Valassis User Conversion Analysis and Model

Duke Datathon 2019

Introduction

Specializing in advertising technology, Valassis works to predict a shopper's purchasing intent based on previous topics of interest to better target specific ads for each person. Any event when a shopper clicks on the ad, registers for an account, or purchases a product is known as a conversion. Because conversion events are extremely rare, it is critical to properly identify shoppers who convert in order to best utilize one's marketing budget. By better identifying shoppers who are more likely to respond to specific ads, Valassis can optimize their ad spending such that customer acquisition is maximized.

The focus of this project was to predict shoppers who are likely to convert. For this problem, we aimed to minimize false positives. The knowledge gained from this analysis will allow Valassis to better understand the conversion behavior of different consumer groups. Understanding consumers will pave the way for the company to maximize customers' return on advertising spending (ROAS) and consequently increase its market share in the marketing technology space.

Engineering Process

A general workflow of our process is shown in the following figure (Figure 1).

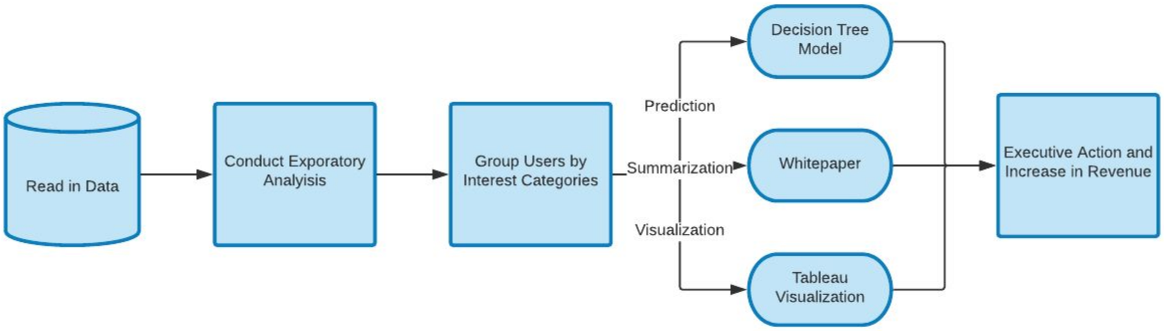
Read in Data
Conduct Exporatory Analyisis
Group Users by Interest Categories
Prediction
Summarization
Visualization
Decision Tree Model
Whitepaper
Tableau Visualization
Executive Action and Increase in Revenue

Figure 1: Data Engineering Process to Predict Shoppers who Convert

The imported training data was first restructured and cleaned in order to simplify the analysis. All "NA" values were assigned a value of 0 since these values were assumed to indicate no interest in the specific topic by the user. No other values were adjusted or normalized. The individual topic IDs were collapsed into larger categorical IDs. For instance, all "Arts & Entertainment" subtopics, such as "Performing Arts" and "Movies," were collapsed into an overall category of "Arts & Entertainment." Topic IDs present in the dataset that were not listed in the interest topics code book were classified into a category called "Unknown." This resulting dataframe was reshaped so that each row represents a unique userID with long and short term interest scores for each category. The columns consisted of the levels of long-term and short-term interest in all of the main categories. These steps were also applied to the validation set data in order to ensure consistency during model testing and validation.

Analysis

For predictor selection, we built a Random Forest classifier in Python with 250 trees optimized on minimizing the entropy in each tree. The Random Forest Classifier allows us to evaluate the categories that most highly influence the conversion rate. We used a class weight based on the proportion of classes in the training dataset. These weights were used to compensate for the relatively low rate of conversion in the overall dataset. We then extracted the top 8 features based on their importance according to the results of the model (Figure 2). We then utilized a decision tree optimized on minimizing the average entropy in the leaf nodes. We set a maximum depth of 5 to maximize interpretability and minimize the rate of false positives. The model was optimized on the transformed training data and evaluated on the test data.

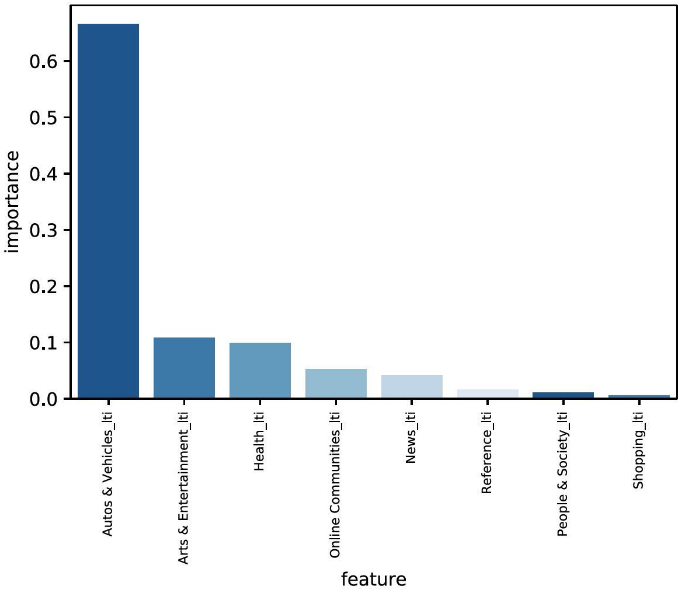

Figure 2: Most Important Features Determined by Model

Findings

The final model had a precision of 55% and F1-score of 52%. Based on our decision tree, we found that the most important topics to consider when predicting if a shopper will convert are long term interest in the following topics: Autos and Vehicles LTI, Arts and Entertainment LTI, Health LTI, Online Communities LTI, News LTI, Reference LTI, People and Society LTI, and Shopping LTI. Of these eight topics of interest, the top five were further analyzed, as these topics were more important in predicting conversion.

After analyzing the results of our decision tree classifier, we found that those who had higher long-term interest scores in Auto and Vehicle and Health topics were more likely to convert. However, those who had higher long-term interest scores in Arts and Entertainment and News were less likely to convert. This finding is also supported through our initial exploratory data analysis; we plotted these four topics against the average long-term interest score for both converters and non-converters. Clearly, for Auto and Vehicle and Health, the average LTI score for converters is higher than that for non-converters (Figure 3). Likewise, for News and Arts and Entertainment, the average LTI score for converters is lower than that for non-converters (Figure 4).

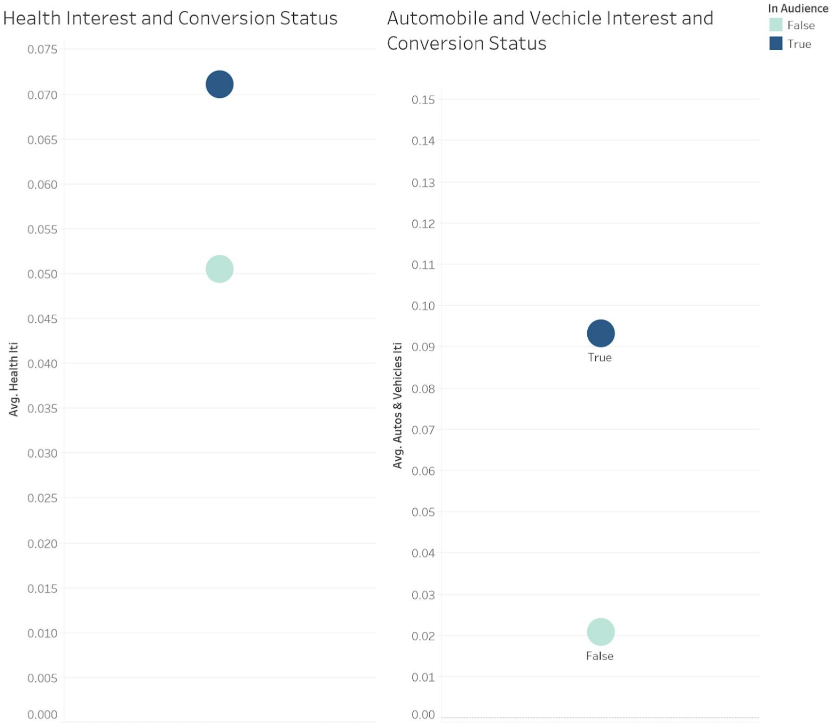

Figure 3: Average long-term interest scores for converters and non-converters in the Health and Automobile and Vehicle interest groups. On average, converters (dark blue) have higher LTI scores than non-converters (light blue) in these two interest categories.

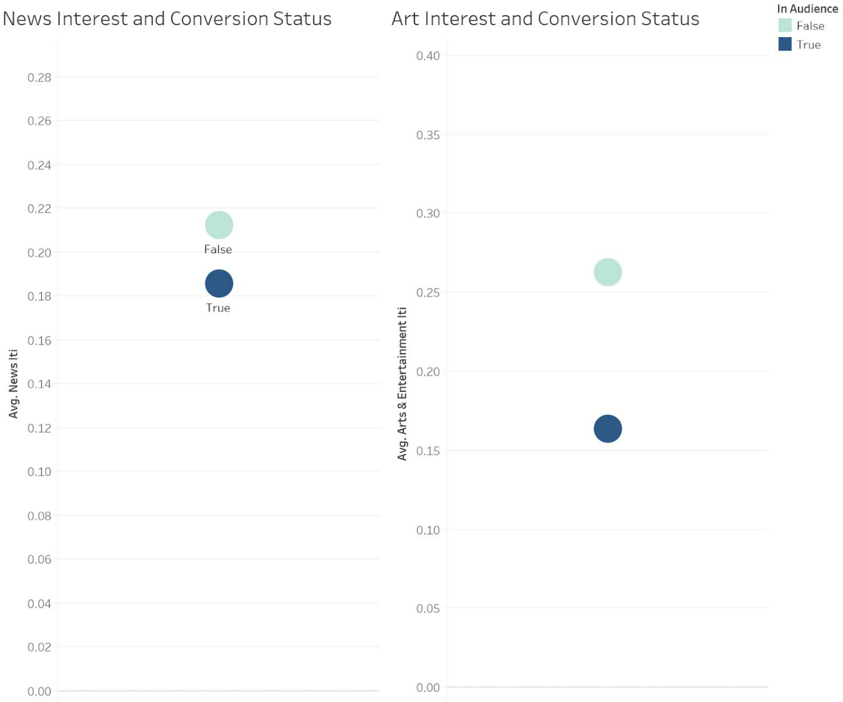

Figure 4: Average long-term interest scores for converters and non-converters in the News and Art and Entertainment interest groups. On average, non-converters (light blue) have higher LTI scores than converters (dark blue) in these two interest categories.

Conclusion

After thorough engineering of the data to facilitate analysis and model building, we found the top eight most important features in predicting conversion by building a random forest model. Using the most important features, we then constructed a decision tree to predict a shopper's conversion given a set of interest categories and interest scores. Overall, we found two interest categories, Automobile and Health, where higher interest scores generally lead to a higher likelihood of conversion. On the other hand, we found two interest categories, Arts and News, where higher interest scores generally lead to a lower likelihood of conversion.

Based on this analysis, Valassis has an opportunity to more accurately identify shoppers who are most likely to convert. This project's segmentation of consumers by interest level reveals specific interest categories that provide the greatest potential for revenue, as well as interactions between an individual's varying interests. Implementation of this model will increase shopper's interaction and conversions with more targeted ads, thereby increasing Valassis' effectiveness. The model's application will be reflected in advertising performance, which will increase the ROAS for partnering organizations and ultimately increase revenue for Valassis overall.