SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

Predicting response time for the first reply after the question is posted in SAS® community forum

#SASGF

Predicting response time for the first reply after the question is posted in SAS® community forum

Praveen Kumar Kotekal

MS in Business Analytics, SAS® and OSU Data Mining Certificate, Oklahoma State University

Abstract

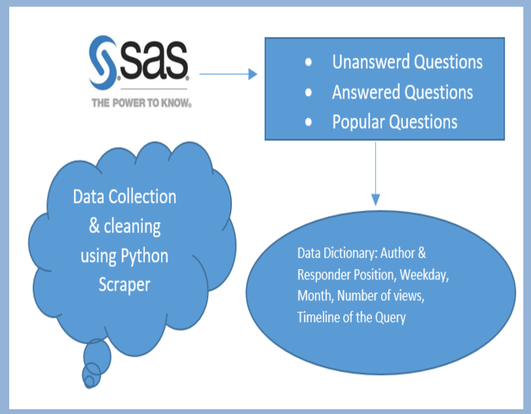
Many inquisitive minds are filled with excitement and anticipation of response every time one posts a question on a forum. This paper explores the factors that impact the response time of the first response for questions posted in the SAS® Community forum. The factors are contributors' availability, nature of topic, and number of contributors knowledgeable for that particular topic.

The results from this project help SAS® users receive an estimated response time, and the SAS® Community forum can use this information to answer several business questions such as following: What time of the year is likely to have an overflow of questions? Do specific topics receive delayed responses? Which days of the week are the community most active? To answer such questions, we built a web crawler using Python and Selenium to fetch data from the SAS® Community forum, one of the largest analytics groups. We scraped over 13,443 queries and solutions starting from January 2014 to present. We also captured several query-related attributes such as the number of replies, likes, views, bookmarks, and the number of people conversing on the query.

Using different tools, we analysed this data set after clustering the queries into 22 subtopics and found interesting patterns that can help the SAS® Community forum in several ways, as presented in this paper.

Background & Data Preparation

Forums are most ubiquitous online participating groups which facilitate users to post questions, interact with people with similar interests and share knowledge in an organized thread layout. One key issue that abates an otherwise interactive query thread from being so is posts not receiving an initial response there by making them dormant among multiple other peer posts and going unnoticed. To further analyse this issue and to determine the factors influencing a post receiving or not receiving timely responses we chose the online SAS® community forum, an ongoing global community effort created by SAS® for SAS® users. In order to perform this analysis, data was extracted from the SAS website using a self-created web crawler tool built on Python. Up to 13443 queries were retrieved beginning from January 2014.



Methodology

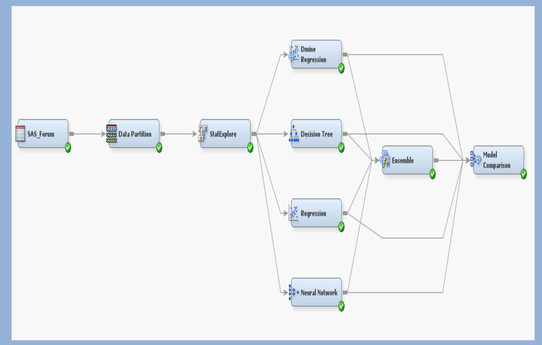
SAS® 12.1 was used to analyze the data. The modeling approach followed for the project is SEMMA (Sample, Explore, Modify, Model and Assess). The data was partitioned into two stratified samples (Training 60%, Validation 40%).

Exploration is used for detecting trends and refining the data. The variables are modified and transformed to adjust skewness and kurtosis values. Complete data for all the fields was available for only 10% of the data. Tree based Imputation methods were used to make necessary imputations. Directly related variables were rejected as part of analysis. Models ranging from decision trees, neural networks, logistic regressions, Dmine regressions were applied as part of the analysis.

Data Exploration & Model Building

Initial Data Exploration was done to identify the relations between the target and inputs. StatExplore was used to identify the usefulness of variables against the targets. Highly correlated input variables were removed.

Clustering node was useful in identifying and picking up the important variables from the initial list of variables. After reducing the input space, the filtered set of variables were used to make the predictions about the responses. Tree models in general were more robust and predicted the responses well in this case. Dmine regression was found to be the best model based on Average Squared Error.



Selected Model	Model Description	Target Variable	Selection Criterion Valid: Average Squared Error	Train: Sum of Squared Errors
Y	Neural Network	Tgt_h	0.192846941	714.8735608
	Ensemble	Tgt_h	0.209291624	988.4622429
	Decision Tree	Tgt_h	0.223644875	1208.712527
	Regression	Tgt_h	0.269457042	1264.744769
	Dmine Regression	Tgt_h	0.283554568	1536.23115

Predicting response time for the first reply after the question is posted in SAS® community forum

Praveen Kumar Kotekal

MS in Business Analytics, SAS® and OSU Data Mining Certificate, Oklahoma State University

Results

- It is found that first reply response time factors vary widely across month based on weekday in which question has been posted. Sub_topic and number of likes to question posted in the SAS® community forum found to be the most influencing factors in distinguishing the quickly replied questions to questions which are answered after 1 day delay. Number of people viewed that question turned to be least likely factor to be considered. First reply respondent position has come out as the most important factor for a quick reply.

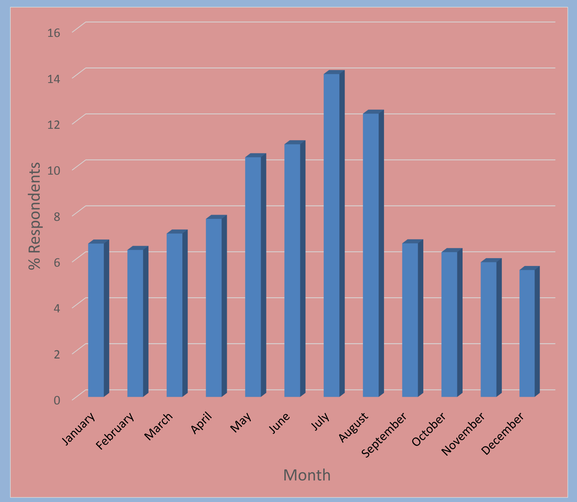


Fig.1 Queries Answered per Month

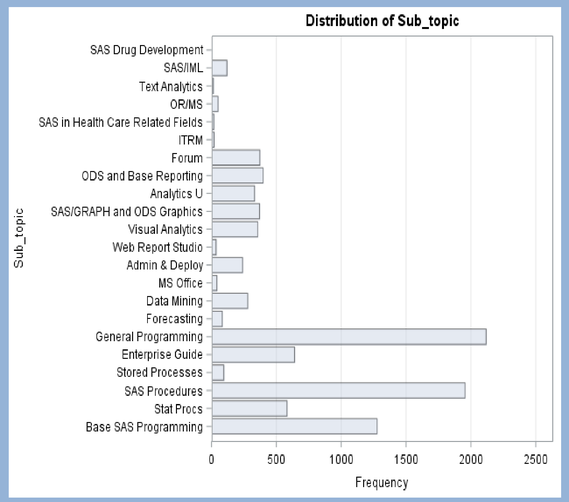


Fig.2 Queries per Sub Category

Some of the months in a year are found to be more active than others. September, October months were identified as high frequency months having 1.8 times more queries than any other month.

Majority of the sub topics having quick response rate had super contributors as responders whereas for sub-topics having delayed response occasional contributors were the respondents.

Around 40% of the questions in Visual Analytics, Stat Procs and Data Mining have a delayed response time of more than one day. Queries are answered more quickly on weekdays rather than on weekends. From the analysis weekday was found to be playing major role in determining the first reply time after the question is posted in community forum. Most of the questions are answered on weekday 4 and weekday 5.

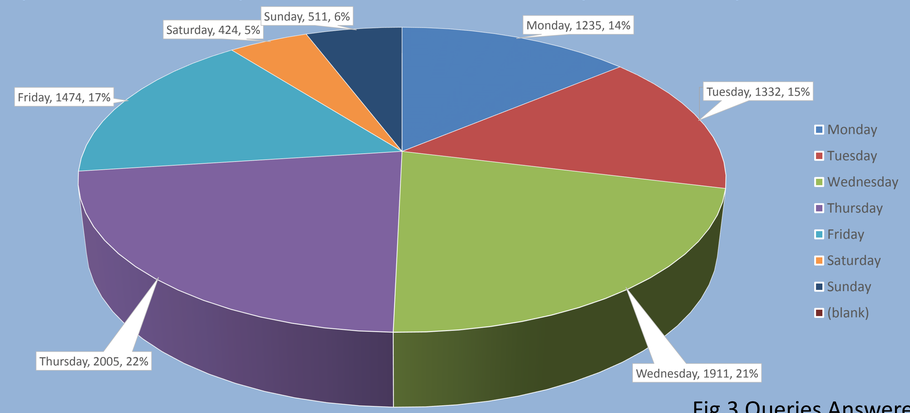


Fig.3 Queries Answered per Weekday

Conclusion

The main objective of the project was to discover the factors that influence the first reply response time. The facts discovered through the process not only give insights to the strategic management of forum but also provide a decision making process backed up by the data. SAS® Enterprise miner is a very powerful tool that helped in finding these patterns given the amount of data thrown in with larger number of variables

Acknowledgement

We thank Dr. Goutam Chakraborty, Professor, Department of Marketing and Director, Master of Science in Business Analytics - Oklahoma State University for his support throughout the research.

Contact

Your comments and questions are valued and encouraged. Contact the author at Praveen Kumar Kotekal, Oklahoma State University, Stillwater, OK, 74075 Email: praveen.kotekal@okstate.edu

SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

LAS VEGAS | APRIL 18-21

#SASGF