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 intend to build a linear regression model which can mimic the behavior of social media users and predict the number of shares of a news article based upon the several statistical characteristics extracted from the news articles. We use Mashable news article dataset from UCI Machine Learning Repository.

To build the regression model, we intend to apply Stepwise Least Square Regression, Weighted Least Square Regression, Regularization and Bootstrap methods to learn the predictor variables and predict the number of shares of an article.

1. **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. 58 predictors contain the statistical characteristic of a news article such as number of image, number of video, title polarity, number of words in the content, rate of positive/negative words in the content etc. 2 non-predictors are URL of a news article and days between the article publication and the dataset acquisition. Among there 58 predictors 3 are categorical predictor - 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 total 39797 instances in the dataset, which are uniquely identified with URL and they were split into training and test data, with 80/20 ratio. All 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 analyzed for consistency.

Online news can use different media, like videos and images, but lot of the predictors are related to the text content, like number and polarity of words. Some of the instances in the dataset are from the news articles without any text and a limited number don’t have any text, video or image. Those instances have 0 values for an important set of the predictors and they were removed from the scope of this analysis. All the data points in the train and the test with this condition were removed from the dataset.

Some inconsistencies in the predictors were found during analysis:

* Rate of number 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 removed from the data set.

The five predictors which represent the closeness of a news article to the five different Latent Dirichlet Allocation (LDA), which are generative statistical models that allows sets of observations to be explained by unobserved groups/topics, have higher number of outlier on 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 of observation. In both the cases to reduce the influence of outlier log transformation has applied.

A wide spread was observed in the values of the response variable. Number of shares for news articles goes from 1 to 843300, with a mean around 3300. This can be serious problem in regression analysis, to deal with this outlier problem we decided to show two separate study, one with a data set which contains the outlier and one with a data set which doesn’t contain the outliers. To remove the outliers, cook’s distance is used and the instances which have cook’s distance greater than , where n is number of instances, were removed from the original train data set and a new train data set is created.

In the both train data set, a closer look at the distribution of target variable shows that it is far from normal, so the Box-Cox transformation was applied to ensure that the distribution of the errors in the regression model follow a distribution close to normal. The distribution before and after applying the transformation can be observed in the following plot.

**2.2 Multi-collinearity**

Multi-collinearity is a serious problem as it can increase the variance in the coefficient estimation and can cause the coefficient to change the sign. To handle the multi-collinearity, we have analyzed the covariance matrix of the continuous predictor variable. The graphical representation of covariance matrix is presented in Appendix x.

Below is a list of predictors which have high multi-collinearity along with the implemented approach to handle it.

1. n\_non\_stop\_unique\_tokens and n\_unique\_tokens has collinearity of 0.887

Fix: Predictor n\_non\_stop\_unique\_tokens is removed from the analysis as both the predictors are semantically similar.

1. self\_reference\_avg\_sharess has high collinearity with self\_reference\_min\_shares and self\_reference\_max\_shares

Fix: After trying all the combination of these three highly correlated predictors, collinearity was not removed so self\_reference\_min\_shares and self\_reference\_max\_shares predictors are removed from the analysis.

1. rate\_positive\_words and rate\_negative\_words has collinearity of -0.997

Fix: Predictor rate\_negative\_words is removed from the analysis.

1. kw\_min\_avg and kw\_min\_max has collinearity of 0.986

Fix: Predictor kw\_min\_max is removed from the analysis

1. n\_unique\_tokens and n\_tokens\_content has collinearity of 0.751
2. title\_subjectivity and abs\_title\_sentiment\_polarity has collinearity of 0.71
3. min\_negative\_polarity and avg\_negative\_polarity has collinearity of 0.719
4. rate\_positive\_words and global\_sentiment\_polarity has collinearity of 0.779
5. kw\_max\_min and kw\_avg\_min has collinearity of 0.901
6. kw\_max\_avg and kw\_avg\_avg has collinearity of 0.899
7. kw\_avg\_max and kw\_max\_max has collinearity of 0.913

Fix for 5 to 11: Added both the predictors and then divided by 2 to create a new predictor.

**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, we have implemented three stepwise regression model: Forward, Backward and Both Directional. In each of the stepwise regression model building process we have used 10-fold cross validation for variable selection. Using 10-fold cross validation, we have trained forward, backward and both directional stepwise regression model on 10 different subsets the dataset.

We have examined the models which are generated in each fold for variable selection and selected those variables which are selected in all the fold models for all three stepwise regressions. Along with that, we have created a list of variables which are selected in at least one of the fold model but not in all the fold models for all three stepwise regressions. After that we have used 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 with and without outliers.

**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? We have explored the interaction of all the continuous variables with these three categorical variables and found the following significant interaction terms based upon the analyses. 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

1. 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

1. 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.

1. 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.

1. 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.

1. 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

1. 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

1. 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

1. num\_self\_hrefs and is\_weekend

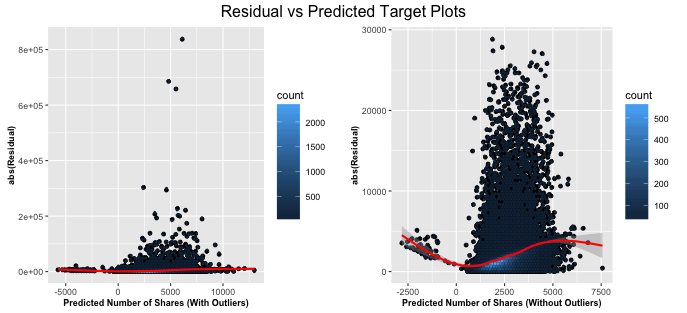
Number of shares are increasing with the increase in number of links to other articles for the news article which are published on weekdays

1. 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 Weighted Least Square**

One of the assumption about linear regression modeling process is that the standard deviation of the error term is constant overall value of the predictor variables. However, that’s not always the case. Sometimes the standard deviation of the error term increases or decreases with change in the predictor or response variables. Weighted least square 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 mentioned below shows the variance in the residuals is increasing with increase in the predictor variable for both the train dataset.



In this analysis we have tried various flavor of weighted regression. To learn the weights, we trained a linear model which includes all the predictors of the dataset and untransformed target variable. After that we used absolute residuals and predicted target values of that model to train a new linear model to learn the weights of each data point. Once the weights of each data points are known, we tried three approaches to train a model using weighted least square.

In the first approach, instead of applying the Box-Cox transformation on the target variable we transformed the target variable instances by their corresponding weights and performed stepwise regression using 10-fold cross validation. In the second approach, in each 10-fold cross validation iteration we only transformed the train dataset target variable instances by their corresponding weights and left target variable instances of the validation dataset unchanged. In the third approach we didn’t transform the target and used the weights of each data point to train a weighted least square model.

**3.3 Regularization**

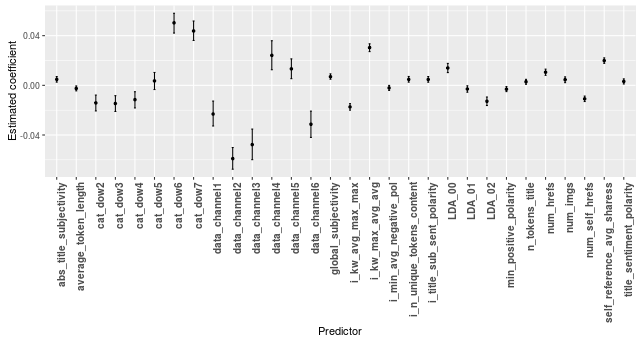
Regularization can be applied to linear models to reduce the variance in parameter estimation. There are two basic ways of doing regularization. LASSO adds a penalty proportional to the sum of absolute values of the coefficients and produces a sparse representation of the predictors. RIDGE adds a penalty 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 regularization. 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 are tested. The results were evaluated on a 10 fold cross-validation, using the RMSE measured on the validation sets.

An interesting plot that can be produced with glmnet showing the effect of the penalty term on variable selection for LASSO. It can be difficult to identify all the predictors in the plot, since there is a large number in the dataset, but we can see the most important ones, that have higher coefficient values and requires a higher penalty to be removed from the model. This picture is presented in Appendix x.

**3.4 Bootstrap**

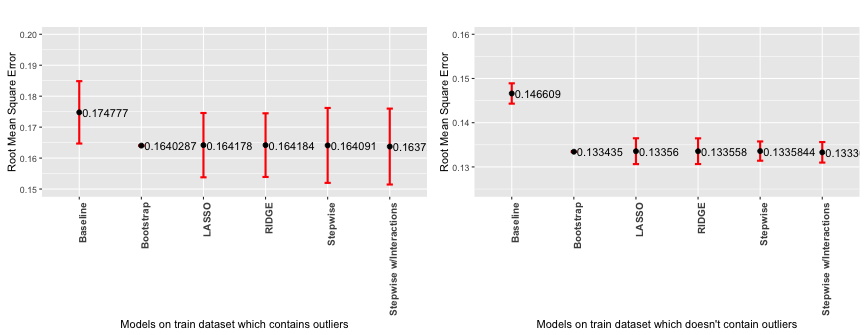
The bootstrap method [ref] can give the precision of the estimated coefficient values for a fitted regression model in complex cases, such as when the error variance is not constant. From the residuals plot showed previously, we concluded that this is the case with our models, so bootstrap was applied. 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.

****

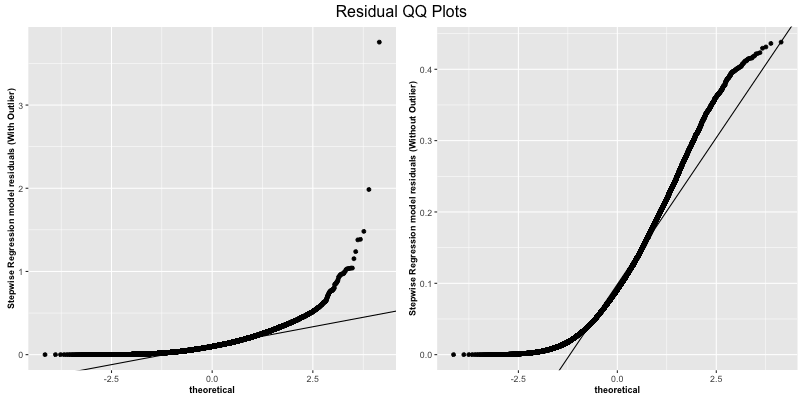
It can be seen that only one predictor has a confidence interval that goes through 0. It means that we don’t have enough evidence to claim that it is significant to the model, but it can’t be simply removed because it is part of the categorical variable “day of week”. Overall, the categorical variables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

1. **Model Comparison**

To select the final model, we have evaluated Stepwise regression, Weighted regression, Regularization and Bootstrap models using the 10-fold cross validation on both train datasets. As weighted regression falls under the case when there is no constant variance in the errors terms, we have compared it with the best model which assumes constant error variance. The following plot compares the Stepwise, regularization and bootstrap models. The mean and variance of RMSE is calculated based upon the value from 10-fold cross validation results.



From above plot, it can be seen that the Stepwise, Stepwise with Interaction, Regularization and Bootstrap; all models perform equally except slightly high RMSE variance in stepwise models. But due to simplicity, we have selected Stepwise regression model without Interaction model as best model in the categories of models which assume the constant variance of error terms. Though Stepwise is the best model but its residuals are not normally distributed so we can conclude that none of the above model is appropriate to predict the linear relationship between the number of shares and predictors.



**5. Conclusion and Future Work**