Yelp recruitment Kaggle competition: a first pass
Advanced Methods IV 140.754

1 Introduction

Yelp [11] is an internet company that provides users with information about businesses in their neighborhood. They are currently hosting a Kaggle competition and offering job interviews to the winner and other top players [12]. The goal of the competition is to predict the number of *useful* votes a review gets. Yelp provides competitors with a training data set with 229,907 reviews and a test data set of 22,957 reviews.

2 Methods

2.1 Data Processing

We initially read the data intoe `R` [8] using the packages `plyr` [10], `RJSONIO` [5], `reshape2` [9], and `doMC` [2] for faster computation times and decomposing the nested lists. We extracted the number of occurrences for the top 200 words from the text reviews using the `tau` [4] package through a process known as tokenizing in Natural Language Processing (NLP) vocabulary. Some of the punctuation characters might be influential in determining how helpful a review is; like the number of spaces compared to the number of lines. For the business address, we extracted the zipcode and discarded the detailed information. Overall, the zip code should help differentiate up and coming neighborhoods from declining ones.

We then proceeded to collapse the information from the four tables (user, business, checkin and review) into a single table where each row represented a unique review. For the purpose of this work, we retained only the top 50 words and discarded the detailed checkin information, although this could be used in more detailed analyses.

To allow for an internal comparison, we further separated at random the training data set into two groups: a validation set and the actual training set with probabilities 0.3 and 0.7 respectively for each review.

2.2 Finding models

Overall, the strategy we followed was to fit a model with the train data set and if possible, perform 10 fold cross-validation on each model. We then evaluated the performance of the model on the validation data set. These results were then used to select the final models to be used in the competition.

2.2.1 Linear regression

A multivariate linear regression was used with all the variables and from this model two other models were determined via step wise selection using AIC and BIC criteria respectively. The three models were then cross validated using the `cvTools` [1] package. We used 10 fold cross validation with 5 replicates.

2.2.2 Generalized Linear Model: Poisson

Similarly to what was done with linear regression, we constructed a model using Generalized Linear Models (GLM) with a Poisson family. Then two other models were selected for using AIC and BIC

criteria. Then, cross validation was performed as in the case of the linear regression model using `cvTools` [1].

2.2.3 Generalized Boosted Regression Model: Poisson

We fit a Generalized Boosted Regression (GBM) model using the Poisson distribution—since we are dealing with counts—using the `gbm` [7] package. We used 10 fold cross validation and chose to use 500 trees, although using more might lead to better results.

2.2.4 Random forest

We built a random forest for regression using the `randomForest` [6] package. Instead of performing cross validation, we sought to build a large forest (2000 trees) using the `foreach` [3] package and a larger number of cores. However, this computation is still ongoing.

3 Results

3.1 Models

3.1.1 Linear regression

The best resulting model, by a very minor margin was the linear regression using the AIC criterion when performing the model selection. Interestingly, all three models performed nearly identically on the validation data set when evaluating using the root mean squared prediction error (RMSPE). However, it is important to note that linear regression will yield predictions below 0, which are not possible. Nevertheless, after rounding all negative predictions to 0, the root mean squared log prediction error (RMSLPE)—which is the one used when evaluating the predictions in this competition—is 0.5907 with a standard error (SE) of 0.001635 in the validation data set. This would currently grant a top 100 (out of 147) spot in the leader board.

3.1.2 GLM: Poisson

Compared to the linear regression model, in this case there was more variation between the three models with the BIC model outperforming the other two. However, when using RMSLPE to evaluate their performance on the validation data set, BIC was the lowest with 0.6504 and SE of 0.03593, which is worse than the linear regression models. However, these results are not definitive as they were fit with a subset of the training data for computation time purposes.

3.1.3 GBM: Poisson

The GBM model performed very similarly to the GLM model when using the validation data set (RMSLPE of 0.6545 with SE of 0.001495). However, it did so by predicting useful votes between 1.2 and 1.78 for each review. That is, with much limited variability.

3.2 Random forest

A smaller random forest was fit on a very small sample of the data, and it's RMSPE is similar to the linear regression model one.

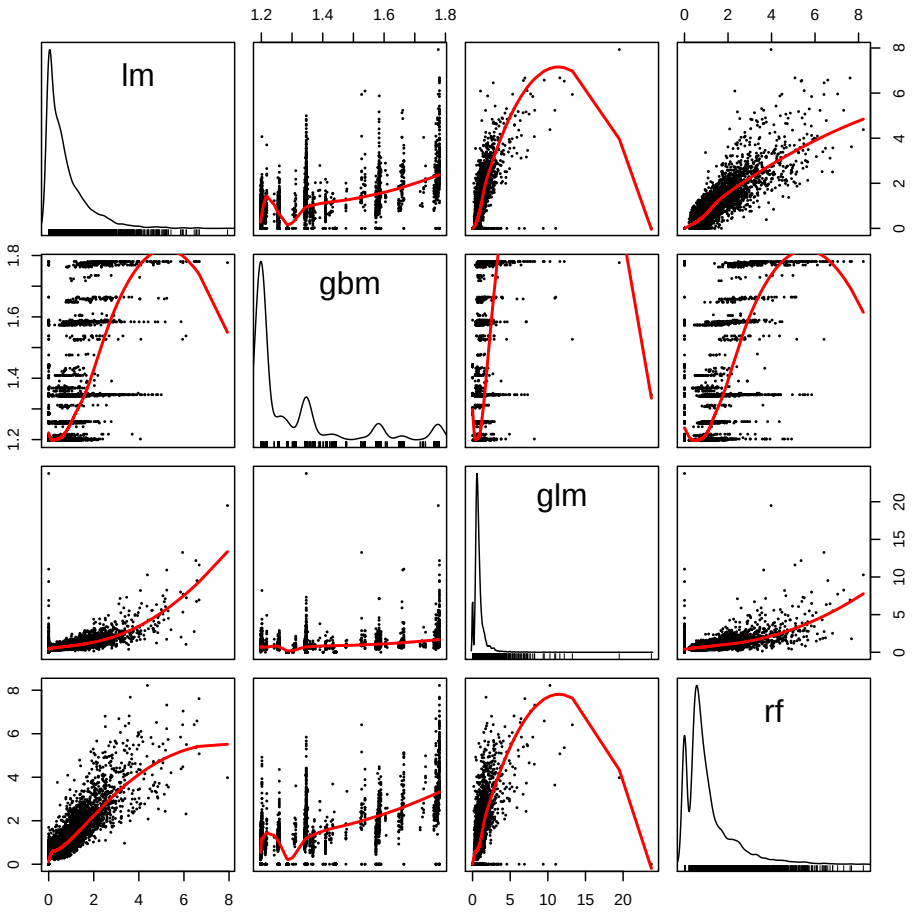

Figure 1: Predictions from the best linear model (lm), best glm, the gbm cross-validated and the random forest (rf) for the test data given by Yelp. Red lines represent lowess smoothers. Showing the first 5000 points.

3.3 Predictions

Using the linear model selected using AIC, the GLM model selected using BIC (*), the GBM and the random forest (*:= using subset of the training data), we calculated the predictions for the test data given by Yelp as shown in Figure 1. Note that for the linear regression, negative values were rounded to 0. Interestingly, each prediction method has a different relationship with one another, where the linear regression and the random forest are the closest to each other.

We thus used the mean of the predictions for each review and submitted this as our first entry to the Kaggle competition, resulting in a score of 0.56667 and placing in position 73 of the leader board. This is above the *all zeros benchmark* and the *global mean value benchmark*, although far from Muschelli's 0.50930 in rank 30.

4 Conclusions

Using the top 50 words, the total number of checkins, zip code and the information given from the user and business tables, we built a prediction algorithm for the Yelp Kaggle contest by combining the predictions from several models into one. While some of them have not completed running with the full data (glm and random forest), the preliminary results are promising by ranking in the top 50% of the leader board.

However, further analysis needs to be carried out to carefully identify if any other of the top words or the finer information of the checkin table can be used to improve the results. In addition, it is important to consider that the linear regression model in theory is not limited to the positive real line and can thus be working poorly. Finally, it will be important to consider the effect of rounding the predictions to integers. For instance, by rounding the predictions the score deteriorates from 0.56667 to 0.62584. In other words, the way Yelp is evaluating the competition promotes submitting non-rounded values which in a way are not realistic.

5 References and acknowledgements

Muschelli told me that you could use `foreach` when running `randomForest`.

The code for this project is available at `https://github.com/lcolladotor/lcollado754/tree/master/hw/hw2`.

References

[1] A. Alfons. *cvTools: Cross-validation tools for regression models*. R package version 0.3.2. 2012. URL: `http://CRAN.R-project.org/package=cvTools`.

[2] R. Analytics. *doMC: Foreach parallel adaptor for the multicore package*. R package version 1.3.0. 2013. URL: `http://CRAN.R-project.org/package=doMC`.

[3] R. Analytics. *foreach: Foreach looping construct for R*. R package version 1.4.0. 2012. URL: `http://CRAN.R-project.org/package=foreach`.

[4] C. Buchta et al. *tau: Text Analysis Utilities*. R package version 0.0-15. 2012. URL: `http://CRAN.R-project.org/package=tau`.

[5] D. T. Lang. *RJSONIO: Serialize R objects to JSON, JavaScript Object Notation*. R package version 1.0-3. 2013. URL: `http://CRAN.R-project.org/package=RJSONIO`.

[6] A. Liaw and M. Wiener. "Classification and Regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22. URL: http://CRAN.R-project.org/doc/Rnews/.

[7] G. R. with contributions from others. *gbm: Generalized Boosted Regression Models*. R package version 2.0-8. 2013. URL: http://CRAN.R-project.org/package=gbm.

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: http://www.R-project.org/.

[9] H. Wickham. "Reshaping Data with the reshape Package". In: *Journal of Statistical Software* 21.12 (2007), pp. 1–20. URL: http://www.jstatsoft.org/v21/i12/.

[10] H. Wickham. "The Split-Apply-Combine Strategy for Data Analysis". In: *Journal of Statistical Software* 40.1 (2011), pp. 1–29. URL: http://www.jstatsoft.org/v40/i01/.

[11] Yelp. *yelp*. URL: http://www.yelp.com/ (visited on 05/01/2013).

[12] Yelp. *Yelp Recruiting Competition*. URL: http://www.kaggle.com/c/yelp-recruiting (visited on 05/01/2013).