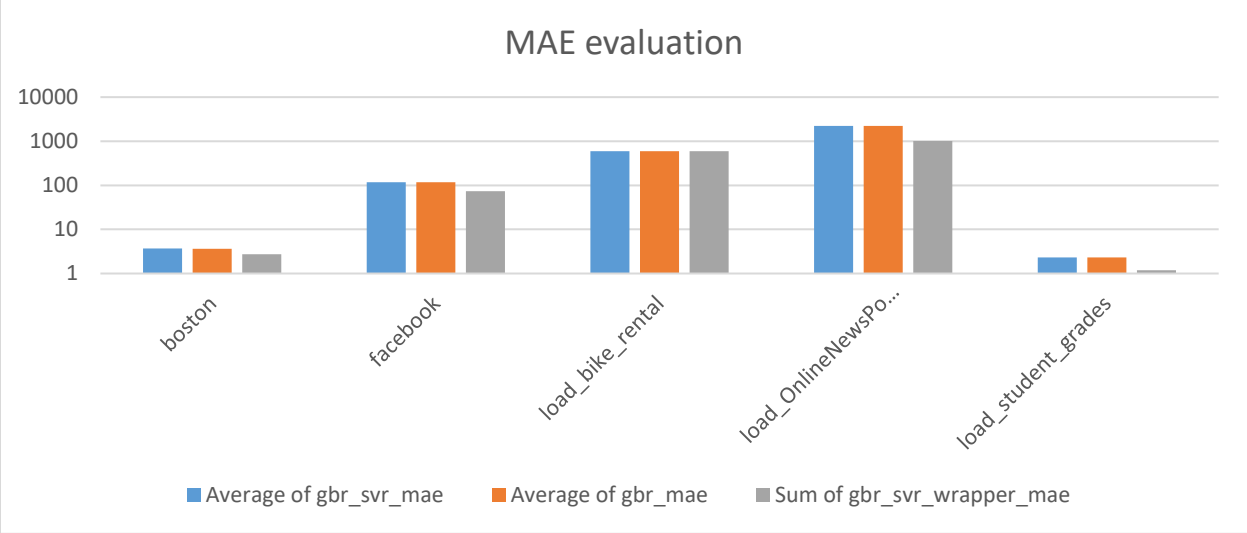
Data Set Characteristics:	Multivariate	Number of Instances:	649	Area:	Social
Attribute Characteristics:	Integer	Number of Attributes:	33	Date Donated	2014-11-27
Associated Tasks:	Classification, Regression	Missing Values?	N/A	Number of Web Hits:	154031

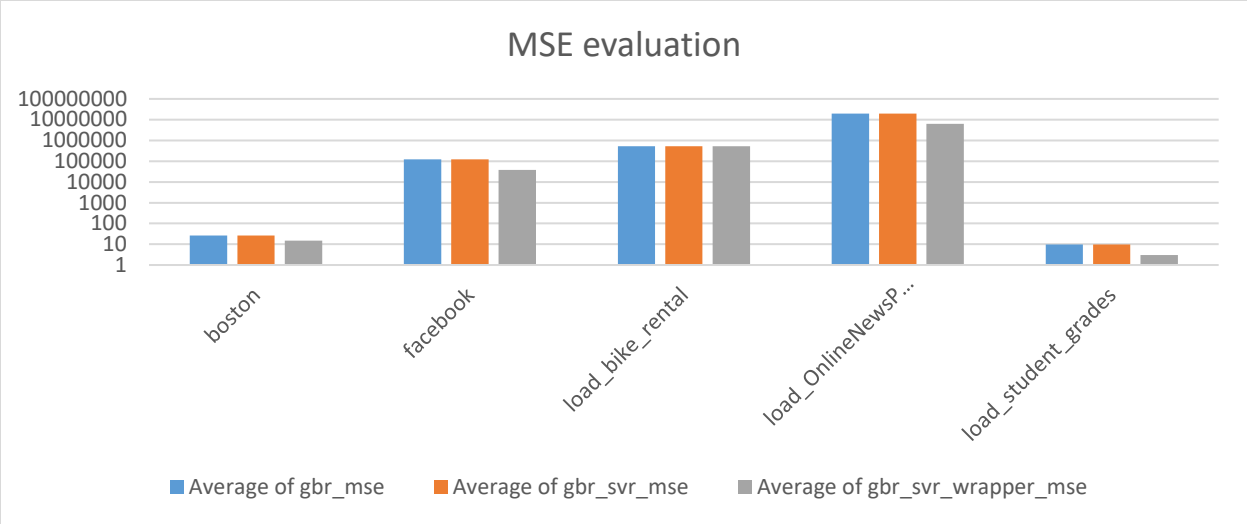
Source:
Paulo Cortez, University of Minho, GuimarÃ£es, Portugal, http://www3.dsi.uminho.pt/pcortez

Data Set Information:
This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

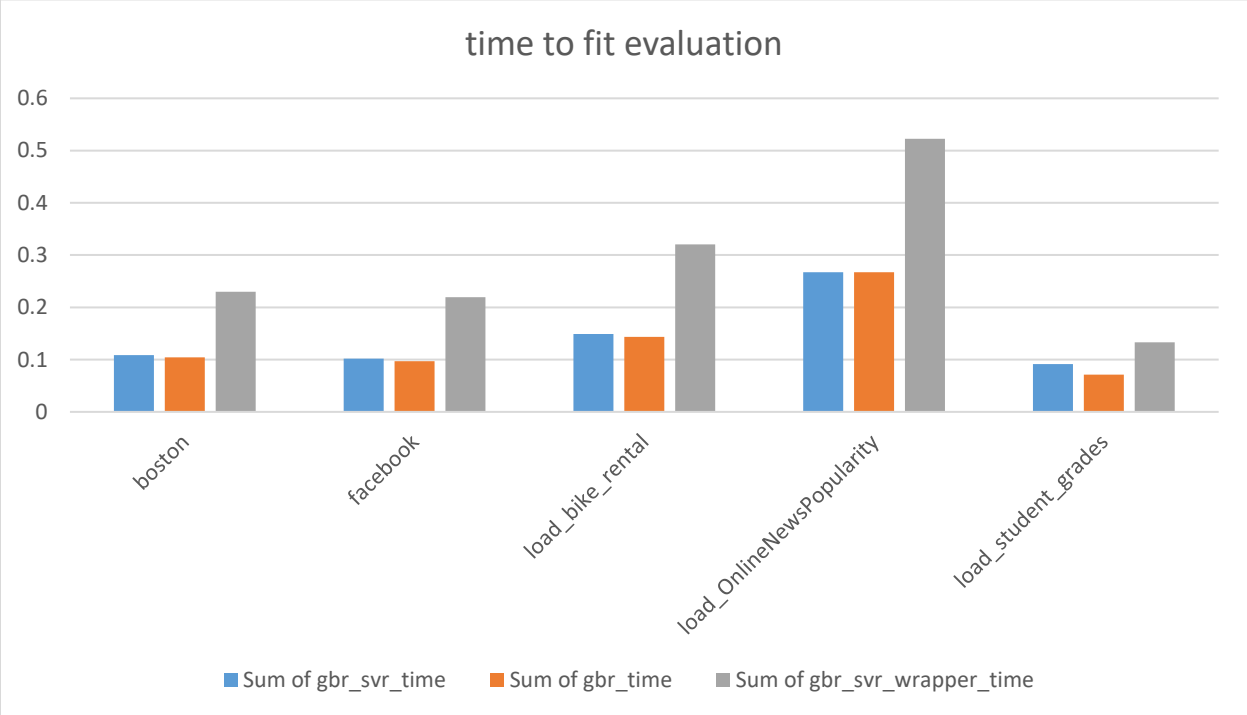
Results

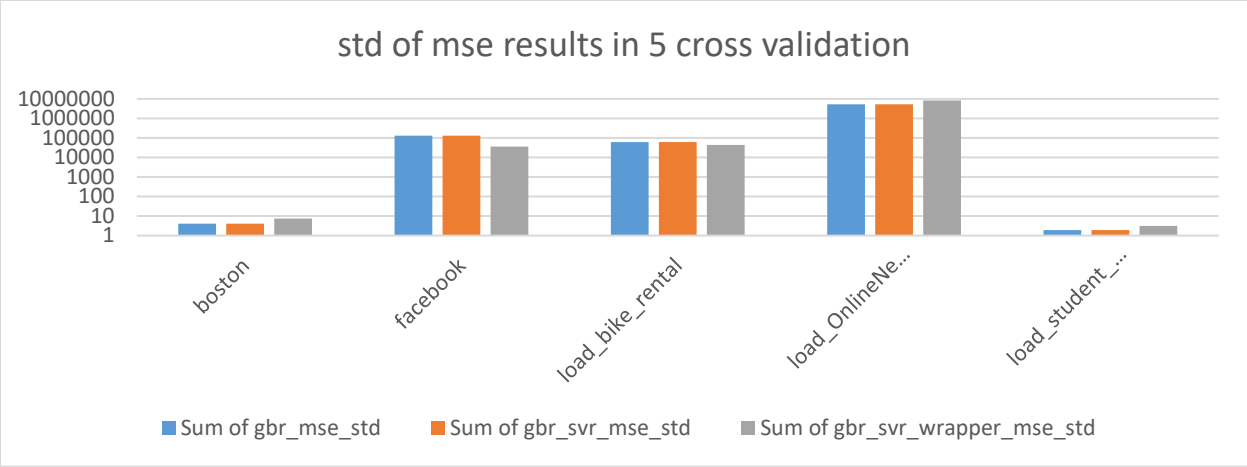
GradientBoostingRegressorSVRSimpleWrapper has a significantly lower error rate compare to the other regressors. We assume the wrapper model showed better results due to the extra train after the gradient boost model-training phase. The non-wrapper models showed more or less similar results meaning that the optimization of the residual on SVR results does not show significant improvement compare to the improvement over the mean.





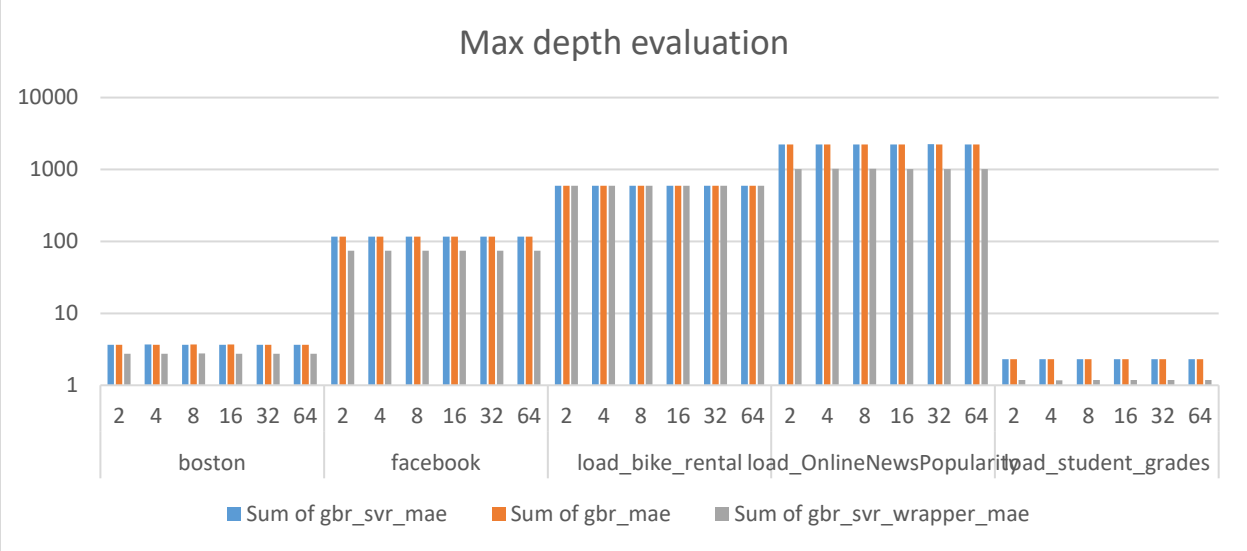
But the GradientBoostingRegressorSVRSimpleWrapper time to train is higher than the others and the standard deviation of its results from the 5 fold cross validation was also higher meaning that the model might be is less stable than the others

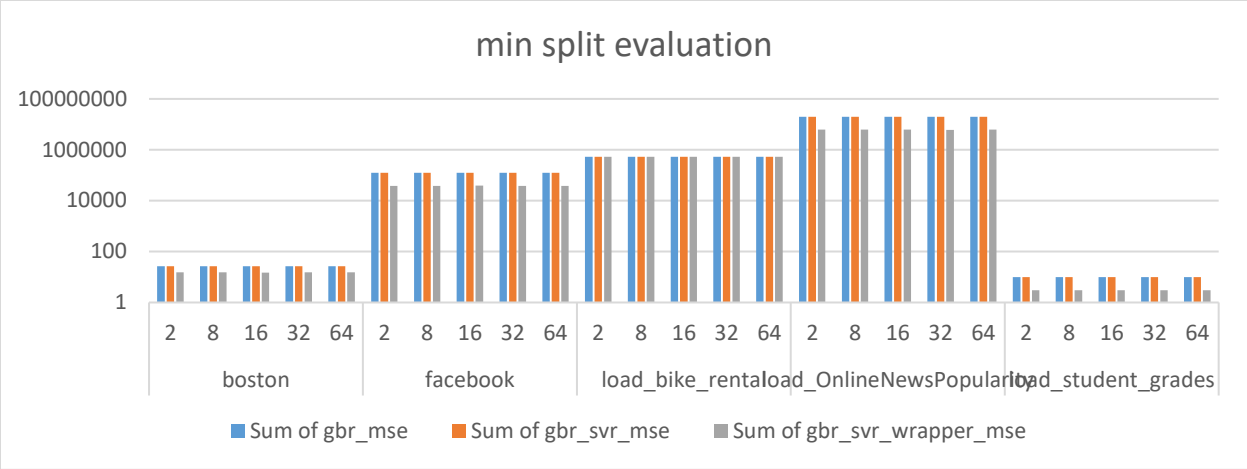




For evaluating the “max depth” and “min spit” parameters we ran experiment with same validation method and with all other configurations presented above.

Unfortunately, we could not find any significant value of “max depth” or “min split” that will decrease the error significantly. We tested their values of 2, 4, 8, 16, 32 and there wasn’t a specific value that outperform the others




min split evaluation
100000000
1000000
10000
100
1
2 8 16 32 64 2 8 16 32 64 2 8 16 32 64 2 8 16 32 64 2 8 16 32 64
boston
facebook
load_bike_rental
load_OnlineNewsPopularity
load_student_grades
Sum of gbr_mse
Sum of gbr_svr_mse
Sum of gbr_svr_wrapper_mse