Assessing the effectiveness of regression models in predicting popularity of online news articles

Arshad Ahmed and Konstantinos Stathoulopoulos, City University, 2015

Introduction

Journalism 2.0 gave to news organisations many tools with which they can expand their audience. Continuing this trend, modelling techniques could be used to understand what inherently makes news popular and could also gain insight into customer preferences and thus aid in greater content personalisation. [1], [2] We decided to apply a regression tree, ensemble regression tree and support machine with a Gaussian kernel on the Online News Popularity dataset from UCI ML Repository [2] in order to predict the popularity of the articles that were uploaded on Mashable website. It should be mentioned that the specific dataset is high dimensional with over 39000 observations and 61 features.

Hypothesis Statement

H0: The content of an article, its keywords, its digital media content, the earlier popularity of news referenced in the article and the day it was published affect and can predict its popularity.

H1: The parameters of H0 don't affect and predict the popularity of an article.

Initial Data Analysis and Work Flow

The visualisation of the initial variables revealed huge variations regarding the scale of the data, which necessitated the need for scaling. We decided to split the research using two different scaling methods in order to investigate the effect on the results. This is outlined in Fig1.

Training and evaluation methodology

- Testing conducted with 80/20 split of features and target data into random samples with 10 fold cross validation
- Optimum model for 3 methods used to calculate errors on test data shown in table 1.

Method Descriptions

Regression Trees are trees where their leaves predict a real number and not a class. The algorithm looks for splits that minimize the prediction squared error. [7]

+ Easily interpretable.
+ If some data is missing, we might not reach a leaf, but we can still make a prediction by averaging all the leaves in the sub-tree we do reach.
+ Overcome the limitations of fitting a global linear model to high dimensional data.
+ Fast predictions
- They can stop growing too early for example if there are variables which are not highly informative

Support Vector Machine Regression utilise the concept of a hyperplane to represent a decision boundary. In a SVM regression this translates to finding a functional form f(x) that can correctly predict new cases [5].

+ Good generalization performance [6]
- Limited by the kernel choices

Ensemble Regression (Random Forest): Ensemble methods make use of the principle that a combination of machine learning models generally yields better results. Random forests form a family of methods that consist in building an ensemble of decision trees grown from a randomized variant of the tree induction algorithm. [8]

+ Robust to outliers. [4]Handles mixed type of data (categorical and numerical) through "Bagging".
+ Handles missing values.
+ Deals well with irrelevant values
+ "Bagging" constructs deep trees which is time consuming and memory intensive.
- Random Forests tend to be less accurate for regression than for classification. [4]

Analysis and evaluation of results

In using a high dimensional dataset our initial expectation was that we would have some strong predictors. We applied dimensionality reduction methods with the following results:

- **PCA**: Less than 60% of the variance in the data was explained by 10 components and we concluded that the result wasn't enough to justify the reduction.
- **Relieff:** Used for feature ranking and selection. The algorithm was applied separately to categorical and numerical variables but the results were unstable.
- **Sequentialfs**: Applied with two different criterion functions, KNN and Logistic regression, but both returned only one column as useful.
- **Stepwisefit**: Used for feature selection and returned a subset of 20 variables which we could use for our machine learning models.

The analysis was split between the different scaling methods and the above subset was as an input to our machine learning models. Using the scaled feature data we trained 3 models and used them to predict target values based on the test data and calculated Mean Square Error, Mean Absolute Error, Median Absolute Error, Error Variance, R Square Coefficient of fit, Correlation between prediction and actual values, RMS Error as well as the cross validation error .

We observed that overall the Random Forest Model performed better over the variety of metrics quoted, with the SVM results being very close to it.

Lessons learned and future work

Feature extraction from this dataset was a major challenge and future work could explore the applicability of different feature selection schemes using LASSO, and Factor Analysis on this dataset and others and evaluate their impact on the final result.

References:

[1] E. Hensinger, I. Flaounas, and N. Cristianini, "Modelling and Predicting News Popularity," *Pattern Anal. Appl.*, vol. 16, no. 4, pp. 623–635, Nov. 2013.

[2] P. V. and P. C. K. Fernandes, "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News," *Proc. 17th EPIA*, 2015.

[3] S. Cosma and C. R. Shalizi, "Lecture 10□: Regression Trees," *October*, pp. 1–7, 2006.

[4] L. Breiman, "Random Forests," *J. Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[5] "Support Vector Machines (SVM)." [Online]. Available: http://www.statsoft.com/Textbook/Support-Vector-Machines. [Accessed: 15-Nov-2015].

[6] "Support Vector Machines: Disadvantages." [Online]. Available: http://svms.org/disadvantages.html. [Accessed: 15-Nov-2015].

[7] L. Rokach, O. Maimon, "Data mining and knowledge discovery handbook", pp. 166-182, 2005

[8] G. Louppe, "Understanding Random Forests", 2014

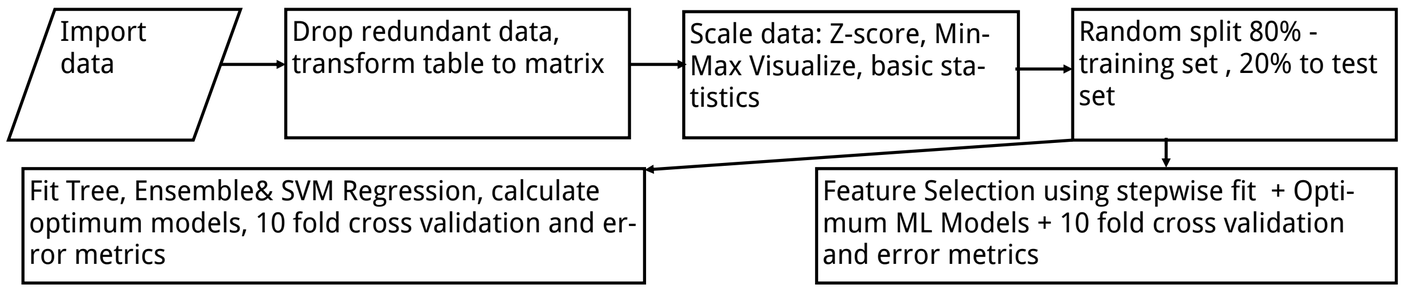

Fig 1: Shows our approach to data loading, pre-processing, model building and evaluation

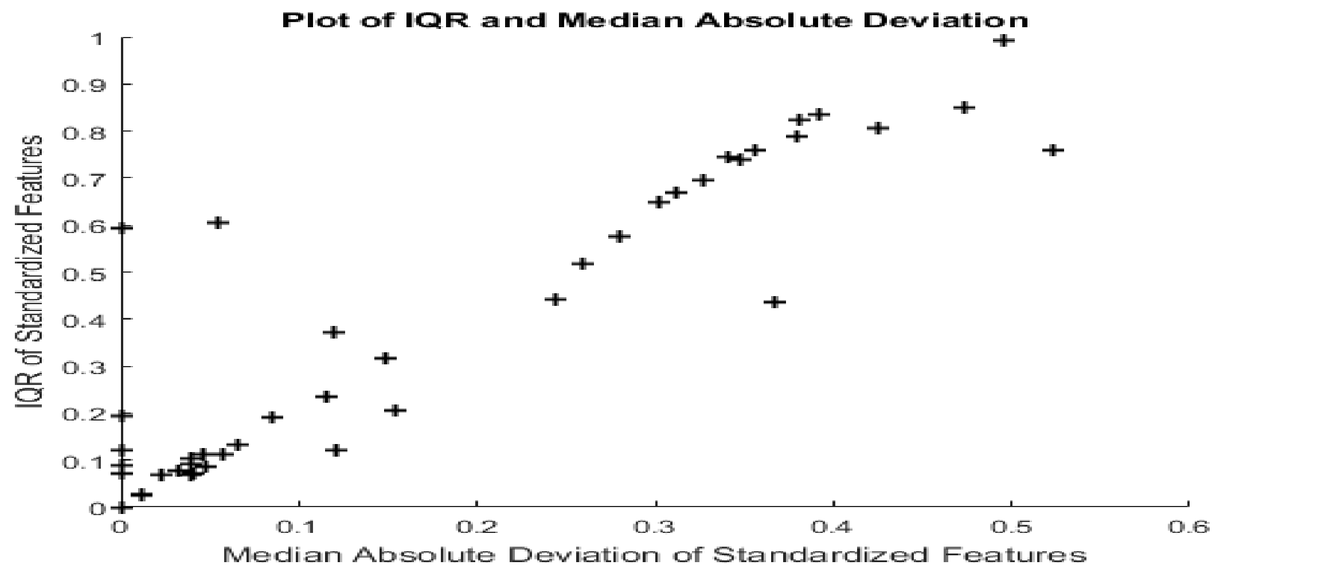

Fig 2: Interquartile Range and Median Absolute Deviation Plot of features. This plot is used as the IQR and MAD reduce each dimension into a point and is easier to visualize. This shows most features are not useful and some have high variance.

	Z-Score Standardization						Min-Max Scaling			
	Rtree	**RF**	**SVM**	**Rtree + FS**	**RF + FS**	**SVM + FS**	**Rtree**	**RF**	**Rtree + FS**	**RF + FS**
CV Error	1.098	1.068	1.074	0.948	0.968	0.949	0.0002	0.0004	0.0002	0.0003
Resub Error	0.658	1.039	1.068	0.919	0.566	0.945	0.0001	0.0001	0.0001	0.0001
MSE	0.395	0.371	0.368	0.858	0.888	0.866	0.0002	0.0002	0.0002	0.0004
Mean ABS Error	0.210	0.201	0.143	0.208	0.219	0.156	0.0038	0.0038	0.0038	0.0040
Median ABS Error	0.077	0.079	0.041	0.077	0.076	0.040	0.0018	0.0018	0.0018	0.0019
Error Variance	0.395	0.371	0.362	0.858	0.888	0.858	0.0002	0.0002	0.0002	0.0004
R Squared	0.141	0.084	0.148	0.082	0.049	0.073	0.0729	-0.2355	0.0823	-1.3069
Correlation	0.081	0.081	0.027	0.076	0.048	-0.011	0.1006	0.0230	0.1157	0.0154
RMS Error	0.609	0.629	0.607	0.926	0.942	0.931	0.0134	0.0155	0.0133	0.0211

Table 1: Shows all the error metrics calculated on test data and used for model evaluation

	Zscore Standardisation	Min Max Scaling without Feature Selection	Min Max scaling with Feature Selection
Regression Tree	Calculated Cross validation error with increasing leaf size of 200 in a logarithmic space	Model fitting. Categorical attributes had to be specified in the model	Model fitting. Categorical attributes of the selected subset had to be specified in the model
	Used a value of minimum leaf size of 180 based on this graph to create a new Regression Tree	Calculated cross validation error and found the optimal pruning point. Pruned the regression tree to level 3450	Calculated cross validation error and found the optimal pruning point. Pruned the regression tree to level 3375
Ensemble Regression	"Bagging" with 200 trees	"Bagging" with 200 trees. Categorical attributes had to be specified in the model	"Bagging" with 200 trees. Categorical attributes of the selected subset had to be specified in the model
	Calculated Cross validation, regression loss and out of bag error for the ensemble over the range up to 200	Calculated Cross validation, regression loss and out of bag error for the ensemble over the range up to 200	Calculated Cross validation, regression loss and out of bag error for the ensemble over the range up to 200
	The plot showed that convergence is achieved around 20. A compact model with minimized errors was created	The plot showed that convergence is achieved around 40. A compact model with minimized errors was created	The plot showed that convergence is achieved around 25. A compact model with minimized errors was created
Regression SVM	Specified the kernel function only and tested Gaussian and RBF functions		
	Experiments were made on both Gaussian and RBF kernels and found that they had near identical performance but the Gaussian kernel converged faster		

Table 2: This table shows initial starting parametrisation of the respective models and their method of optimisation also the final values chosen to derive the results on test data.