Email

Supposed that Microsoft would like to update the functionalities of the Outlook Emails, and the software can better differentiate the regular email from spams. Data scientists assess the existing problems and decide to introduce a new classification model to solve the problem.

First of the all, in order to attain what can be referred to as a business understanding, data scientists begins with spending the time talking to the stakeholders and software users, they find the biggest problem of unsatisfying user experience is that users are always missing the important emails because the email filtering is out of performance. Once the goal is clarified, the next piece of the puzzle is to figure out the objectives that are in support of the goal. Now the question being asked is: What is the best way to improve the user experience of the Outlook Email? The solution to this problem is focused on improve the accuracy of spams classification.

Followed by that, the appropriate analytic approach for the problem is Naïve Byers since we need to improve the percentage of the accuracy of the email filtering. Prior to undertaking the data collection and data preparation stages of the methodology, it is vital to define the data requirements for Naïve Byers classification. This includes identifying the necessary data content, formats and sources for initial data collection. In our case, we can use word counts function to split the email contents into the words, then use calculated probability of certain words appeared in health mails and spam to determine mails.

After collecting data and do the preparation, we can build the Naïve Byers model. To be noticed that since our model is predictive, we need to partition the data into training set and test set, which can be used to evaluate the model. When the evaluation is accomplished, depending on the purpose of the model, it should be rolled out to a limited group of users in a test environment, to build up confidence in applying the outcome for use across the board. In preparation for solution deployment, the next step was to assimilate the knowledge for the business group who would be designing and managing the intervention program to filter the mail. The model should be deployed on the existing server of outlook email.

Last but not the least, the plan for the feedback stage included these steps: first, the review process would be defined and put into place, with overall responsibility for measuring the results of a “flying to risk” model of the email filtering. Second, spams would be tracked and recorded. Third, the intervention would then be measured to determine how effective it was in reducing spams appearing in the inbox.

After the deployment and feedback stages, the impact of the intervention program on wrong classification rates would be reviewed after the first year of its implementation. Then the model would be refined, based on all of the data compiled after model implementation and the knowledge grained throughout these stages.