Consumer Data Research Centre

An ESRC Data Investment

SHOP DIRECT

Propensity modelling – moving beyond standard regression models & explorations into online engagement

Mary E. Lennon[1], Guanpeng Dong[1] and Tony Birch[2]
[1]University of Liverpool, [2]Shop Direct

Project Background

Online retailers (e-tailers) have not only cast aside brick-and-mortar locations but also established new methodologies for customer focused initiatives. In their place they are exploring Big Data and associated statistical techniques and machine learning algorithms, to understand their customer's engagement. Broadly, engagement is a customer's spend (transactional) and emotional (non-transactional) commitment to a brand that can be quantified for customer acquisition and retention efforts. However, engagement tends to be ethereal and is difficult to define and leverage. This dissertation has consequentially situated itself between theoretically defining and using engagement for statistically driven customer focused initiatives. To achieve these goals, this study had two aims:

1. Explore random forest, a machine learning alternative to traditional regression framework, for propensity modelling in the context of online engagement.

2. Improve current theoretical models on engagement through the addition of individual contextual information. Specifically, through the inclusion of demographic and geo-demographic variables.

Data and Methods

This study primarily used data from Shop Direct (SD), a prominent UK e-tailer, specifically, data from their current engagement model. Each observation was an active customer account between 1st January – 1st April 2017, for their largest brand, totaling 2.2 million observations. This study also considered the age and gender of the customer's from SD's customer account file. The Internet User Classification (IUC) and Retail Attractiveness Scores supplemented SD's internal data to test the effects of geographic contexts, such as proximity to retail centres, on customer engagement. The data were included in a derivation of the random forest algorithm entitled 'Bag of Little Bootstraps' (BLB). BLB was coded in R and employed to tackle the memory and computational issues that arose when analysing SD's Big Data. The technique saves time and memory by running optimised bootstrapping across parallel cores on a computer while producing results that are as robust and reliable as the traditional random forest algorithm. To assure robustness, the results were cross validated.

Key Findings

This study had three key findings. First, BLB random forest offered notable improvements to the logistic regression baseline, with an average increase in performance of 1%. Second, the results were robust. The model error rates had a fluctuation of +/- 0.15% between each validation test. Third, the top performing model was that which included all available demographic and geo-demographic variables indicating that individual context matters. Beyond model performance, logistic regression's coefficients and random forest's variable importance shed light on the relationship of the variables to engagement. Key among these are that proximity to retail centres affects a customer's propensity to engage with an e-tailer, such as SD. Further, there is complicated relationships between IUC and the SD data. Customers in all IUC groups except for the most unlikely, elderly and unconnected, had a negative relationship with customer engagement. The results for IUC may be muddled due to the varying nature of the relationships caught by both the classification system and engagement model. For effective use of IUC additional exploration of these relationships is necessary. Regardless of the nature of the relationship, IUC and primary retail catchments were important contributions to the top performing model.

Logistic Regression v. Random Forest

Variables	*Model No.*	*Training*	*Testing*
SD Variables	M_1	0.2284	0.2273
SD Variables + Demographics	M_2	0.2281	0.2273
SD Variables + IUC	M_3	0.2289	0.2278
SD Variables + Retail Catchments	M_4	0.2289	0.2276
SD Variables + Demographics + IUC + Retail Catchments	M_5	0.2282	0.2270

Table 1. Comparing Model Performance

Value of the Research

Online engagement is critical to many e-tailers. Despite this, theory on how to measure it is limited and non-theoretical examples of its implementation are currently unknown. This project offers a starting point for crucial discussions into defining and predicting engagement. Further, it has proven the potential of machine learning techniques as well as the benefits of demographic and geodemographic contextual information to engagement models. They provide notable improvements in model performance and increase insights into consumer behaviours.