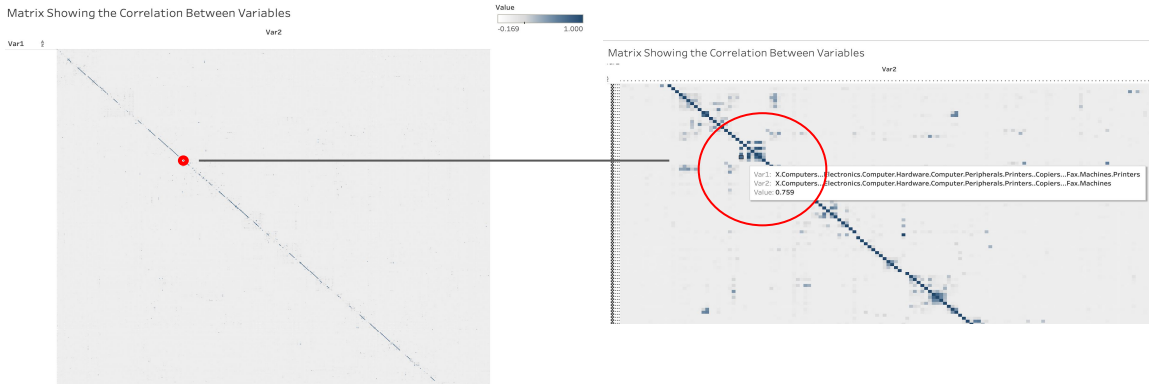
Predicting Shopper Conversion For Valassis

Daniel Sprague, Eric Tay, Jane Zhang, Ethan Shen

Team 75

EDA & Data Cleaning

- Initially focused on LT interests
- Heat Map
  - Subcategories with same category are correlated
- Transforming Data
  - Dictionary to vector
  - Cleaning data
  - Normalized data

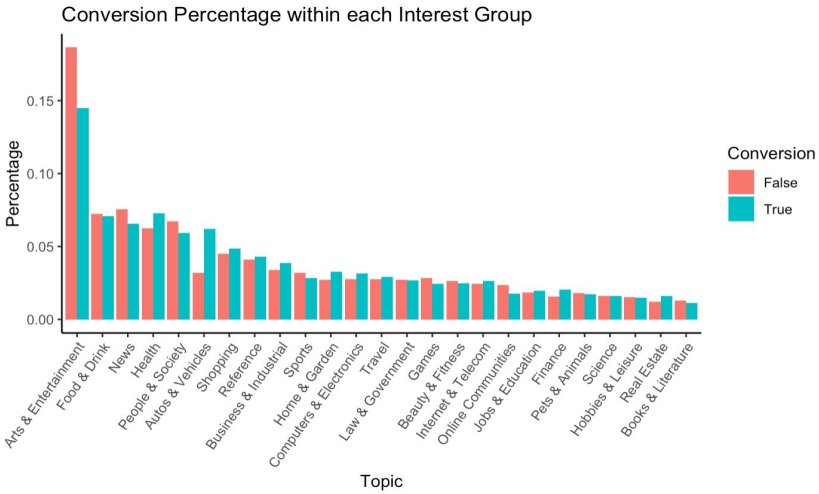


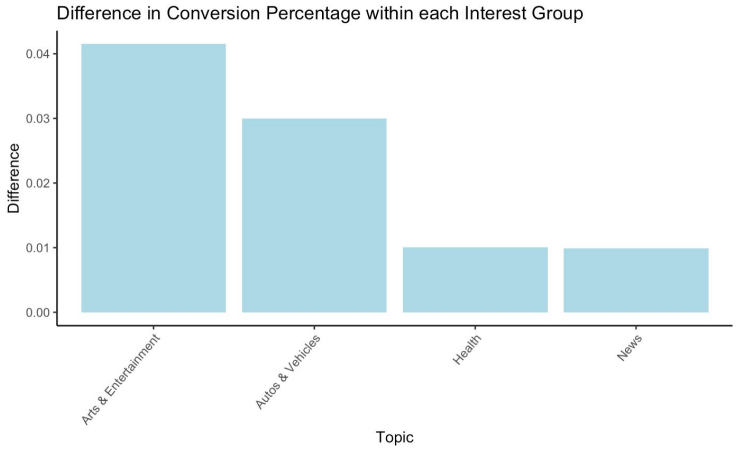
	userID	inAudience	ltiFeatures	stiFeatures
0	1	True	{'45': 0.020536141517834786, '47': 0.003117529...	{}
1	2	True	{'45': 0.001158253110658664, '592': 0.01546380...	{}
2	3	True	{'908': 0.002470851264264668, '590': 0.0021402...	{}
3	4	True	{'1187': 0.001127974558171163, '1780': 0.00117...	{}
4	5	True	{'907': 0.025339209040149392, '1187': 0.006020...	{'907': 0.10445132121076425, '908': 0.05651522...

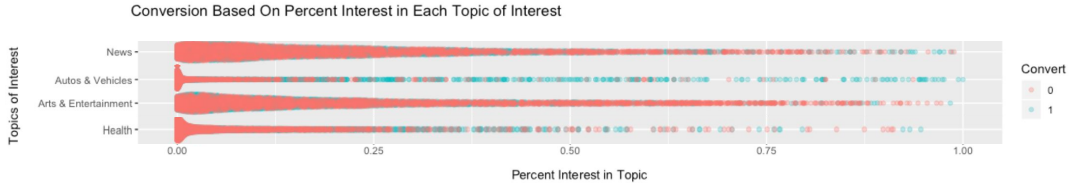
	Arts & Entertainment LTI	...	Pets & Animals LTI	Travel LTI	Food & Drink LTI	Reference LTI	Science LTI	Convert
0	0.017921	...	0.00000	0.023954	0.000000	0.017892	0.000000	1
1	0.099311	...	0.00000	0.000000	0.221425	0.001066	0.001579	1
2	0.224136	...	0.00891	0.009268	0.076458	0.032466	0.005204	1
3	0.046099	...	0.00000	0.000000	0.006055	0.015876	0.000000	1
4	0.000000	...	0.00000	0.000000	0.000000	0.000000	0.000000	1

EDA & Data Cleaning

- Visualizations
  - Signalled that interests such as Arts & Entertainment, Autos & Vehicles, News and Health could have significant predictive power

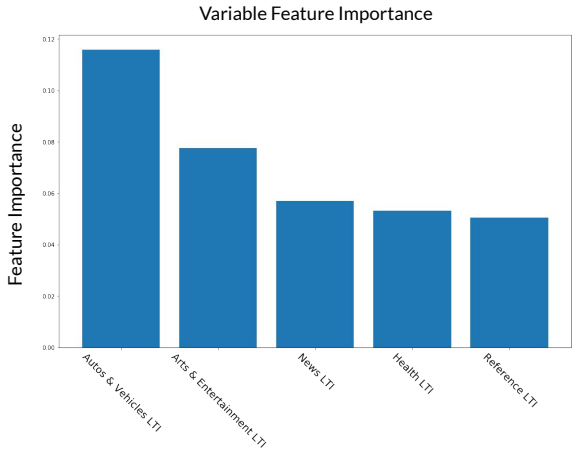


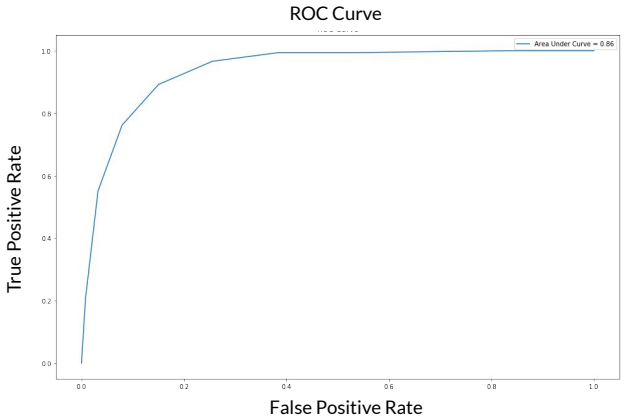




Model Selection, Findings & Improvements

- Model 1
  - Focused on LTI data for both training and testing
  - Resampled data to account for class imbalance
  - Tried boosted decision trees, SVMs, random forests (10 trees)
  - Picked random forests! (Best performance, overfitting, ease of tuning, variable importance)
  - Implemented a 3-fold Cross-Validation (CV)
  - Training: Accuracy - 0.637, SD - 0.0103
  - Test: Accuracy - 0.722, AUROC - 0.65
  - Findings match EDA, which predicted Arts & Entertainment,
- Model 2
  - Focused on data with STI features, reformatted vector, and ran same process as model 1
  - Training: Accuracy - 0.650, SD - 0.0117
  - Test: Accuracy - 0.743, AUROC - 0.86
  - STI adds enormous predictive power and significantly decreases false positives
  - 2 different models for Valassis to use





Final Model & Conclusion

- Model 3
  - Retrained model that included data with LTI but no STI data: Suboptimal results with significant FPR
  - Implemented 3-fold nested cross validation
  - Optimized hyperparameters for random forest
    - Optimized for maximal AUROC
  - Training accuracy increased from 0.6153 ± 0.011 to 0.6399 ± 0.003
  - AUROC increased from 0.62 to 0.64