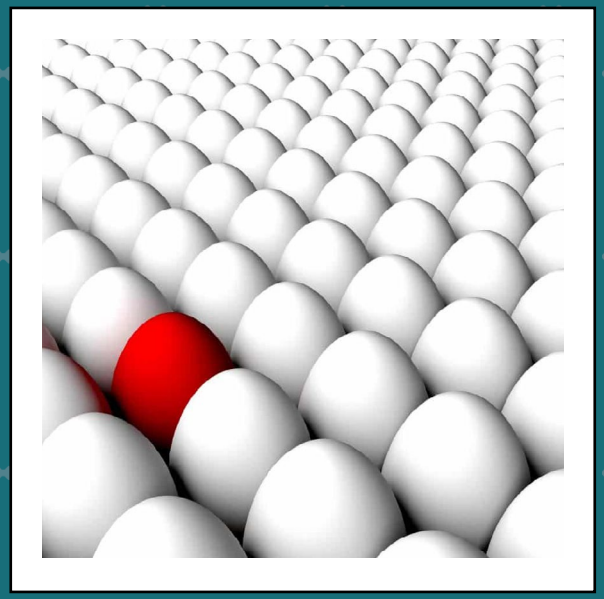

mosaic
DATA SCIENCE

http://mosaicdatascience.com/

info@mosaicdatascience.com

Data Science Case Study

Purchase Order Email Classifier

For one of the US's largest healthcare systems, Mosaic designed and deployed a custom machine learning tool to identify anomalous purchase orders.

Industry

Hospitals

Use Case

Automating Purchase Order Review

Techniques

NLP, Logit, Analytics Adoption & Scale

Outcome

Tool flags anomalous purchase order over entire hospital supply chain.

BACKGROUND | OPPORTUNITY FOR AUTOMATION

Many hospital systems receive thousands of purchase orders (PO) each week to keep the facilities stocked with the products and equipment necessary for healthcare professionals to perform their day-to-day functions, from administrative tasks to sophisticated surgeries. For the one of the largest healthcare system in the United States, this is no small feat. For this hospital system, hospital staff place POs through a variety of modes, including phone, fax, email, and Electronic Data Interchange (EDI). Phone orders are confirmed by voice at time of placement; fax and email orders are confirmed by a reply email to a designated mailbox; and EDI orders are confirmed electronically by the vendor. The Supply Chain department for this healthcare system has to process these POs for 22 hospitals and over 500 care facilities.

Staff members of a small customer service team within the Supply Chain department manually review the PO confirmations received in the designated email account to ensure that exceptions are routed appropriately and promptly addressed. Examples of common exceptions include backorder or discontinued product notifications, price changes that need to be acknowledged before the order will be processed, or errors in the PO, such as unrecognized product numbers, that require the healthcare provider to submit a revised PO. Due to the sheer volume of POs placed on a weekly basis and the team's other competing duties, there are often exceptions that slip through unnoticed. Missed exceptions cause downstream issues through the supply chain that can have very high costs – such as last-minute surgery postponements due to missing equipment that can negatively affect patient health and outcomes and are costly to a hospital's bottom line.

The Supply Chain department determined that a machine learning tool for automatically flagging emails requiring attention as they arrived in the PO confirmation mailbox would reduce the risk of missed exceptions and significantly reduce the workload on the Supply Chain department. The hospital system needed an analytics consulting partner who had experience implementing Natural Language Processing (NLP), text analytics, and production machine learning tools, and Mosaic Data Science was poised to assist with their PO challenges.

APPROACH

The Supply Chain department had already collected a large training dataset of 19,000 email confirmations that had manually tagged as to whether the confirmations required attention and the type of attention required. These would be leveraged to develop the initial NLP development that would form the core of the tool, and new emails would be added to the training set over time as they were automatically tagged by the model and verified by Supply Chain associates.

The machine learning consultants at Mosaic decided to attack this project in two phases. The first phase of the project focused on data analysis, NLP development, and performance testing of a machine learning email classification algorithm in an offline, experimental setting. In the second phase of the project, Mosaic implemented an email classification tool based on the developed algorithm in the healthcare system's production environment.

NLP DEVELOPMENT

Mosaic developed and evaluated multiple candidate machine learning models for performing email classification. Machine learning models are a class of automated pattern recognition algorithms that can identify patterns in a "labeled" training dataset – in this case, the list of emails manually tagged as to whether or not they need a response – and apply those patterns to add labels to a previously unlabeled dataset – in this case, new PO emails as they are received.

Overall, approximately 7% of tagged emails were labeled as needing attention. The classification models developed by Mosaic were evaluated and compared across performance metrics established based on the intended use of the model. The two primary metrics used for comparison were precision and recall.

- Precision is the percentage of emails flagged by the model as "Needs Attention" that are accurately labeled. A high precision means a relatively low number of false positives (emails flagged as "Needs Attention" that do not actually need attention).
- Recall is the percentage of all emails that need attention that are accurately flagged by the model as "Needs Attention." A high recall means a relatively low number of false negatives – few emails that need attention will be bypassed by the model.

An ideal classification model will have both high precision and high recall, meaning that almost all emails that need attention are captured by the model with very few false positives. However, real-world classification models are never perfect. A particular model must be tuned to adjust the tradeoff between precision and recall based on the business use case for and objectives related to the model. A model that prioritizes precision will only

flag the emails that most clearly need attention but will likely miss many other emails that need attention, meaning that it will have a lower recall. A model that prioritizes recall will more aggressively flag emails as needing attention but will likely capture some emails that do not need attention, meaning that it will have a lower precision.

The models were developed and analyzed in Python using well-established open source NLP development libraries for text analytics and machine learning. Python was also used to develop the production software in the second phase of the project.

NLP MODEL SELECTION

After a comparison of a variety of machine learning models, Mosaic's team of analytics consultants selected a logistic regression (logit) model to perform the email classification task. The model proved to have excellent performance based on the precision-recall tradeoff while also providing good interpretability and robustness. The team performed some NLP development variable transformations before running the training set through the algorithm. This included term frequency-inverse document frequency (tf-idf)[1], a normalization technique popular in NLP development, that expresses term occurrence relative to "typical" occurrence frequency across the full set of emails.

Once Mosaic's data scientists had trained and validated the logit model, they could begin to run new emails through the model in a scoring mode. The trained model can assign a "score" between 0 and 1 to new emails as they arrive in the confirmation inbox. Scores closer to 1 imply a higher likelihood that the email needs attention, while scores closer to 0 imply a higher likelihood that the email does not need attention. Based on analysis of model scores on previously labeled emails, a threshold for flagging emails as "Needs Attention" was selected to balance the precision and recall of the model.

MODEL PERFORMANCE

The logistic regression model described above was evaluated based on an out-of-sample performance analysis. A portion of the labeled emails (in this case, all labeled emails dated 5/1/2017 or later) was set aside as a holdout set. The logistic regression model described above was trained on the remaining emails. This trained model was then applied to the holdout set to generate scores and flags, and the accuracy of the scores and flags were assessed by comparing them to the known labels for the holdout emails. This analysis approach provides the best estimate for how well a model will perform on new, unlabeled data and helps to ensure that the final model is not overfit to eccentricities unique to the training data.

Figure 1 below shows the holdout set score distribution for the two classes of emails: “Needs Attention” and “Good/Other.” The separation of the two distributions with “Needs Attention” emails (blue bars) clustered toward higher values and “No Attention Needed” emails (yellow bars) clustered toward lower values indicates that the model does a very good job of distinguishing between the two classes.

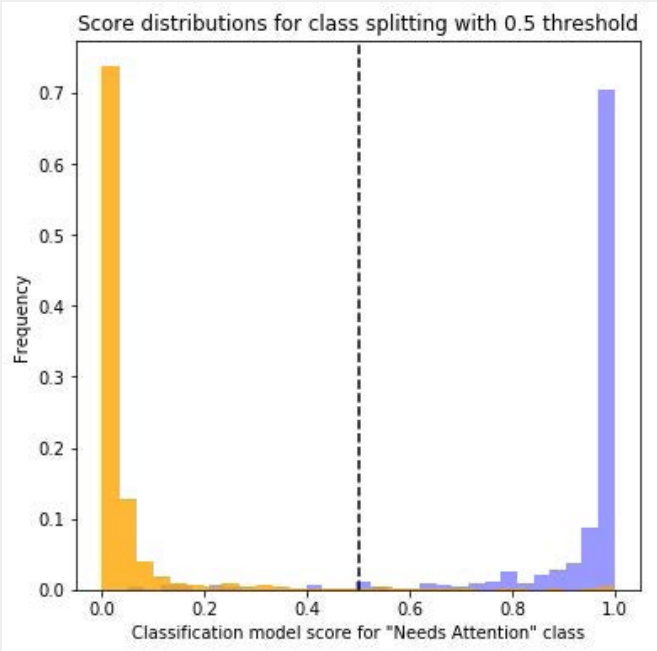

Figure 1. Score distributions for logistic regression model on holdout emails. Blue bars represent scores for emails with actual label “Needs Attention”; yellow bars represent scores for emails with actual label “No Attention Needed.”

By setting a classification threshold between 0 and 1, the scores can be used to flag emails as needing attention or not. Emails scored above the selected threshold are flagged by the model as needing attention. Higher thresholds will reduce incorrectly flagged emails (increasing precision) but will increase the number of emails that need attention but are not flagged (reducing recall). Lower thresholds will correctly flag more emails that need attention (increasing recall) but will incorrectly flag a higher number of emails (reducing precision). Table 1 quantifies this precision-recall tradeoff for key recall thresholds.

Threshold	Recall	Precision
0.88	85%	91%
0.78	90%	85%
0.58	95%	79%
0.30	98%	60%
0.15	99%	48%

Table 1. Precision-recall tradeoff for logistic regression model

By consulting the precision-recall chart, the Supply Chain personnel managing the production model can select a threshold that best matches business objectives according to the resulting precision and recall values.

RESULTS

Precision and recall can also be used to translate model results into business terms. For example, with a model threshold of 0.15, approximately 99% of emails that need attention will be correctly flagged (meaning that 1% of emails that need attention will not be flagged). The corresponding precision of 48% means that for every 1 email that is correctly flagged, 1 additional email will be incorrectly flagged. Thus, in a batch of 1,000 emails, of which 70 need attention, we would expect that approximately 144 emails will be flagged, 69 of the flagged emails will correctly need attention, and 1 email that needs attention will be missed by the model. Using the model with this threshold, the Supply Chain department customer service employees would need to evaluate fewer than 15% of the arriving emails to respond to exceptions, with a risk of only 1% that any individual email that needs attention is missed. Similarly, if a threshold of 0.58 were used, fewer than 9% of emails would require manual review if the department decided to accept a risk of 5% that any individual email needing attention would be missed by the model. With these thresholds in hand, company leadership can now weigh the risks of missing PO problems against benefits of having staff in this small customer service group focus on other priorities.

Without the machine learning model developed and implemented, the hospital system used to budget time and resources for the customer service team to pour over all emails to identify POs needing attention. Mosaic's tool not only provides a reduction in time spent evaluating PO confirmations, but also eliminates almost all of the missed exceptions due to human error in reviewing thousands of purchase orders on a weekly basis, resulting in a number of beneficial downstream financial and patient experience effects.

Endnotes

1. https://en.wikipedia.org/wiki/Tf%E2%80%93idf