Blog Feedback Prediction

Rajal Nivargi

March 23, 2020

Contents

1 Abstract

This project aims to develop an automatic analysis method of blog posts by predicting the number of comments it is expected to receive. This also includes understanding possible insights from the predictions for social media marketing using blog posts.

2 Introduction

Today social media has been used extensively. Around 45% of the world population uses social media.[1] Previously, social media platforms like Facebook, YouTube, Instagram, etc. were meant for entertainment and connectivity. However, this has evolved over the years widening the scope and functionality of these websites/applications. These include marketing, publicity, customer reviews, etc.

A blog or weblog is a discussion or informational text entry published on the internet.The author of these weblogs can be a normal person, professional writer or a celebrity. It is freely written and accessible to all. 'Today's consumers are more intelligent and are utilizing platforms such as blogs, content sharing sites, wikipedia and social networking to create, converse and share Internet content.'[2]

The opinion in the form of feedback on such public sharing platforms can affect an establishment's reputation or sales. 54% of the social browsers use social media to research products and 49% depend on influencer recommendations. [1] The number of comments on a blogpost can be an indication for the popularity of the post. This may depend on the keywords used in the title which ensures better search-ability, time the post was published, the day the post was published, etc. Since this data is available as a part of the dataset, it will be possible to understand and relate these attributes to analyse many aspects. This can aid social media marketing. Thus, analysis of social media content becomes important. Manual analysis of the massive amount of content the internet produces is practically impossible. Therefore, there is a need of an automatic analysis method of such documents.

2.1 Goals

1. Analyse the multivariate data set
2. Understand and implement state of the art regression models
3. Achieve better performance than the models previously used
4. Explore deeper insights from the prediction and data attributes

3 Related Work

3.1 Related Work on the data set area

A number of studies use machine learning models for social media analysis. Asur and Huberman used tweets from the popular website Twitter for constructing a linear regression model to predict box office revenues of movies in advance of their releases. [3] Another study shows how the emotions of the users can be predicted by the comments on a Facebook post. This was done using CNN and RNN/LSTM models (for prediction).[4]

This project is based on the paper by Krishtian Buza [5].It predicted the number of expected feedbacks on a blog using a different techniques such as multi-layer perceptron (MLP), support vector machine (SVM), RBF-networks, regression trees (REP-tree, M5P-tree), nearest neighbor models, multivariate linear regression, and bagging approaches.The study by Uddin shows that a robust Ada-Boost approach on Buza's dataset give a decent prediction rate using discriminative and relevant blog content features.[6] **The baseline considered is the project by Sotiris Baratsas[7]. "The Random Forest Algorithm produced the best results, with 23.2699. The Gradient Boosting algorithm produced equally good results, with RMSE = 23.7586."**Both the models along with others will be implemented and comparative analysis will be drawn. Another work closely related to this field was by Yano and Smith[8]. They applied Naive Bayes, linear and elastic regression, and topic-Poisson models to predict the amount of feedback received by a political blog posts.

3.2 Related work on the Pattern Recognition Approach

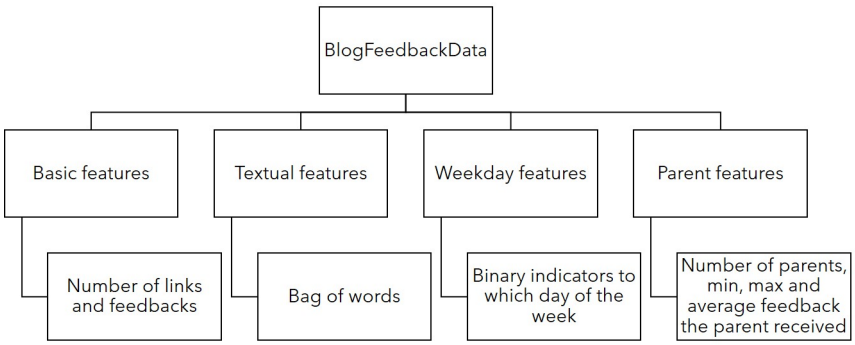
Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. 'Wu et al. [9] compared RF with linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-nearest neighbor (KNN) classifier, bagging and boosting classification trees, and support vector machine (SVM) for separating early stage ovarian cancer samples from normal tissue samples based on mass spectrometry data. RF outperformed the other methods in terms of prediction error rate.'[10]Therefore, it seems like random forests is a promising model for prediction in context of this project.

Gradient Boosting Machines are a collection of weak learners ensemble into one model to create a strong learner. Gradient Boosting Machines can be used for both regression or classification tasks. In a study by Alonso,Torres, and Dorronsoro [11], a comparative analysis was drawn between the random forests regression and gradient boosting regression models on wind energy prediction. The gradient boosting method was shown to be slightly better than random forests in some cases. It will interesting to understand the behavior of these ensemble models on the blog data set and observe their performance.

4 Data

The data is from the UCI machine learning repository. It is provided by Professor Kristian Buza at the Budapest University of Technology and Economics [5].The data set is multivariate data with Integer/Real attribute characteristics. It has 60021 instances and 281 attributes.The data set satisfies the rule of 10 and is sufficient for a regression task. The last column of training data is the dependent variable that is the number of comments in the next 24 hours.

Figure 1: Features in the dataset



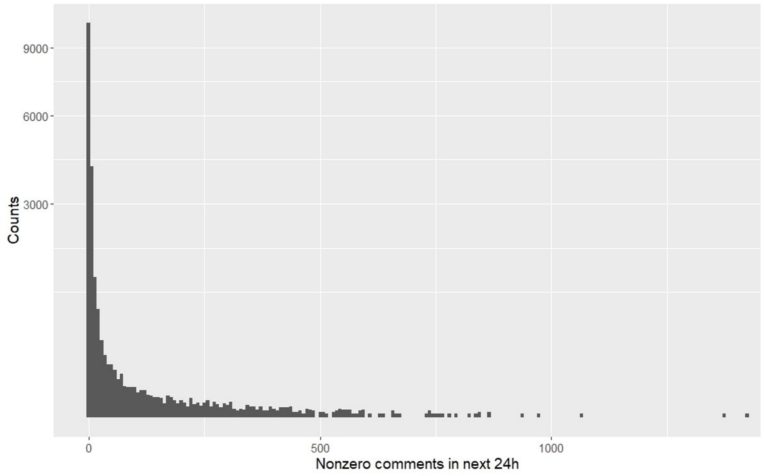
The data is obtained by crawling Hungarian blog sites and processing 37279 raw HTML document pages from around 1200 blog posts. The basetime(in the past) of selection of these documents was at most 72 hours before the selected base date/time. In the train data, the base times were in the years 2010 and 2011. In the test data the base times were in February and March 2012. This simulates the real-world situation in which training data from the past is available to predict events in the future.

The train data has 52397 observations, and the test data has 7624 observations. The dependent variable is highly skewed. 64.05% of the dependent variables in train data are zero, and among the non-zero ones, about 75% of them are less than 10, however, the dependent variable can also be as high as 1424. This can be seen in the figure 2.

Table 1: Basic information about training and validation sets[6]

	Training set	Validation set
Total crawled blog post or instances	52397	7624
Total extracted blog content features	280	280
Total Basic features	62	62
Total Textual features	200	200
Total Weekday features	14	14
Total Parent features	4	4

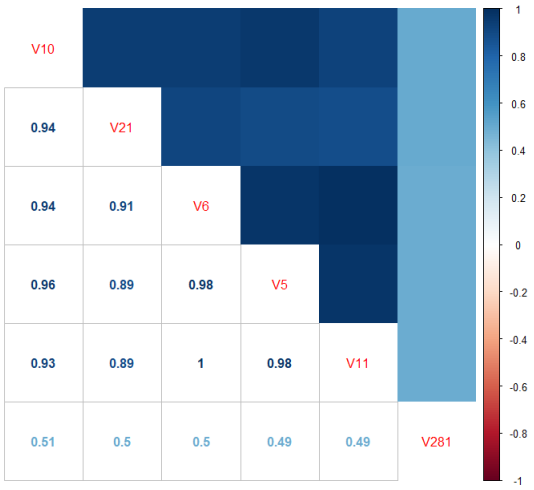
Figure 2: Data distribution



5 Data analysis

The different features are shown in the chart 1. Using correlation, it can be determined that the 10th, 21st, 6th, 5th, 11th independent features have the highest correlation with the target values. Those features are medians or averages of the comments from the source in different time periods. However, many of the first 60 features are highly correlated with each other. Feature selection ,thus, becomes an important part of the methods that are suggested.

Figure 3: Correlation map



Understanding the social media aspects of the data for the prediction of the number of feedback a blogpost is expected to receive is important to understand as a part of feature selection. We could choose features based on the features observed in the correlation map in figure 3. However,the correlation between different feature sets is also possible area of interest. Since the correlation map consists of the basic and parent feature sets, the analysis of the two other sets is observed in the following sections.

5.1 Analysis of weekday features

The weekday features are recorded in the columns 270-276 as binary indicators of the day the blog post was published. It is believed that the best day to post a blog is during the weekday such as Monday and Tuesday. People usually tend to invest their time on blogs after the weekend. To relate this to the data available, the frequency chart of the days of the week and the number of blogs published can be seen in figure 4[1]. The blogs were posted most frequently during the weekdays. However, the correlation between the dependent variable i.e. the number of comments and the weekday features is not high. The features may not useful for the purpose of machine learning using the suggested regression models.

5.2 Analysis of the textual features

The textual features are recorded in columns 63 to 262 as bag of words features for 200 most frequent words. The frequency chart of the words is seen in figure 6. As seen in the figure, some words have higher frequency than most. These could possibly be stop words. The next step was to observe the relationship of the textual features with the

[1]The days are denoted as 1=Monday to 7=Sunday

number of comments. One way of deriving that relationship is to calculate the conditional probability of the number of comments being greater than zero given the particular word of the 200 columns is present in the blog. This is shown in the figure 7. The probability was observed to be around 0.4 fairly uniformly for all the 200 words. This suggests that these features may not be very useful to predict the number of comments due to low dependence of the variable.

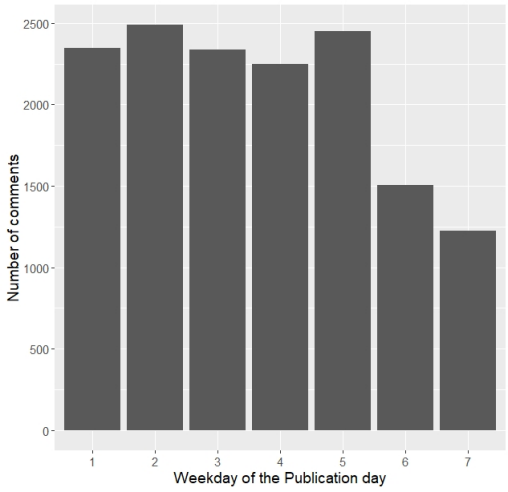


Figure 4: Frequency of blogs published on the days of the week

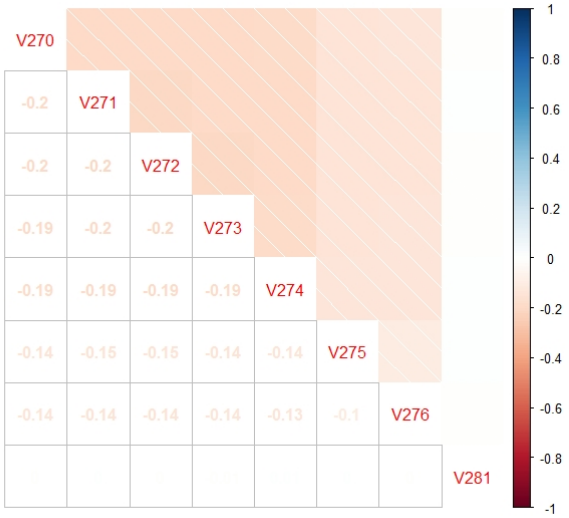


Figure 5: Correlation of dependent variable with the weekday features

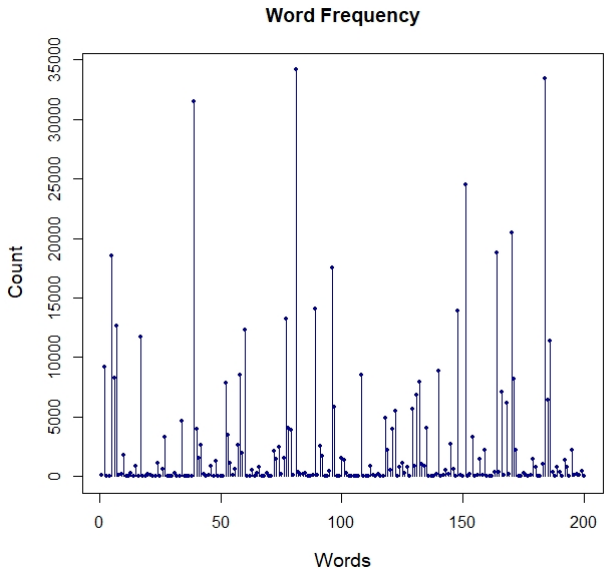


Figure 6: Word frequency

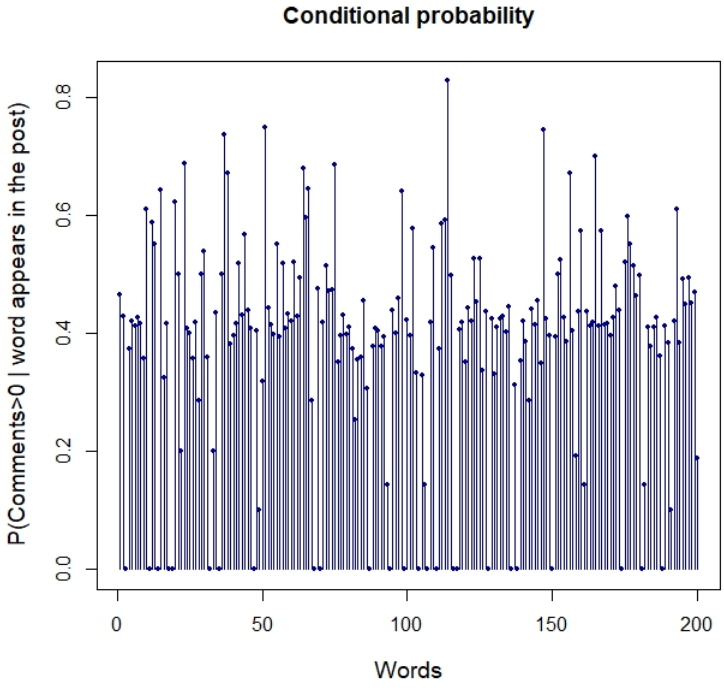


Figure 7: Conditional probability

6 Methods

In this project,five different regression models are proposed to predict the number of feedback on the blog posts in 24 hours after the basetime. There are three methods which are also a part of the original study by Buza[5]. The WEKA classifiers[2]:REP tree and M5 tree were shown to have better performance of all the models implemented in the initial study. The baseline for the project,however, uses Random Forests and Gradient boosting to show good performance as well. Since these were simple models using decision trees,Adaboost is another model suggested by this project in addition to baseline. The linear regression model will also be used for comparative analysis. Hence, the methods the project aims to implement are:

1. Linear Regression: Linear approach to modeling the relationship between a scalar response and one or more explanatory variables.

2. Random forests: Ensemble learning method that is constructed by training on a number of decision trees and predicting the mean of the individual trees.

3. Gradient boost: Produces a prediction in the form of an ensemble of weak prediction models in a stage-wise fashion along with optimization of an arbitrary differentiable loss function.

4. Ada boost: Subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. It works in conjunction with other type of machine learning algorithms to boost performance.

7 Quantitative Validation Method

The validation method this project will use is different o the ones adapted by the initial study. The evaluation methods were Hit@10 which consdiers 10 pages with highest predicted number of feedbacks, count how many received the same in reality and AUC@10 which considers the 10 pages with highest number of feedback in reality ,rank the pages according to their predicted number and area under curve. However, to keep the analysis conventionally understandable, this projects uses the Root mean squared error(also used in [7]) and R2 score majorly.

1. Root mean squared error: Measures the square root of the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

[2]WEKA - an open source software provides tools for data prepossessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems.

2. Mean absolute error: Risk metric corresponding to the expected value of the absolute error loss

3. Median absolute error: Median of all absolute differences between the target and the prediction.

4. Max error: Maximum residual error , a metric that captures the worst case error between the predicted value and the true value.

5. R2 score: Represents the proportion of variance (of y) that has been explained by the independent variables in the model. It provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model, through the proportion of explained variance.[12]

8 Approach

8.1 Fundamental method

In the first part of the project, regression was performed on the complete feature dataset using 4 regression models discussed above: Linear regression, Random forests, Gradient boost and Adaaptive Boost. The evaluation criteria for this purpose was root mean squared error, mean absolute error, median absolute error, max error and R2 score. The default models from Sci-kit learn were used for this implementation. [12]. The models were tested on 5 observation datasets collected at the interval of 15 days over 2 months. Along with that, a randomly permuted feature dataset was also used to observe any effect on the performance.

The performance of the 4 models were determined using the evaluation criteria in 7. The values were as shown in the table. The average and minimum value of error for all the 10 observations is recorded along with the maximum R2 score in the table. The bar graph of the score of the errors shows illustratively the comparison between the different models. From the outputs, it can be inferred that the Random Forests and Gradient boosting algorithms have the best overall performance. This brings a comparison between the two types of ensemble learning models used.

Code 1	Average values over 10 observations					
Regression models	Mean Absolute error	Median Absolute error	Mean Squared error	Max Error	R2 Score	RMSE
Linear	9.904423049	3.76093548	1045.368449	288.3494604	0.698226555	32.33215812
Random Forests	6.492058247	0.416666667	859.5358361	249.4	0.71573993	29.3178416
Gradient Boosting	6.485996368	0.706912889	994.0803989	305.2235143	0.557637721	31.52904056
AdaBoost	9.656520434	3.876278237	1116.026268	303.8274972	0.596202872	33.40697933

Figure 8: Output table: Average error and max R2 values

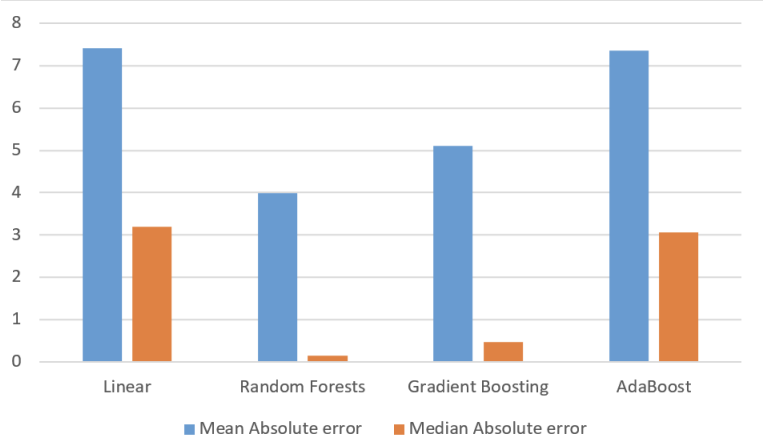


Figure 9: Min of mean absolute error and median absolute error

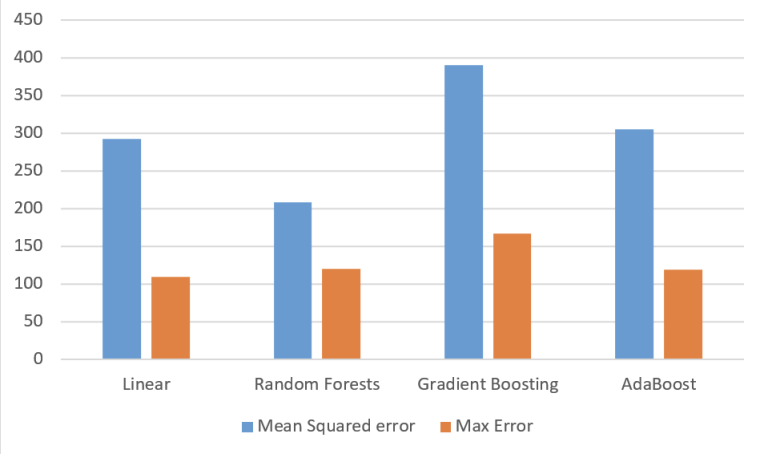
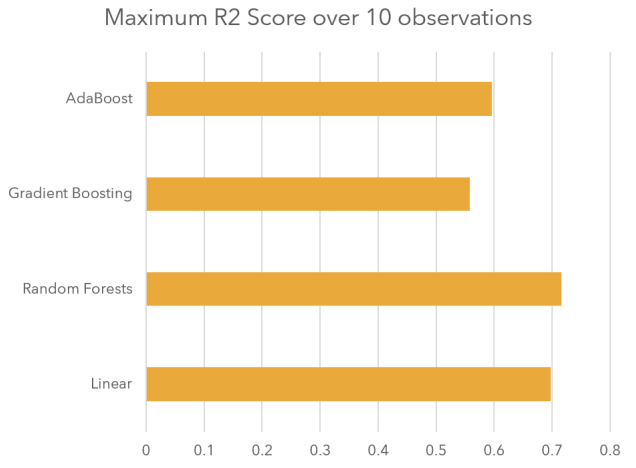
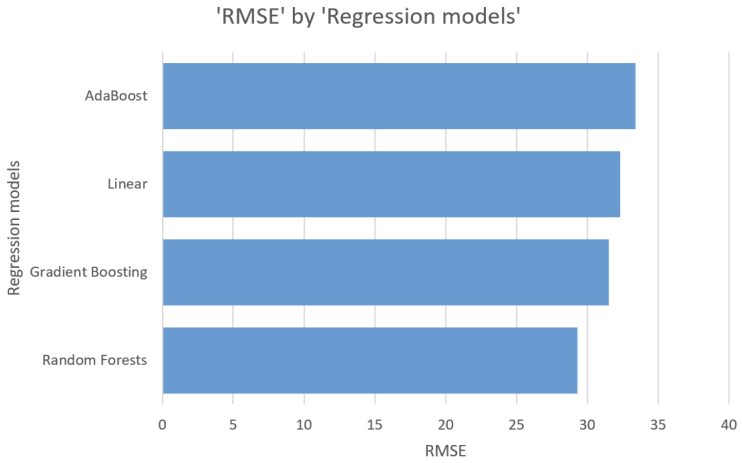


Figure 10: Min of mean squared error and max error





The performance of the random forests and gradient boost models are close but still greater than compared to the baseline which has values RMSE= 23.2699 and RMSE=23.7586 respectively. Further, the steps will focus on increasing the performance of the best performing model in this attempt. This can be done using feature selection based on state of the art algorithms and also only considering a certain group of features to observe of effect of each group on the predictions.

8.2 Feature selection

Feature selection is done by reducing the number of features that are used for the machine learning algorithm. It uses the most relevant features for the process. It is proposed that undertaking this approach will improve the performance of the predictions.

8.2.1 Filter

In the filter method, a subset of features are selected such that the value of criterion function is optimized. This can also result the algorithm to be computationally less expensive and fast. The criterion used for evaluation in this project is correlation of the dependent variable i.e the number of comments that the post is expected to receive in the next 24 hours with the independent variables or features. The features are ranked using a ranking subroutine such that the top 10% of the total features that have highest correlation values are selected.

As discussed in 5, the basic and parent feature sets have features highly correlated to the dependent variable. Thus, as expected, the features that were selected from the filter method were the average, standard deviation, min, max and median of the number of comments before recording the basetime of the post or of the source of the blog post. The models discussed in 6 were trained using only the top 10% of the total features(280) i.e. the highly correlated 28 features.

8.2.2 Wrapper

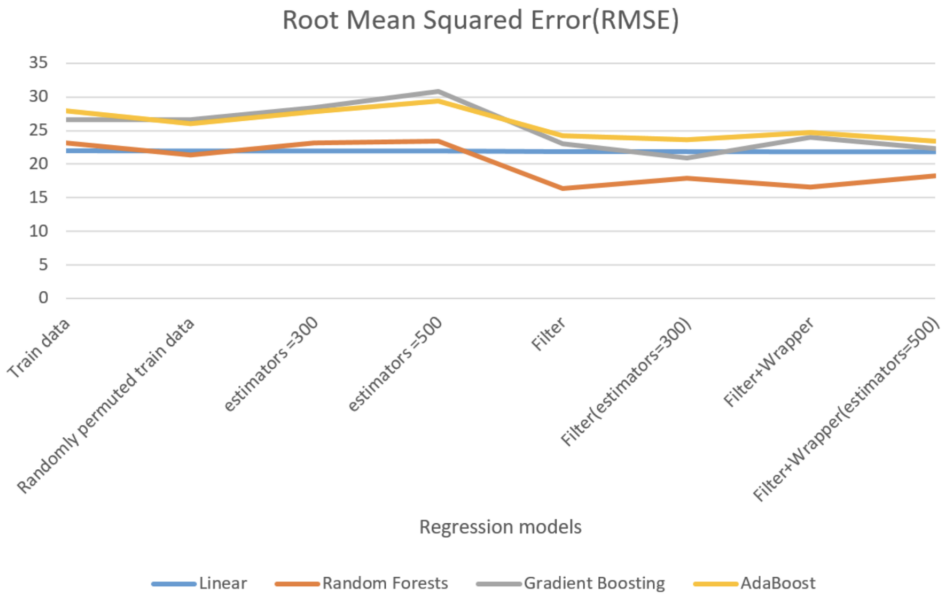
For the subset of features selected in the filter method, the wrapper selects on the basis of quantitative rates. This separates the class distribution. The search strategy adopted for the purpose of this project is Recursive Feature Elimination with Cross-Validation(RFECV)[12]. The features are assigned importance and ranked by a supervised learning estimator. The Random Forests Regressor [12] is used as estimator. The search criterion used by random forests for measuring the quality of the split is Mean squared Error(MSE). A 5-fold cross validation starergy is used in this method. The RFECV ,thus, ranks the features with recursive feature elimination and cross-validated selection of the best number of features by the random forests estimator on the basis of feature importance defined by the mean squared error.

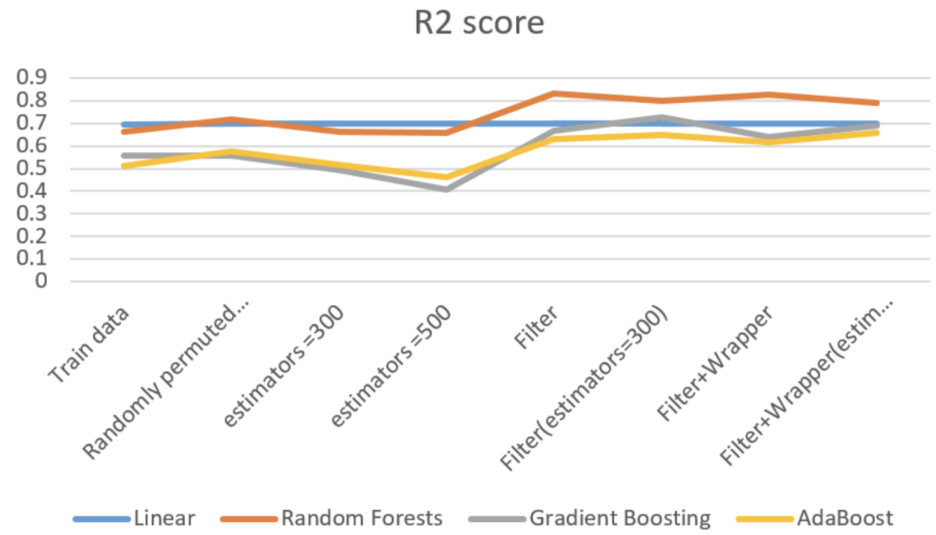
The 28 features obtained in the filter method were then recursively eliminated using the wrapper method. The number of features finally, were reduced to 9. Only a few of these features were parent features that are related to the source of the blog. Mostly, these were from the basic features which are the values of the past number of comments on the blog post.

9 Results

The evaluation results were recorded for the test data noted on the first of the month February i.e. the first of the test datasets available. For the final results, the evaluation criteria used were Root Mean Squared Error and R2 score. The RMSE or Root Mean Square Error is often used to determine the difference between the predicted value of the model or the estimator and the ground truth. R2 score is also called the coefficient of determination. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. The best possible value of R2 score is 1.0. It can also be negative if the model performs worse arbitrarily. Outputs from the fundamental method in 8.1 are shown to observe the difference between the initial approach and after feature selection.

The number of estimators or tress in the forests were by default 100. The random forests regressor seemed to perform worse when the number of trees were increased to 300 and 500. This may be due to overfitting of the model. Simialrly, the gradient boost model had 100 boosting stages by default. However, gradient boosting is fairly robust to overfitting. So, better performance was observed when the number of boosting stages were incresed to 300 and 500. The adaptive boost model had 50 number of estimators at which boosting is terminated. The model may be terminate earlier in case of a perfect fit. Similar to the gradient boost model, Adaboost saw a better performance after increasing the number of estimators. Thus, the ensemble learning models are seen to perform better with more number of boosting trees.





The line graphs for both the criteria show that the performance of the regression improved by using feature selection. The values of the root mean square error are seen to decrease while those of the R2 score are seen to increase.The linear regression model had a fairly consistent output with RMSE = 21.92339372 and R2 score= 0.699198485 on an average. The feature selection method saw no significant increase in its performance. The adaptive boost regressor ,however, performs better with less number of features after the filter and wrapper feature selection and more number of boosting stages i.e 500. The lowest RMSE was 23.39061415 and maximum R2 score of 0.657590011 for this method. Gradient boost perform better using the filter method with 300 boosting stages or estimators with RMSE=20.8880533 and R2 score=0.726939356. The Random forests regressor is seen to perform best according to both the criterion with the least RMSE value of 16.33829997 and R2 score maximum at 0.832938296 using the filter method with 100 estimators or decision trees. The RMSE value is lower than the baseline values of 23.2699

for Random forests and 23.7586 for gradient boosting which used the entire dataset with all features. Thus, the random forests regressor with 100 decision trees gives the best performance followed by gradient boosting with 300 estimators ,both using filter method.

10 Conclusion

Through this project, a better understanding of the regression models on a real world data set was achieved. These models were based on decision trees and ensemble learning. The relationship between using feature selection as well as number of estimators on the performance of these models was observed. All the models used were seen to have performed better after feature selection. The Gradient Boost and Adaboost regressors saw an improved performance with incresed number of estimators. However, the random forests regressor showed the best perfomance while using less estimators.

A study of the blog feedback data set helped understand the relationship of a prediction model in terms social media analysis. The project attempted to understand the possible effect of keywords used in the blog, day of publishing, initial comment thread for the post to be 'popular' and the source of the blog post. The aim was to understand the applications in social media marketing. The data analysis of the Weekday features shows that the most posts were published on weekdays which checks out with other studies on the same question. However, the correlation between these features and the number of comments was low ,so, the features were not used for prediction. The Textual features identified some words having higher frequency than others. However, the probability of the posts getting more comments based on the usage of a word was around 0.4 for all the words. Thus, these features were not used for the prediction process. The basic and parent features with the average, median, min and max values of the comments on the blog posts as well as the sources of the posts were selected in the filter method. This method gave the best performance in this project. The wrapper method selected fewer of the same features as well. Therefore, the popularity of the blog posts in term of number of feedback on the post in the post as well the popularity of its source are both important factors in determining the number of comments on the blog.

Further work The next step would be implementing a neural network for this data predicting the number of comments. The parent paper by Buza [5] used a Multilayer Perceptron for this purpose. So, a deep learning approach may be one of the possible ways to better predicts the number of comments.

11 Updated timetable

Table 2: A Simple weekly timetable for the project.

Week	Info
Week 1 (3/21-3/27):	Proposal
Week 2 (3/28–4/3):	Analysing the data set
Week 3 (4/4–4/10):	Implement regression models on different test sets
Week 4 (4/11–4/17):	Data analysis
Week 5 (4/18–4/25):	Feature selection
Week 6 (4/26–4/30):	Report and presentation

References

[1] M. Mohsin, *10 Social Media Statistics You Need to Know in 2020 [Infographic]*, 7 Feb, 2020. [Online]. Available: https://www.oberlo.com/blog/social-media-marketing-statistics 3

[2] M. Kaur and P. Verma, "Review on comment volume prediction," 2016. 3

[3] S. Asur and B. A. Huberman, "Predicting the future with social media," *CoRR*, vol. abs/1003.5699, 2010. [Online]. Available: http://arxiv.org/abs/1003.5699 4

[4] F. Krebs, B. Lubascher, T. Moers, P. Schaap, and G. Spanakis, "Social emotion mining techniques for facebook posts reaction prediction," *CoRR*, vol. abs/1712.03249, 2017. [Online]. Available: http://arxiv.org/abs/1712.03249 4

[5] K. Buza, *Feedback Prediction for Blogs*, 10 2014, pp. 145–152. 4, 5, 10, 16

[6] M. T. Uddin, "Automated blog feedback prediction with ada-boost classifier," 06 2015, pp. 1–5. 4, 6

[7] S. Baratsas, "Predicting blog comments with machine learning methods." [Online]. Available: https://www.baratsas.com/blog-comments-prediction 4, 10

[8] T. Yano, W. W. Cohen, and N. A. Smith, "Predicting response to political blog posts with topic models," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. NAACL '09. USA: Association for Computational Linguistics, 2009, p. 477–485. 4

[9] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 09 2003. [Online]. Available: https://doi.org/10.1093/bioinformatics/btg210 4

[10] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323 – 329, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0888754312000626 4

[11] Á. Alonso, A. Torres, and J. R. Dorronsoro, "Random forests and gradient boosting for wind energy prediction," in *Hybrid Artificial Intelligent Systems*, E. Onieva, I. Santos, E. Osaba, H. Quintián, and E. Corchado, Eds. Cham: Springer International Publishing, 2015, pp. 26–37. 4

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. 11, 14

[13] B. L. Peiran Cao, Zhoutao Pei, "Blogfeedbackproject." [Online]. Available: https://github.com/hncpr1992/BlogFeedBackProject