INTRODUCTION

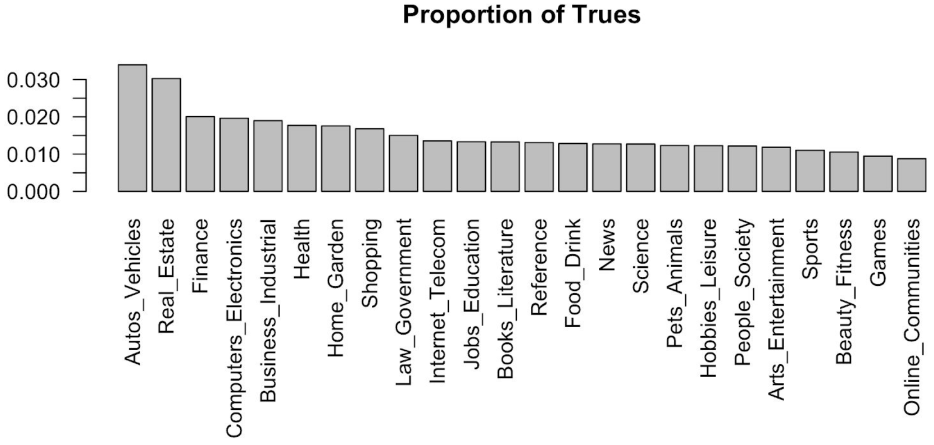
In digital advertising, a “conversion” refers to the event when the shopper clicks on an ad and performs a valuable action such as signup, registration, or make a purchase. Marketers like Valassis aim to optimize ad spending so that customer acquisition is maximized. Obviously, the type of content and categories put out in the ad will cater in various ways to different users in their ultimate decision to "convert" or not. Our team sought to analyze the most important categories in ads that galvanize shoppers to "convert" - an indicator of success for the ad campaign. We also wanted to gain insight into common interests shoppers have among different categories such as those who are interested in health may also tend to be interested in diet and nutrition.

DATA CLEANING

For the interests topics, we first grouped the *topic_name* variable into a new, more general variable *topic_category*. For example, all *topic_name* values that started with “/Arts & Entertainment” were assigned the *topic_category* value: “Arts & Entertainment”. This reduced the 1411 topics to a more manageable 25 categories.

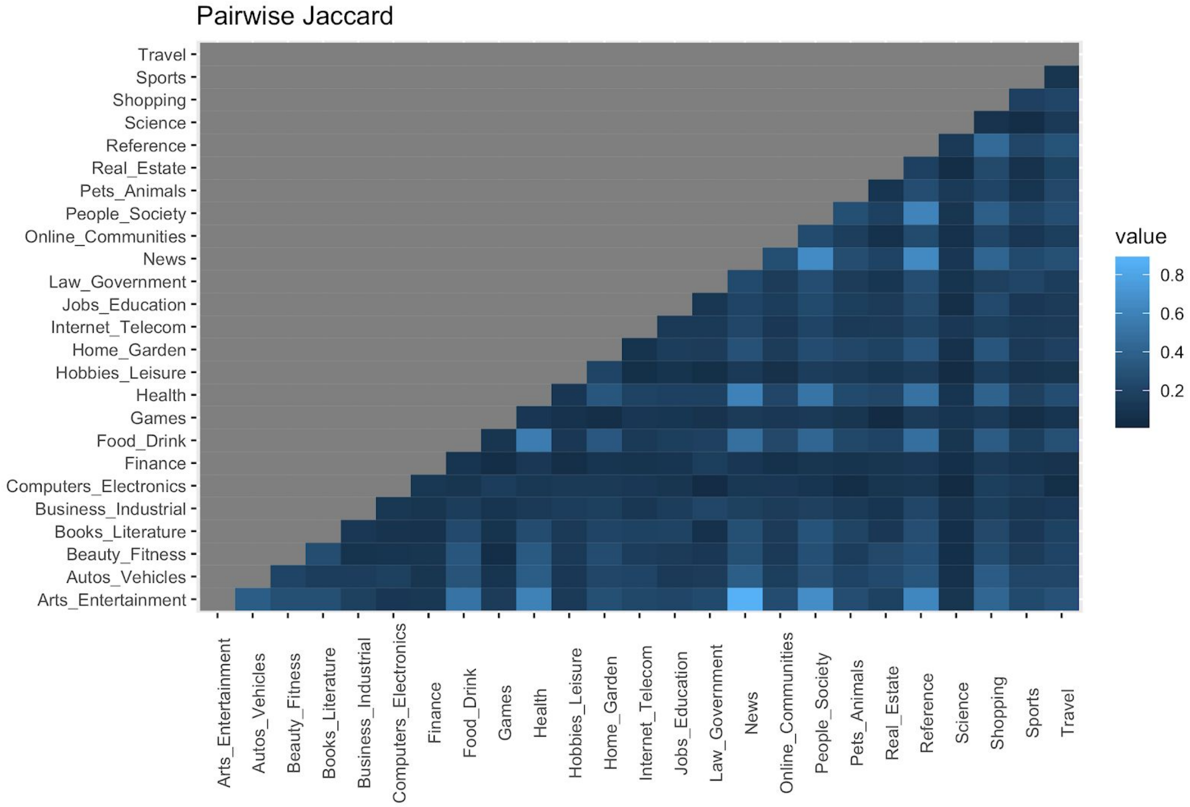
We then assigned the topic_category values to the training and validation test set samples. We took the aggregate proportion for each sample for each category. We also decided to split the two datasets into four. A training set with only long-term interest proportions and one with only short-term, and the same with validation. We also found that many of the aggregate short-term proportions were greater than one, and therefore ultimately decided to only work with long-term interests.

EXPLORATORY DATA ANALYSIS

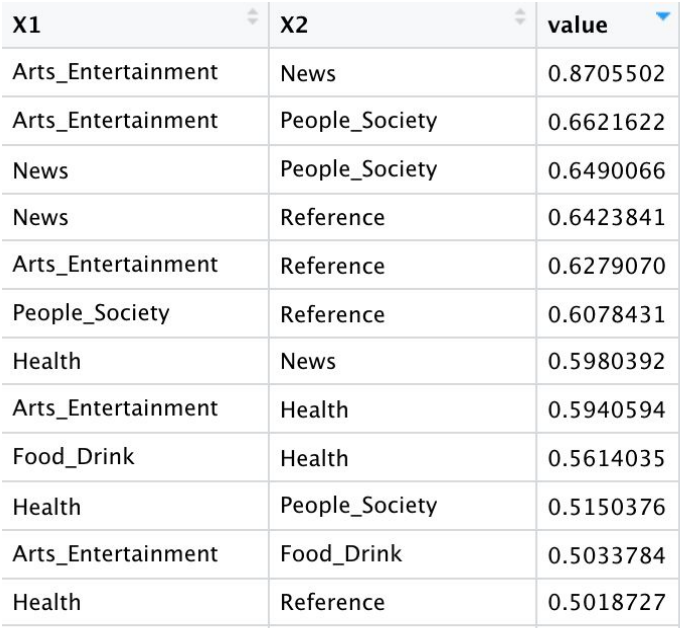


“Convert” to “Not-convert” Ratios When Category is Present in User Interests

We used Jaccard similarity to gauge the pairwise relationships between categories that impact user interests, i.e. for which pairs of categories will users show interest in both. For each category, we essentially list all the users who have an interest in that category that exceeds a threshold of 1%. Then, by pairwise comparisons of category sets of users, we found the ratio of number in the intersection and number in the union. This produces a heatmap Jaccard matrix:



In the heatmap, we see that the lighter category pair depict higher ratios. If we list out the top ratio values and their corresponding pairs, we see:

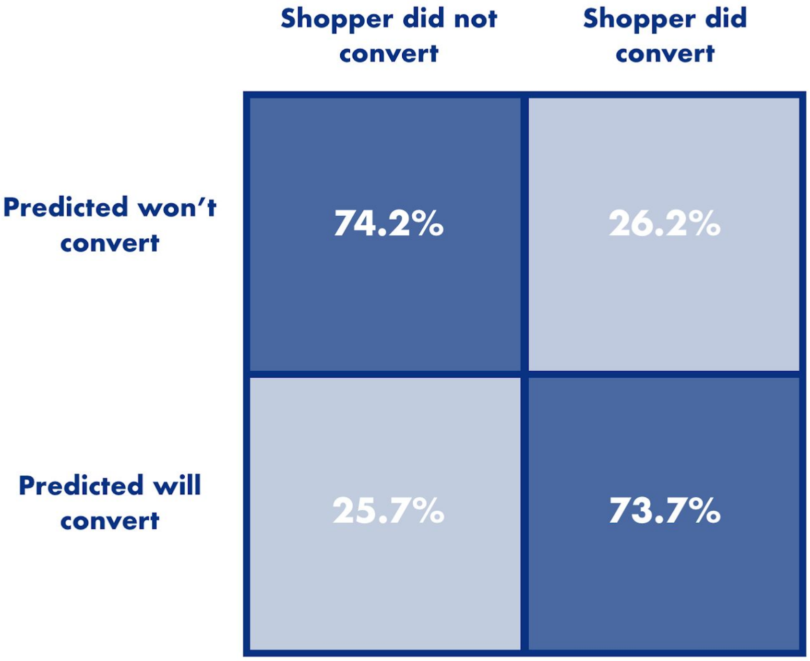


X1	X2	value
Arts_Entertainment	News	0.8705502
Arts_Entertainment	People_Society	0.6621622
News	People_Society	0.6490066
News	Reference	0.6423841
Arts_Entertainment	Reference	0.6279070
People_Society	Reference	0.6078431
Health	News	0.5980392
Arts_Entertainment	Health	0.5940594
Food_Drink	Health	0.5614035
Health	People_Society	0.5150376
Arts_Entertainment	Food_Drink	0.5033784
Health	Reference	0.5018727

From this we see that if a user is interested in art and entertainment, they will also be inclined to view other topics, like news and people & society. In essence, we see that the most overlapping categories that users find interest in are: Arts & Entertainment, News, People & Society, Reference, Health.

RANDOM FOREST MODEL

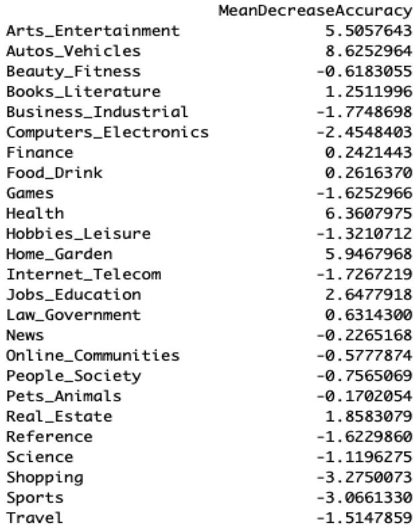
We then created a random forest model that trained on the long term interaction data to predict whether a shopper would convert based off their interests. Most of the shoppers do not convert and therefore the dataset is heavily imbalanced. When we first fit our random forest model on the data, we found that the model always predicted the shopper would not convert due to the imbalanced data. We then used undersampling to make sure the sample training data that the random forest trained on was evenly split between shoppers who converted and shoppers who did not. Our next goal was to identify the optimal hyperparameter value for the variable *'mtry'*, which corresponds to the number of features randomly sampled as candidates at each split. We originally were planning on using the grid search function in the *caret* package to optimize our hyperparameter; however, there were issues installing the *caret* package on the local version of RStudio. We decided to adapt by iterating through values of *mtry* and calculating the value that had the lowest misclassification error. Once we found the optimal hyperparameter we calculated the accuracy of our random forest by randomly sampling (using undersampling to ensure an even mix of test data) the validation set for long-term interactions. We then ran this ten times in order to gain insight into how accurate our model is. We found that our model has a mean accuracy rate of 70.4% with a standard deviation of 2.52% and had a max accuracy rate of 74% with a minimum of 66%. Below is an example of the confusion matrix generated from our model testing.



	Shopper did not convert	Shopper did convert
Predicted won't convert	74.2%	26.2%
Predicted will convert	25.7%	73.7%

CONCLUSION

We performed feature analysis on our random forest to understand what the profile of a shopper who converts is. Below is a screenshot of the results:



```
                      MeanDecreaseAccuracy
Arts_Entertainment               5.5057643
Autos_Vehicles                   8.6252964
Beauty_Fitness                  -0.6183055
Books_Literature                 1.2511996
Business_Industrial             -1.7748698
Computers_Electronics           -2.4548403
Finance                          0.2421443
Food_Drink                       0.2616370
Games                           -1.6252966
Health                           6.3607975
Hobbies_Leisure                 -1.3210712
Home_Garden                      5.9467968
Internet_Telecom                -1.7267219
Jobs_Education                   2.6477918
Law_Government                   0.6314300
News                            -0.2265168
Online_Communities              -0.5777874
People_Society                  -0.7565069
Pets_Animals                    -0.1702054
Real_Estate                      1.8583079
Reference                       -1.6229860
Science                         -1.1196275
Shopping                        -3.2750073
Sports                          -3.0661330
Travel                          -1.5147859
```

From our analysis, we can see that variations in any of the predictors do not heavily correspond to a change in accuracy suggesting that no predictor is extremely strong which makes sense considering as conversion moments are very rare. However, we can suggest from this information that marketers, such as Valassis, should focus their ad spending on shoppers who are interested in automobiles, health, home gardens, and arts & entertainment. Interests in these subjects lead to a relatively strong correlation that the consumer will convert on an ad. This makes sense, for the most part, as these are hobbies in which hobbyists are willing to spend money on a new offerings such as new suspensions for their trucks or a new brand of fertilizer for their vegetable garden.

We can also gain some insight into what other interests a shopper may have based of a given interest. From our Jaccard similarity analysis we see that consumers who are interested in arts/entertainment also tend to be interested in news and people/society, and those interested in the news also tend to be interested in people/society and reference. This is logical as people who tend to follow the news and are interested in entertainment and often times are also interested in the lives of celebrities and popular culture. Therefore one suggestion for Valassis could be to focus more strongly on the arts/entertainment as these consumers are more responsive to ads and we also have some insight into what some of their other interests could be. Overall, our model performed quite well, with an average of 70%+ accuracy, and can help marketers optimize their ad spending by understanding which segments of the population they should focus their resources on.