Yelp Elite Tag Equation

DrMuhsin

Tuesday, November 10, 2015

Title:

Yelp Elite Tag Equation

Introduction:

There is some debate on the internet over the "Elite" status recognition given by Yelp to the active users and apparently the scoring criteria is remain unclear. The aim of this research is to use the Yelp user data to find the equation for Elite status. Find the weightage that being used and perhaps do a prediction to guide Yelp user to obtained the Elite Status. By using the equations obtained at the end of the research, user will have a better guide on how to qualify themselves for the exclusive Yelp Elite Tag.

The Research Objective:

1. To understand Yelp Elite's behaviour pattern

2. To build an equation model to replicate the criteria and weightage for the Elite's Tag.

The Research Hypothesis: Since yelp is a social applications it will rewards the users based on their actitivity on the Yelp apps. **Initial hypothesis based on Yelp User Data are the following data most probably have the high factor weightage on the predicted equation:**

"Yelping Since", "review counts", "Number of fans", "votes" and also "compliments". For votes and compliments, they have different types of votes and compliments. Hence we also will find out which type of votes or compliments have a better weightate on the Elite Tag Algorithm.

Methodology:

Dataset used

To identify the weightage of the Yelp Elite Tag, **"yelp_academic_dataset_user.json"** was used. This particular dataset was chosen due to the data **describe the behaviour of the non-elite and elite yelp user.**

Data Preparation

The dataset consists of 23 variables and 366,715 observations. Out of 23 variables, only a few variables are maintained and used to identify the pattern. Variables such as name and user ID are removed due to its uniqeness which will not provide any correlation with other variables.

Next step, we replaced the "Yelping Since" which have format of month/year into "Yelping period" which calculate how many months the user have been yelping. This will help to identify whether loyal existing users period of yelping have any significant weightage on Elite Status.

Another data that need to be replaced are the NA-s value that appears mostly on "Votes" and "Compliment" section. The right way to handle this is by **replacing the NA with the mean of respective columns.**

For the elite tag, we classify each user that have been or currently tagged in Elite as 1 and Non-Elite users as 0. Using the prep dataset, **behaviour of non-elite will be classify as 0 and behaviour of Elite users will be tag as 1.**

Cross Validation Techniques

The idea is first to train the machine with behaviour and pattern of Elite and Non-Elite. After training of the modeal has been completed. The testing of classification model can be carried out using Cross validation techniques (90/10).

Eureqa Tools

Once we satisfy with the classification of users. **Eureqa tools** has been choosen to process and comes out with the algorithm to represent the Elite Tag models. The results of this process is elaborate further in Results section.

[link] http://www.nutonian.com/

Following are the snapshots on Eureqa "Eureqa is a Machine IntelligenceT application that automates much of the heavy lifting inherent in analytics and data science. Leveraging automated evolutionary algorithms, Eureqa churns through your data to create accurate predictive models in minutes rather than months.

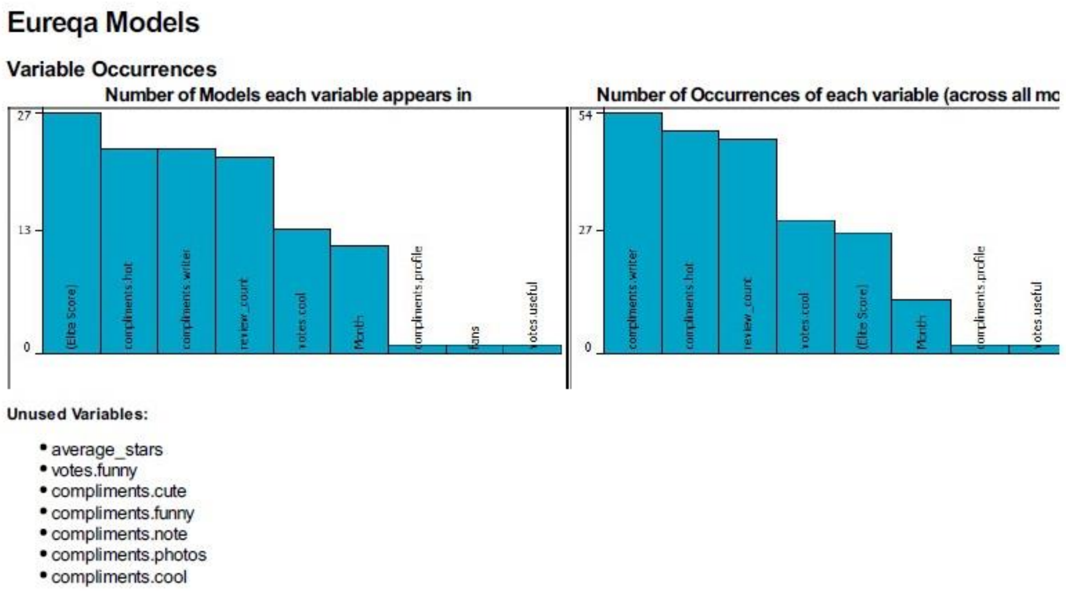
Current modeling techniques require users to choose from a set number of predetermined algorithms (K Nearest Neighbor, Support Vector Machine, etc.). Eureqa builds numeric, time series, and classification models from the ground up, generating and updating models automatically""

Results & Discussion

A. Based on the results from the Eureqa Models, top three variables that have high number of occurance are **"compliments.hot","compliments.write" and also "review coutns"**. "votes cool" and "Yelping period" is among the top variables occured as well.

B. Interestingly, the following 7 variables are unused suggesting that they have no correlation with the Elite Tag classification:

average_stars, votes.funny, compliments.cute, compliments.funny, compliments.note, compliments.photos, compliments cool.



Eureqa Models

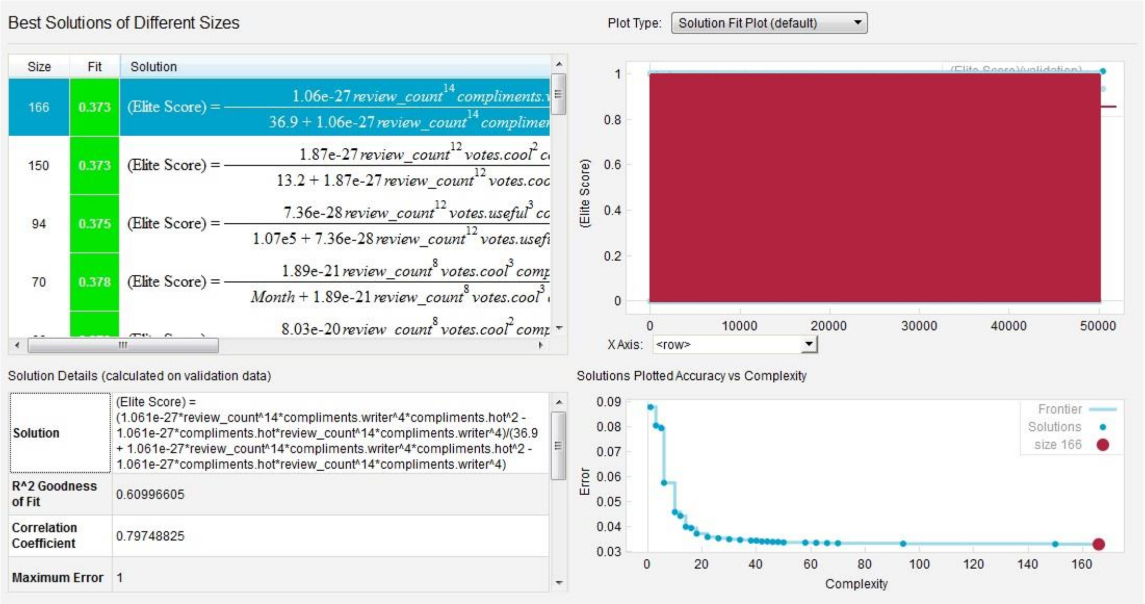
Variable Occurrences

Unused Variables:

- average_stars
- votes.funny
- compliments.cute
- compliments.funny
- compliments.note
- compliments.photos
- compliments.cool

C. After 8 hours of processing the best fit, we stop at Fit "0.373" which have the following details:

- Corrrelation Coefficient: 0.7974
- Mean Squared Error: 0.0312
- Mean Absolute Error: 0.03277



Best Solutions of Different Sizes

Size	Fit	Solution
166	0.373	$(\text{Elite Score}) = \frac{1.06\text{e-}27\, review\_count^{14}\, compliments.}{36.9 + 1.06\text{e-}27\, review\_count^{14}\, complimen}$
150	0.373	$(\text{Elite Score}) = \frac{1.87\text{e-}27\, review\_count^{12}\, votes.cool^{2}\, c}{13.2 + 1.87\text{e-}27\, review\_count^{12}\, votes.coo}$
94	0.375	$(\text{Elite Score}) = \frac{7.36\text{e-}28\, review\_count^{12}\, votes.useful^{3}\, co}{1.07\text{e}5 + 7.36\text{e-}28\, review\_count^{12}\, votes.usef}$
70	0.378	$(\text{Elite Score}) = \frac{1.89\text{e-}21\, review\_count^{8}\, votes.cool^{3}\, comp}{Month + 1.89\text{e-}21\, review\_count^{8}\, votes.cool^{3}}$
		$8.03\text{e-}20\, review\_count^{8}\, votes.cool^{2}\, comp$

Solution Details (calculated on validation data)

Solution	(Elite Score) = (1.061e-27*review_count^14*compliments.writer^4*compliments.hot^2 - 1.061e-27*compliments.hot*review_count^14*compliments.writer^4)/(36.9 + 1.061e-27*review_count^14*compliments.writer^4*compliments.hot^2 - 1.061e-27*compliments.hot*review_count^14*compliments.writer^4)
R^2 Goodness of Fit	0.60996605
Correlation Coefficient	0.79748825
Maximum Error	1

The Elite Tag Algorithm

The resulting equation from this research as follow:

inline equation: $\mathrm{EliteScore} = (1.06065563183796\mathrm{e}^{-27} * \mathrm{review.count}^{14} * \mathrm{compliments.writer}^{4} * \mathrm{compliments.hot}^{2} - 1.06065563183796\mathrm{e}^{-27} * \mathrm{compliments.hot} * \mathrm{review.count}^{14} * \mathrm{compliments.writer}^{4})/(36.9037761517007 + 1.06065563183796\mathrm{e}^{-27} * \mathrm{review.count}^{14} * \mathrm{compliments.writer}^{4} * \mathrm{compliments.hot}^{2} - 1.06065563183796\mathrm{e}^{-27} * \mathrm{compliments.hot} * \mathrm{review.count}^{14} * \mathrm{compliments.writer}^{4})$

Cross Check The Elite Tag Algorithm

Using the algorithm we obtained above, we cross check with the random user data (50,000 sample out of 366,715) to classify the Elite Tag users.

4090 out of 4415 Elite Tag is identified using the algorithm above resulting Accuracy of Elite Tag Algorithm = 4090 / 4415 = 92.6%.

Conclusion

The research question was to find the weightage that have influence on the "Elite Status" on Yelp. Based on the preliminary results , it shows that we can obtained the Elite Tag algorithm with the accuracy up to 92.6%.

Interestingly, eventhough 15 variables feed into the Eureqa tools, the algorithm generation provides us with the Elite Tag algorithm which are based on 3 important variables which are:

inline equation: $\mathrm{EliteScore} = (1.06065563183796\mathrm{e}^{-27} * \mathrm{review.count}^{14} * \mathrm{compliments.writer}^{4} * \mathrm{compliments.hot}^{2} - 1.06065563183796\mathrm{e}^{-27} * \mathrm{compliments.hot} * \mathrm{review.count}^{14} * \mathrm{compliments.writer}^{4})/(36.9037761517007 + 1.06065563183796\mathrm{e}^{-27} * \mathrm{review.count}^{14} * \mathrm{compliments.writer}^{4} * \mathrm{compliments.hot}^{2} - 1.06065563183796\mathrm{e}^{-27} * \mathrm{compliments.hot} * \mathrm{review.count}^{14} * \mathrm{compliments.writer}^{4})$

- Review_count
- Compliments_writer
- Compliments_hot

The possible reason is maybe these top 3 variables have the highest weightage on the algorithm. The algorithm obtained was based on 8hours processing time for Eureqa tools. This is not a complete finalized results from the tools. The processing power is based on the CPU capabilities and time, Eureqa will provides a much better algorithm which have a better fitting and low Mean Squared Error level over time(or one can opt for Eureqa Cloud Service with a fee).

The initial hypothesis is concreted further with the finding that "Review count" and "compliments" type is the highest weightage on the equation. However it is also noted that **"votes_type" is less significant compared to "compliment type"**

The Algorithm shows that for new users, high activity in **"review counts"**,**"compliments.hot"** and **"compliments.write"**will get you to Elite tag faster than any other activities in Yelp.

Further works

To optimize the accuracy, several action can be taken to obtain a better Algorithm to describe Elite user in Yelp. Following are some of the recommendation for future works:

- To run Eureqa for a longer time period to obtain the most optimized algorithm. (or to utilize server to help with the processing). This might take days with current laptop specifications i have. A better processing power will help to reduce the times needed to complete the process.
- Enhance Algorithm will have lower M.S.E value which reuslts a much better accuracy.
- To consider pre-processing filtering techniques to improves the datasets quality.