Predicting Shopper Conversion for Valassis

Introduction

Valassis provides a powerful service that shows advertisements based on a user's interests from his or her browsing history. Our goal as data scientists is to help Valassis determine which users are most likely to convert into buyers when shown a company's ads, optimizing Valassis' marketing costs and effectiveness. To do so, we built a model to classify whether or not shoppers would convert based on their long and short-term interest in specific topics. Because our dataset has many unconverted shoppers, we did not want to use misclassification error as a metric. Rather, we wished to maximize the probability of predicting a conversion accurately (true positive rate) while minimizing the probability of wrongly predicting a conversion (false positive rate), as this would drive the most revenue for Valassis. As such, we used the area under the curve (AUC) of the receiving operator characteristic (ROC) curve as our main metric to measure model success.

Exploratory Data Analysis and Data Engineering Process

We noticed that the initial list of 1411 interests consisted of a number of general categories (i.e. Travel, Arts & Entertainment) with many additional subcategories (i.e. Travel/Tourist/Destinations). Using a heat map (Figure 1), we noticed that subcategories with the same general category are correlated, and thus would create issues with classifying data. Hence, we grouped interests into the 25 general category groups, eliminating the subcategories, therefore reducing correlation amongst features. We also decided to first focus on the long-term interest data, since every user had long-term interests (LTIs) but not necessarily short-term interests (STIs).

Figure 1

We transformed each user into an entry in a panda dataframe with 26 features: the first 25 features contained summed up relative importance values for each of the 25 interest categories and 1 feature for whether or not that user converted. There were also a number of topic IDs in the original dataset that did not correlate to a known topic, and thus, we discarded them because they could not be grouped with any other general category of topics. We then normalized each user so that their relative importances for each feature would sum to 1.

Amongst the users that did and did not convert respectively, we then visualized their distributions across each interest group (Figure 2 & 4), and took the absolute difference of those percentages (Figure 3). Figure 3 indicated that Arts and Entertainment, Autos & Vehicles, Health and News could possibly have the most predictive power. In particular, analyzing Figure 2 & 4, LTI in Autos & Vehicles and Health seemed that predict that a user would convert, while LTI in Arts and Entertainment and News seemed to predict that a user would not convert. Our group hypothesizes that LTI in Autos & Vehicles and Health could either signify niche interests in these fields, or perhaps imply an older age group, which therefore has more purchasing power and may be more inclined to convert. On the other hand, many individuals have LTI in News and Entertainment, and such users likely be browsing the site with no intent of purchasing/converting. However, our data did not include the kinds of ads that users clicked on, nor the demographic of users for us to confirm our inferences.

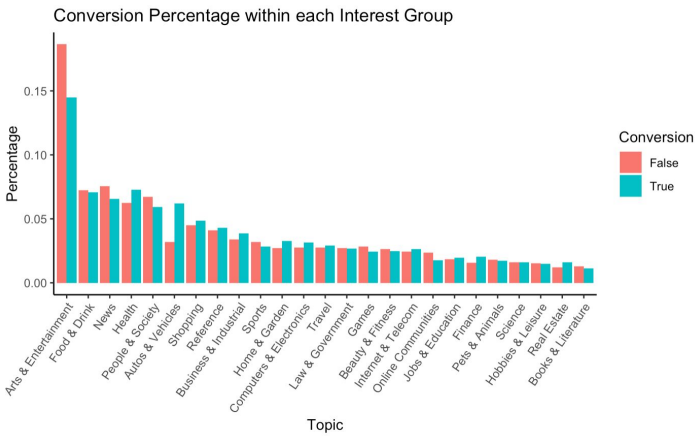

Figure 2

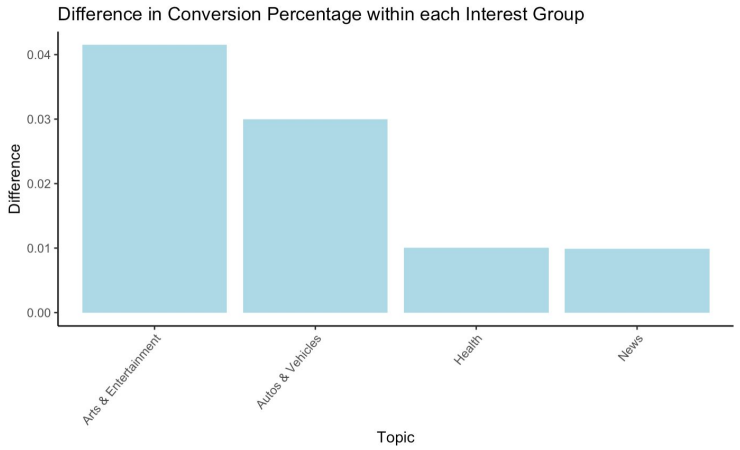

Figure 3

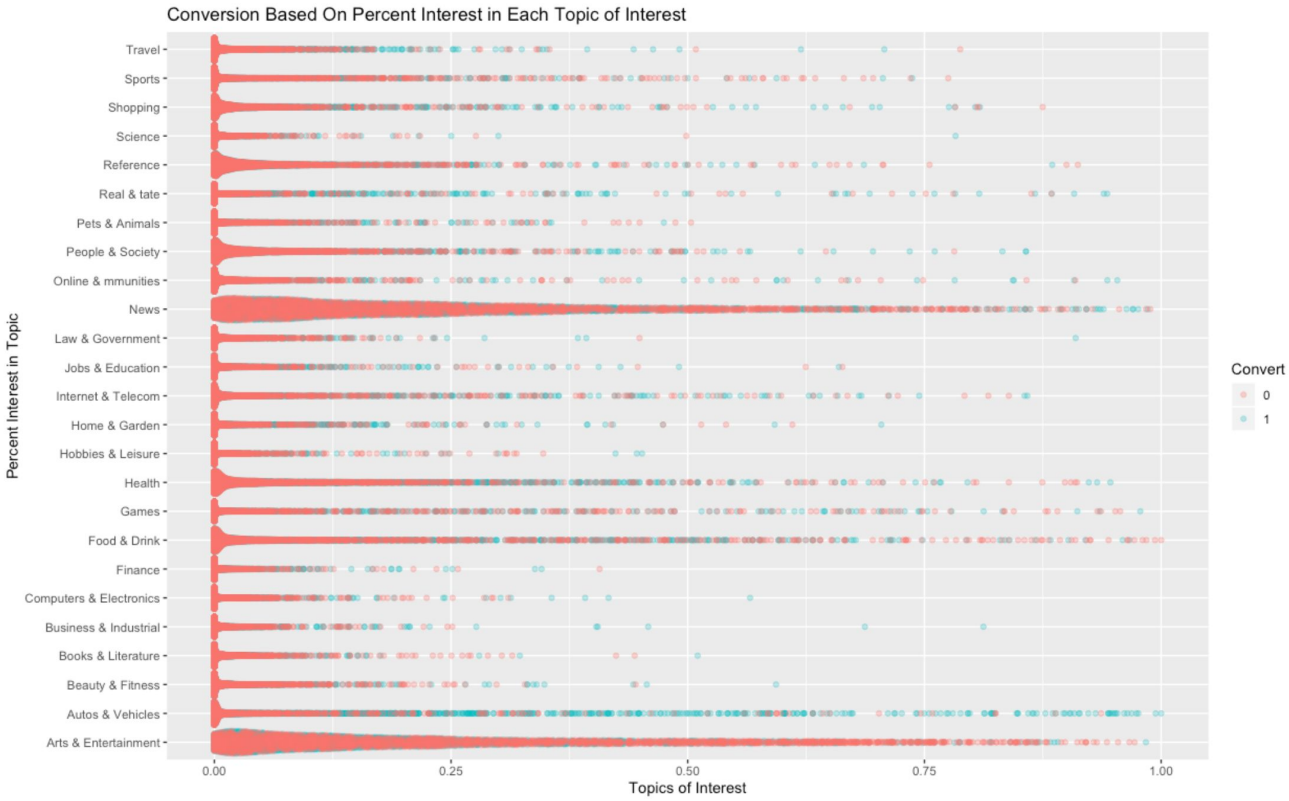

Figure 4

Model 1 - Focusing on Long-Term Interests

We first began by focusing on the user's long-term interests. We used a random sampler on our data to get an equal representative sample from each class, since our data is severely imbalanced. Then, to classify our training data, we built multiple models, such as boosted decision trees, support vector machines, and random forests (using 10 trees as a default). We then transformed our test data, normalizing it too such that the sum of % interest for each user would be 1. Applying our trained models to the test data, we noticed that the random forest model performed the best. Despite efforts to balance the data, both boosted decision trees and support vector machines tended to predict that all users would not convert. As such, due to random forest's lower propensity to overfit, the ease of tuning its hyperparameters quickly, and the ability to extract variable importance from the model, we eventually chose to implement the random forest model.

Initial Findings

Using our random forest model gave us an accuracy of 0.637 and standard deviation of 0.0103 on the training data, when we implemented a 3-fold cross validation. Applying it to the test data gave us an accuracy of 0.722 and ROC AUC of 0.65. We extracted the variable feature importance from this model, and noted, as expected, that the Long Term interest topics most important in classifying users were Autos & Vehicles, Arts & Entertainment, News and Health (Figure 5).

Model 2 - Incorporating Short-Term Interests

After analyzing the data using only the long-term interests we began to look at how STIs might play a role in predicting conversion. Only ~27% of the data had any information about STIs, so we decided to evaluate these users separately. We trained a new random forest model using now 50 features including both the LTI and STI features, and only data that had STI information. We then tested it using only the test data that had STI information. We also normalized the STI features in the same way we did for the LTI features such that for each user the relative importance values would sum up to 1. We call this model, "Model A".

Performance and Further Improvements

Using our random forest model gave us an accuracy of 0.65 and standard deviation of 0.0117 on the training data, when we implemented a 3-fold cross validation. Applying it to the test data gave us an accuracy of 0.743, an AUROC of 0.86 (Figure 6) and a recall of 0.97. We extracted the variable feature importance from this model, which we will discuss on the next page. Given that our model performed exceedingly well for users with STI and LTI data, we wished to provide Valassis with two models, one to use when STI data was available (Model A), and one to use when STI data was not available (Model B). To obtain Model B, we re-trained a new model as per Model 1, this time just using training data whereby STI information was not available. We then tested this model on data where STI information was not available, but realized that Model B had a higher FPR than Model A, with an AUROC of 0.62.

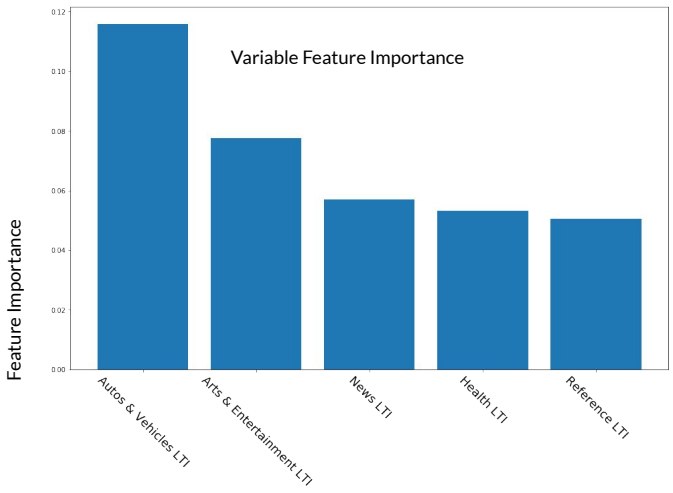

Figure 5

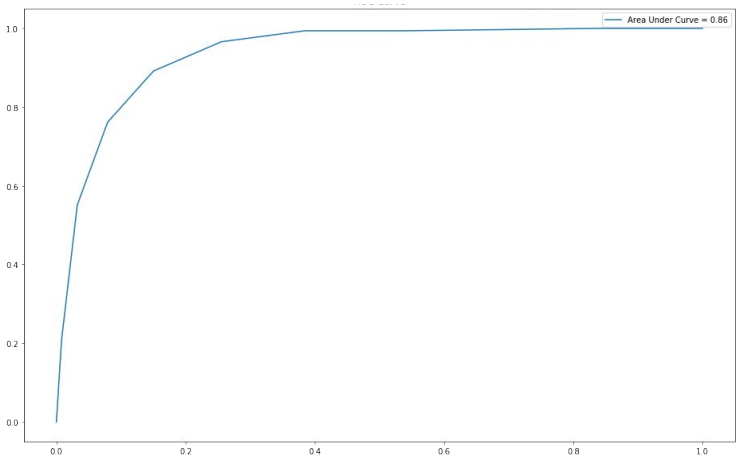

Figure 6

Model 3 - Tuning Hyperparameters with Nested Cross-Validation

As noted before, Model B was unable to classify test data with only LT interests very well. We hence decided to tune our hyperparameters with a 3-fold nested cross-validation on our training data to improve our models' predictive power. We varied the number of trees, the maximum depth of trees, the minimum samples needed to split an internal node, the minimum samples in a leaf, and whether we do a bootstrap sampling of data points. This allowed us to obtain an optimal value for our hyperparameters for both models.

Final Findings

Using these optimal hyperparameters, we re-trained our models and applied it to the test data. While model A did not see significant improvements (as it already held significant predictive power), model B saw marked improvement, with the AUC of the ROC increasing from 0.62 to 0.64.

In addition, with the optimal hyperparameters of model A, we obtained the variable importance (VI) of both LT and ST interests (Figure 7). We then visualized the difference between the VI of ST and LT features (.. We noted that topics that had predictive power in the LT tended to have predictive power in the ST too. However, interestingly, topics like Finance and Computers & Electronics seemed to have surprising predictive power in the ST, even though they did not have much importance in the LT.

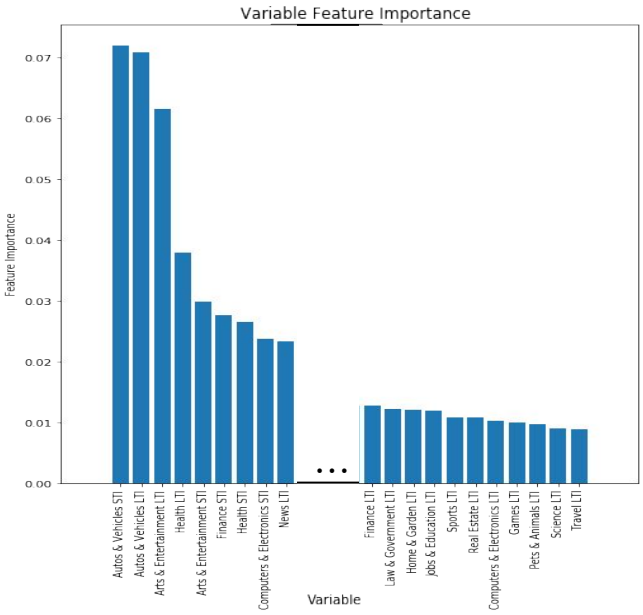

Figure 7

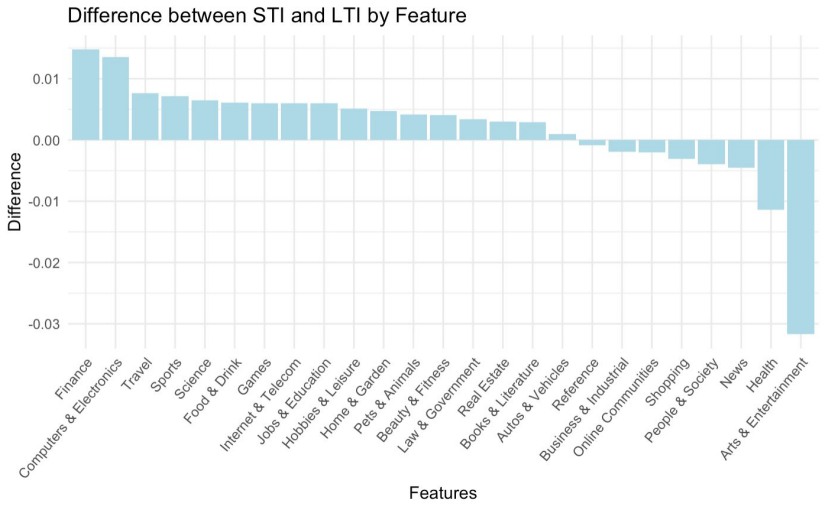

Figure 8

Conclusions

In conclusion, we have two models for Valassis to use. If Valassis can access both ST and LT data, it should use model A, and if it only has LT data, it should use model B. We note that we are best able to predict conversion for users that provide both ST and LT data, and it might be most worthwhile for Valassis to focus on users which they are able to obtain ST interest data on. Looking at LT interests, Autos & Vehicles, Arts & Entertainment, News and Health have the most predictive power, and ST interest overlap with these categories pretty heavily. However, in the ST, it is also important to consider interests in Finance and Computers & Electronics. Aside from the data, these patterns seem to make sense from a logical perspective. For example, it makes sense that people who very recently gained an interest in computer & electronics would likely be searching for a new device to purchase, and thus be more prone to conversion. Our recommendation to Valassis is that they should target their advertising most heavily to those who show a long term-interest in Autos & Vehicles, Arts & Entertainment, News, and Health, but that they should also be aware of when short term interests change in a few select categories and target those individuals when that interest peaks.