CS 7280 - Project Report
Analysis and Prediction on Online News Popularity
Group: G6

1. Introduction

In recent years, popularity of social media has increased drastically. More and more people are joining social media platforms to share their views, likes, dislikes, interests and many other things with their family, friends and colleagues. Due to this, tons of news, stories, articles, images and videos are being shared on social media every day. Recent studies show that social media has become an epicenter of online news distribution and consumption. More than half of the social media site users have shared news stories, images or videos and nearly as many have discussed the news on social media [1]. As a result, there is an increased interest in identifying the news articles that will receive a significant amount of user attention.

In this project, we build linear regression models to try to explain the behavior of social media users and predict the number of shares of news articles based upon several features extracted from the news articles. The Mashable news article dataset from UCI Machine Learning Repository is used in the analysis [2]. We explore the relations in the dataset, build and compare stepwise regression and regularized linear regression models.

2. Data Analysis

The Online News Popularity Data Set contains 58 predictors, 2 non-predictors and a response variable, which is number of shares of an online news. [Appendix Table-1] The 2 non-predictors are the unique URL and number of days between the article publication and the dataset acquisition. The predictors contain information extracted from news article, such as number of images, number of videos, title polarity, number of words in the content, rate of positive/negative words in the content etc. 3 predictors are categorical: type of article (data channel), day of week when the article was published and whether the article was published on weekend.

2.1 Data Cleaning

There are a total of 39797 instances in the dataset, which are uniquely identified with URL. The dataset was split into training and test sets, with an 80/20 ratio. All the 58 predictors and the response variable were explored for the detection of missing values, outliers and collinearity between predictors. The dataset offers a description for each of the predictors and some properties can be inferred from that information; based on it, the values in the predictor were also checked for consistency.

Online news uses different media, like videos and images, but a lot of the predictors are related to text content. Some data points don't have any text and a reduced number does not have any text, video or image. All those instances present 0 values for a high number of the predictors and they were removed from the scope of this analysis; they represent near 3% of the train dataset.

Some inconsistencies in the predictors were found during analysis:

- Rate of unique tokens higher than 1.
- Predictor (Minimum number of shares for a worst keyword) contains -1 value for more than 50% of the instances.

· Rate of nonstop words in the content become constant predictor after removing instances which doesn't have text in it.
All the instances with those conditions were also removed from the data set.

The five predictors which represent the closeness of a news article to five topics using Latent Dirichlet Allocation (LDA), have a high number of outliers on the positive side. Along with it, the predictors which represent the number of minimum, average and maximum shares of a worst, average and best keyword in a news article have wide spread. In both the cases, to reduce the influence of outliers, the log transformation is applied.

A closer look at the distribution of the response variable shows that it is far from normal. Number of shares for news articles goes from 1 to 843300, with a mean around 3300. This can be a problem for linear regression models, which have the assumption of a normally distributed error term. The Box-Cox transformation was applied to ensure that the distribution of the errors in the regression model follow a distribution close to normal. To evaluate the impact of the outliers in the model, we decided to perform two separate studies, with and without outliers. To remove the outliers, cook's distance is used. Cook's distance is really a good estimator as it can detect the data points which have large residuals and high leverages.

The distributions before and after applying the transformation can be observed on the following plot.

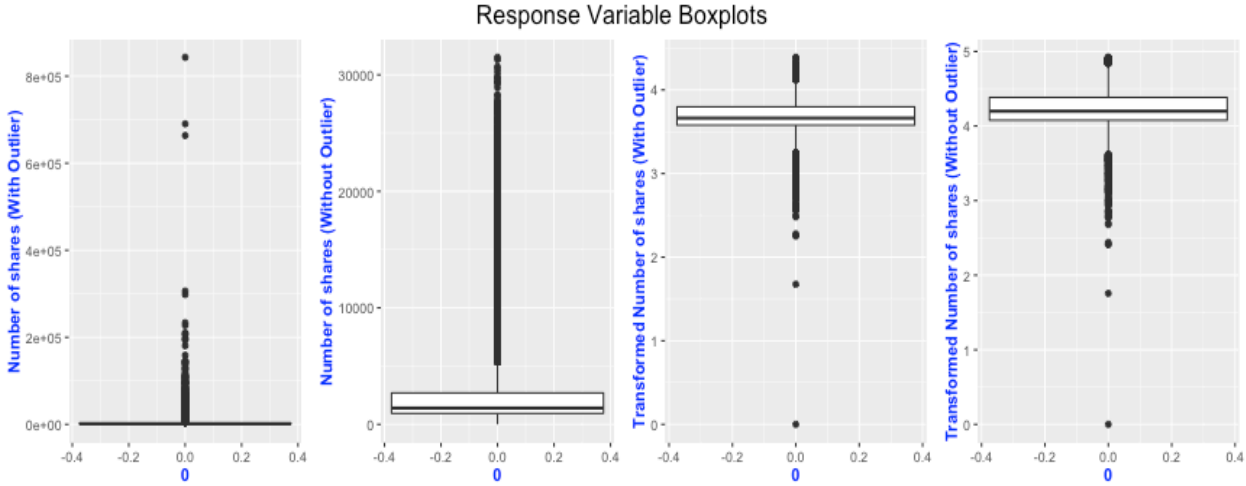

Figure 1: Response Variable Boxplots for the two train dataset with/without Box-Cox transformation

2.2 Multi-collinearity

Multi-collinearity is a serious problem as it can increase the variance in the coefficient estimation of a predictor and can cause the coefficient to change sign. To handle multi-collinearity, the pairwise Pearson correlation matrix of the continuous predictors is analyzed. [Appendix Figure 1]

The list of predictors which have high correlation, along with the approach to handle them is presented below:

1. *n_non_stop_unique_tokens* and *n_unique_tokens* have correlation of 0.887.

Those two predictors represent similar information, so n_non_stop_unique_tokens is removed from the analysis as both the predictors are semantically similar.

2. *self_reference_avg_sharess* has high correlation of 0.967 with *self_reference_min_shares* and 0.994 with *self_reference_max_shares*

   Predictors *self_reference_min_shares* and *self_reference_max_shares* are removed.

3. *rate_positive_words* and *rate_negative_words* have correlation of -0.997.

   Predictor variable *rate_negative_words* are removed from the analysis.

4. *kw_min_avg* and *kw_min_max* have correlation of 0.986.

   Predictor variable *kw_min_max* was removed from the analysis.

5. *n_unique_tokens* and *n_tokens_content* have correlation of 0.751.
6. *title_subjectivity* and *abs_title_sentiment_polarity* have correlation of 0.71
7. *min_negative_polarity* and *avg_negative_polarity* have correlation of 0.719
8. *rate_positive_words* and *global_sentiment_polarity* have correlation of 0.779
9. *kw_max_min* and *kw_avg_min* have correlation of 0.901
10. *kw_max_avg* and *kw_avg_avg* have correlation of 0.899
11. *kw_avg_max* and *kw_max_max* have correlation of 0.913

    In items from 5 to 11, interaction terms were added by taking the mean of the predictors.

3. Model Building

3.1 Stepwise Regression

Stepwise Regression is a handy method for regression problem as it automatically builds a model by successively adding or removing the variables based upon the supplied criteria such as t-statistic, AIC or BIC. In this analysis, three stepwise regression models are implemented: Forward, Backward and Both Directional. In each of the stepwise regression model building process, 10-fold cross validation is used for variable selection. Using 10-fold cross validation, forward, backward and both directional stepwise regression models are trained on 10 different subsets of the dataset.

For variable selection, the models generated in each fold are examined for all of the three stepwise regressions process. First hand selected only those variables which are selected in all the fold models for all the three stepwise regressions. After that created a list of variables which are selected in at least one of the fold model but not in all the fold models and applied best subset selection method to select the best subset of the variables from that list using Mallow's cp criteria. For both the train datasets with and without outliers, initially 19 predictors are selected by all the fold models for all three stepwise regressions and 1 predictor variable is selected by best subset selection method from the remaining variables using Mallow's cp criteria. Note: Selected variables are different for both the train datasets.

3.1.1 Interaction Terms

An interaction occurs when a predictor has a different effect on the response variable based upon the different values of other predictor. As mentioned in the first section, in this dataset there are three categorical variables - type of article (data channel), day of week when the article was published and whether articles was published on weekend? Interaction of all the continuous variables with these three categorical variables are analyzed and found the following significant interaction terms based upon it. Using these interaction terms new models for both the train dataset are created considering best stepwise regression model as baseline.

1. *num_hrefs* and *data_channel_is_socmed*

   Number of shares are decreasing with the increase in the number of links for the news articles which are related to social media where as its reverse case for other categories
2. *num_imgs* and *data_channel_is_socmed*

   Number of shares are decreasing with the increase in the number of images for the news articles which are related to social media where as its reverse case for other categories
3. *num_imgs* and *is_weekend*

   Number of shares are drastically increasing with the increase in the number of images for the news article which are published on weekdays.
4. *global_subjectivity* and *data_channel_is_socmed*

   Number of shares are decreasing with the increase in text subjectivity for the news articles which are related to social media where as its reverse case for other categories.
5. *avg_positive_polarity* and *data_channel_is_socmed*

   Number of shares are decreasing with the increase in average polarity of positive words for the news article which are related to social media where as its reverse case for other categories.
6. *i_n_unique_tokens_content* and *data_channel_is_bus*

   Number of shares are drastically increasing with the increase in the rate of positive words in the content and text sentiment polarity for the news articles which are related to business compare to other categories of news article
7. *min_positive_polarity* and *data_channel_is_entertainment*

   Number of shares are increasing with the increase in minimum positive polarity for the news articles which are related to entertainment compare to other categories of news article
8. *n_tokens_title* and *weekday_is_Tuesday*

   Number of shares are increasing with the increase in number of token in the title for the news articles which are published on Tuesday compare to other days
9. *num_self_hrefs* and *is_weekend*

   Number of shares are increasing with the increase in number of links to other articles for the news article which are published on weekdays
10. *max_negative_polarity* and *is_weekend*

    Number of shares are increasing with the increase in maximum negative polarity in the news articles which are published on weekends.

3.2 Regularization

Regularization can be applied to linear models to penalize more complex models and reduce the variance in parameter estimation. There are two basic ways of doing regularization. LASSO adds a penalty proportional to the sum of absolute values of the coefficients and produces a sparse representation of the predictors. RIDGE adds a penalty proportional to the sum of squares of coefficients and shrinks the values. Elastic net combines the two forms of regularization, using an extra parameter to measure the weight of each penalty term.

The glmnet package[ref] was used for building regularized models. In order to find the best values for the parameters lambda (weight of the penalty term) and alpha (elastic-net mixing parameter), a grid search strategy was performed: a set of possible values was selected for the two parameters and all the possible combination between them were tested on a 10 fold cross-validation, using the RMSE measured on the validation sets. There was no significant difference between the models and we opted to use only one pure LASSO and one pure RIDGE models, with the best value of lambda found on cross-validation. The LASSO model selected with this method is keeping most of the predictors. <more>

4. Model Comparison

A baseline model was added to the set of models for evaluation; it simply predicts the mean number of shares in the training data. Once all the models described above were built, a 10-fold cross validation, using the same seed, was performed and the RMSE was measured for the validation set in each fold. At the end, there are 10 measures for each model and a 95% confidence interval was built using Student's distribution with 9 degrees of freedom. The models were run twice on the dataset with and without outliers and the results are shown above with error bars.

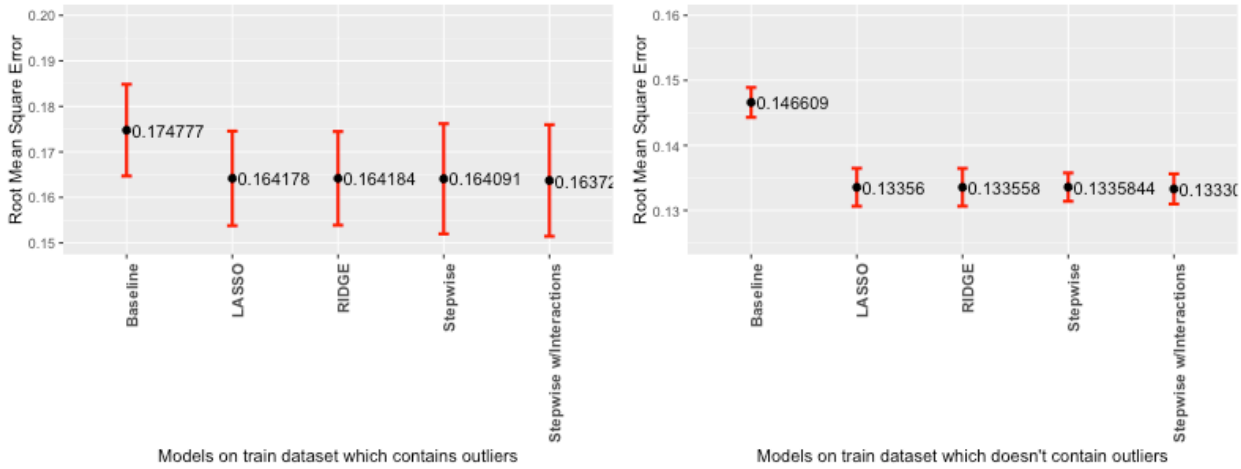

Figure 2: Various Model performance comparison on both the train data sets

Without removing outliers, the performance of the models is quite similar; there is no significant difference even when comparing to the baseline model. When outliers are removed, we see a better performance for the models, but there still isn't any significant difference between them (except for the baseline). Based on the results, we decided to use the model from stepwise feature selection, which is simpler and easier to understand as it has less predictors.

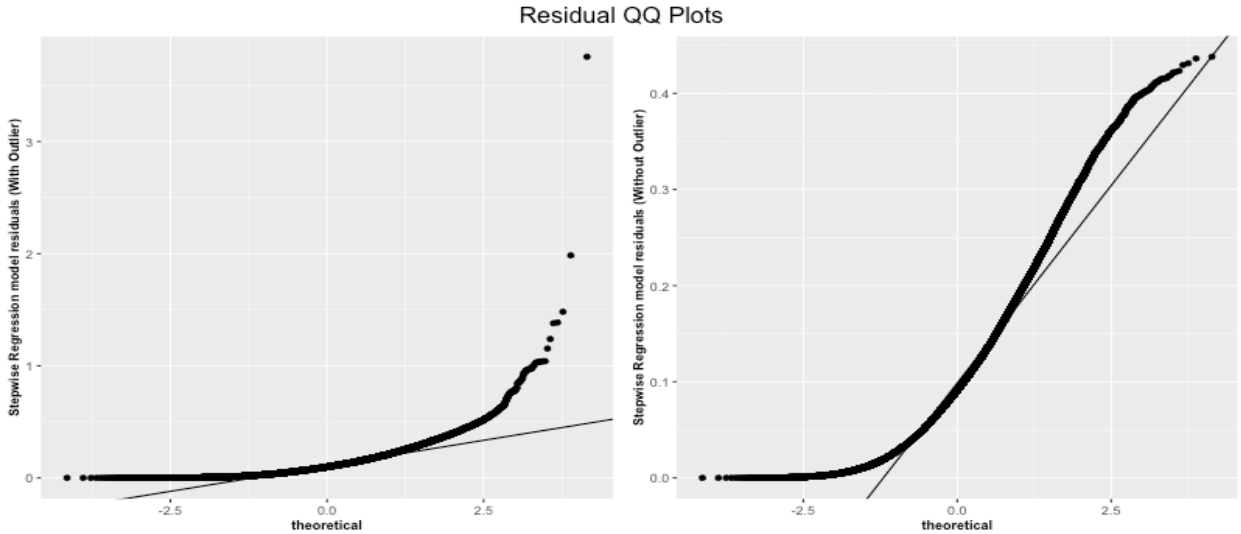

Figure 3: Various Model performance comparison on both the train data sets

4.1 Weighted Least Square

One of the assumptions for linear regression models is that the variance of the error term is constant over all values of the predictors variables. In our analysis, the variance has different value in different ranges of the predictor or response variables. Weighted regression can handle this situation as it weights the observations proportional to the reciprocal of the error variance of that observation to overcome the issue of non-constant variance. The plots below show that the variance in the residuals is increasing with increase in the predictor variable for both the train datasets.

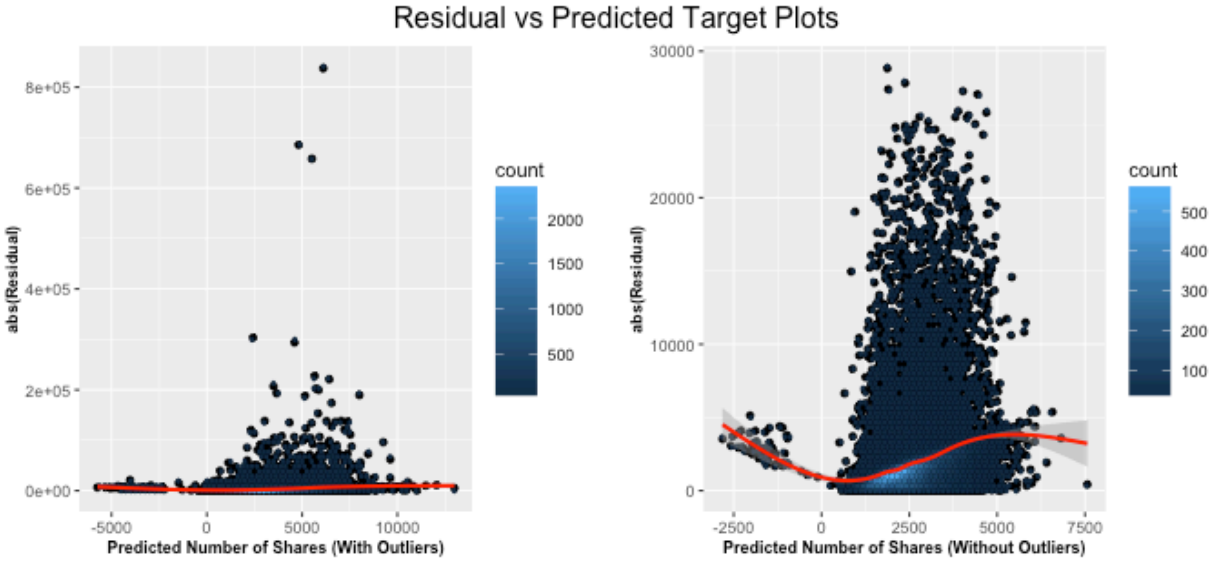

Figure 4: Residual vs Predicted response using full model on both the train datasets

To model the variance in the error term, a linear model is fitted on all the predictors of the dataset and an untransformed response. After that absolute residuals and predicted response values of that model are used to train a new linear model to predict the variance of each data point. Once the variances are known, they are converted into weights by taking inverse of squared variance for each data point. Once the weights of each data points are known two different approaches are tried to train a model using weights.

In the first approach, instead of applying the Box-Cox transformation on the response variable, weights are used to transform the response variable and performed normal stepwise regression using 10-fold cross validation. In the second approach, instead of applying weighted transformation on the full dataset, it was applied on train dataset of each fold of cross validation and left the validation dataset response variable untransformed. In the third approach weights are directly used to train a weighted least square model. Performance of first two approach is equal to the baseline but third approach turns out to be very effective.

4.2 Bootstrap

The bootstrap method [ref] can give the precision of the estimated coefficient values for a fitted regression model in complex cases, such as when the error variance is not constant. From the residuals plot showed previously, we concluded that this is the case with our models, so bootstrap was applied. Multiple samples are taken from the observed data, with replacement, using the same number of observations. For each sample, a model is fitted and the estimated

regression coefficients are calculated. Due to computational limitations, we used 300 bootstrap samples.

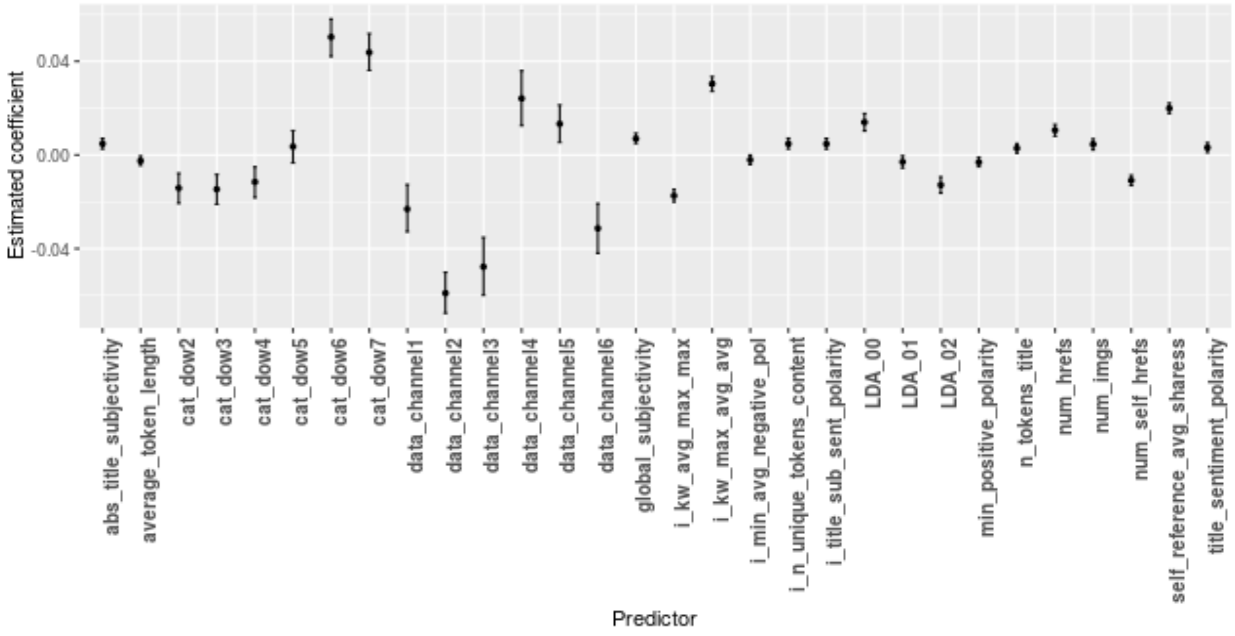

Figure 5: Predictor estimation using Bootstrap

It can be seen that only one predictor has a confidence interval that contains 0. It means that we don't have enough evidence to claim that it is significant to the model, but it can't be simply removed because it is part of the categorical variable "day of week". Overall, the categorical variables have a high influence in the model, which can be seen by the higher absolute values in the estimated coefficients.

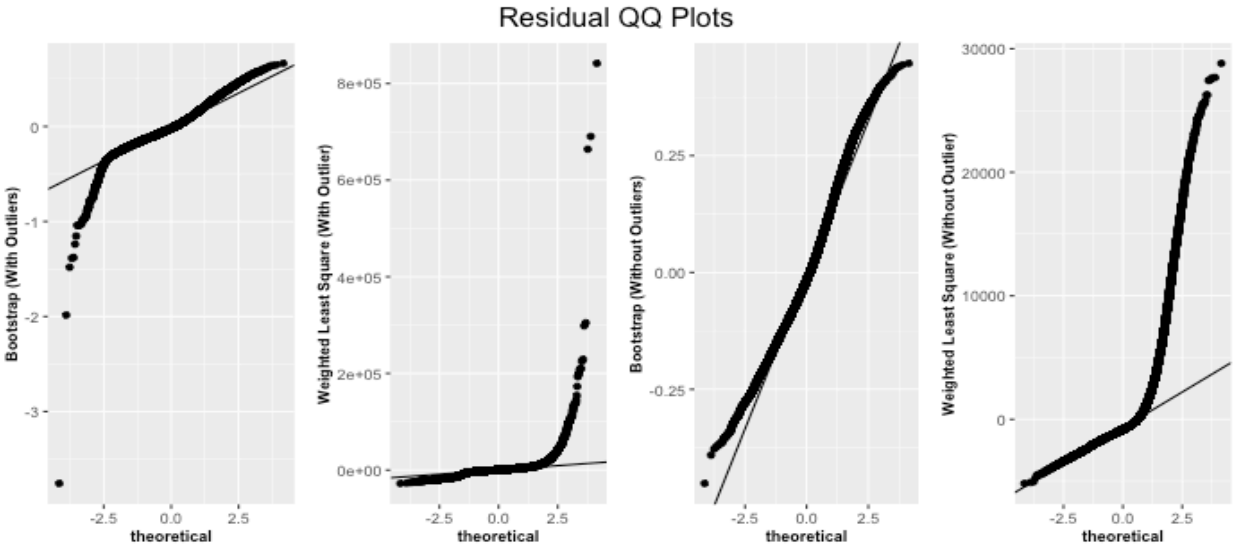

Figure 6: Residual QQ Plot for the Bootstrap and Weighted Least Square model on both the train datasets

5. Evaluation

The final model used for evaluation has the features selected in stepwise procedure and is bootstrapped; 300 models are trained on different bootstrap samples and the mean prediction of the models is used. The same preprocessing steps were performed on the test data and the predictions of the model are inversely transformed (from the Box-Cox transformation) to reflect the same magnitude of the original response variable. The error, measured with the root mean square error is presented when using the training data with and without outliers:

<results here>

6. Conclusions and Future Work

Predicting the number of shares for online news using linear regression models was a challenging problem. We tried several methods for fitting a model, but there was a negligible improvement in the predictions and overall the models performed badly. The predictions were in a limited range and couldn't accommodate the points with very high number of shares. The residual plots show that a small number of datapoints dominate the errors. Other aspects of online news have to be investigated to see if new features can be extracted to help predicting better those instances with very high number of shares.

The assumption that the variance of the error term is constant did not hold for the simpler regression models. A weighted regression with a linear model for the variance was proposed, but there was no way to compare its results with the other models and the errors still do not show normal distribution. A more complex fit for the variance could be tried in future work.

Another approach for this dataset is to turn the problem into classification. The prediction could be whether the online news articles will be very popular or not; there could also be different classes represented by different intervals in the number of shares. A logistic regression would be appropriate for that analysis.

7. References:

[1] http://kng.ht/1R5VqIE
[2] K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal

glmnet: https://cran.r-project.org/web/packages/glmnet/vignettes/glmnet_beta.html