Data Mining, Fall 2017

Analytical Project Assignment

Modeling quality of suggestions

Associated data sets

suggestions.csv

Data description

This dataset is a 'scrape' from an online forum of a large human resource company. The purpose of the forum is to provide a way for employees to give suggestions to the upper management about a variety of topics. Each suggestion corresponds to a particular thread on the forum (designated by suggestion ID). Other posters can respond to, vote on, or simply view the suggestions in the forum. There is also information about the author of the suggestion such as how many posts he/she has on the site and how long he/she has been an employee (in terms of the number of days with the company).

A group of interns have painstakingly gone through the suggestions and highlighted the ones they would recommend to follow up on (about 500 of them). However, this process is not scalable or sustainable. They have approached your team to develop a method that will narrow the scope of their future search so they do not have to sort through the 'bad' suggestions.

Your tasks

1. Determine which combination of attributes of the suggestion (and maybe the person who wrote it) can be used to predict a 'good' suggestion. Does number of views matter more or less than votes?

2. How much does the 'age' of the employee matter when it comes to their ability to make a good suggestion? Are the employees with longer tenures making better suggestions than those with shorter ones?

3. Can the same data be used to rank employees based on their demonstrated ability to make predominantly good suggestions? Can it be used to identify groups of employees whose suggestions could be aggregated to provide more reliable suggestions than made by the best individuals?

We have aggregated the data based on Author_Id - through which we got the following columns like Responses_mean, Views_Sum, Votes_Up_mean etc. To determine whether the data will be appropriate to rank the employees on their ability to offer quality suggestions, we formulates a new column 'ratio_recommended'. This column is factor of the proportion of good recommendations based on which we will be able to rank authors on their suggestions. Yes, we can extend the same data to rank the employees. Our feature selection process has results in the following features being most significant:

<di>
<li>Votes_Up_Mean
<li>Responses_Mean
<li>Responses_Sum
<li>Votes_Up_Sum
</di>

The above features are not author specific i.e., they will get aggregated when certain employees are taken as a group. This means that certain employees can be aggregated to perform equal to good authors. Our process doesn't select features which are based on an individual like Tenure. Hence, we conclude that by taking employees in groups which have their 'ratio_recommended' higher than a certain threshold will improve the quality of recommendations in the company.

4. Make recommendations to your IT department about better ways they could collect this data in the future. What other attributes would prove useful and why? Would it be possible to build a completely automated suggestion ranking system?

Attributes:

- The can capture the sequence of the suggestions, whether some are followups to a certain or are new ones
- The category under which these suggestions fall
- The time the suggestion was made and hence we can calculate if a suggestion has higher views as it been there for a long time or was it recently posted
- Capture the authors who are voting and apply a weightage to vote depending on the rank of the author.