Consumer Data Research Centre

An ESRC Data Investment

Sainsbury's

Predicting customer quality based on early shopping behaviour in online grocery retail

Sophie de Kok[1], Daniel Hulme[1] and Graham Johnson[2]
[1]University College London, [2]Sainsbury's

Project Background

Online grocery shopping is growing rapidly. However, there are almost no barriers for customers to switch to competitors which results in a high churn rate. Predicting customer quality can be used to target and retain valuable customers and to adapt marketing spend based on a customer's quality. This research focuses on predicting quality in terms of customer lifetime value and churn based on a customer's early shopping behaviour. It is a challenge to obtain reliable predictions of the quality of new customers based solely on their first few shops. However, being able to predict customer quality as early as possible is important for businesses to adapt their marketing spend accordingly. We also investigate the trade-off between the amount of data used and the predictive performance of the models. Lastly, a customer segmentation is created to understand the different type of customers.

Data and Methods

The data used in this research has been made available by Sainsbury's. It consists of 209,592 transactions from 24,319 unique customers from orders placed between March 2010 and May 2018. For each online transaction, purchase information is available and so-called RFM features are computed. To investigate the trade-off between the amount of data used and the model's performance, five subsets were created. Three subsets were based on the order number, e.g. a subset containing only the first order of each customer. The other two subsets were based on the number of days, e.g. all orders placed within 14 days after their initial purchase.

To predict customer lifetime value and churn, linear/logistic regression, random forest, SVM, XGBoost and a neural net were implemented. In addition, these individual models were combined into a stacking model to increase the performance. Lastly, a two-stage algorithm was created which combined the churn and CLV predictions to improve the performance even further. Each model was trained on the various data subsets. To determine the customer segmentation, LRFMP variables were computed and a K-means algorithm was employed to find the various segments.

Key Findings

Accurate predictions of customer quality were in fact found using the early shopping behaviour of customers. The neural net was the best performing individual machine learning model for predicting both customer lifetime value and churn. Moreover, the stacking model significantly outperformed each individual model. However, the two-stage method only had a very minimal performance gain compared to the stacking model which is not justified by the extra amount of training required. Therefore, the stacking model is preferred when predicting customer quality. We also found that the optimal amount of data depended on the outcome variable and the aim of the company. For example, the mean absolute error of the customer lifetime value predictions decreased by 14% when looking at 30 days of data instead of 14 days. However, the error only decreased by 0.01% when employing 14 days of data compared to one day. Churn prediction gained most by employing more data, with the accuracy improving from 78% to 85% when looking at 30 days of data instead of 14 days. Lastly, distinct customer segments were found using K-means clustering. The customer segmentation is visualised below using t-SNE.

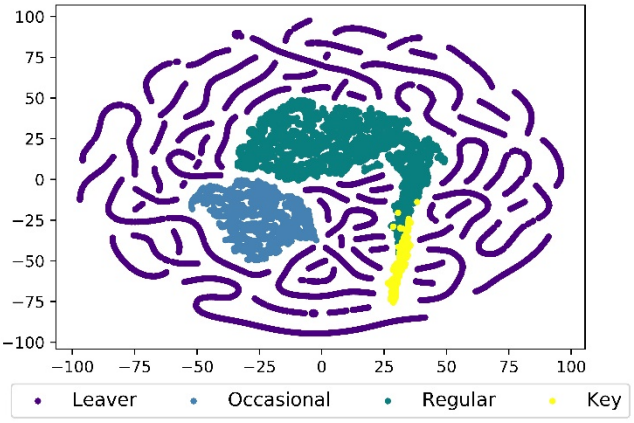

Figure 1. K-Means customer segmentation

Note: these results are not representative of Sainsbury's customer mix as specific sample data was used.

Value of the Research

Being able to get an early estimate of customer quality is very valuable to determine how to distribute marketing spend and reduce unnecessary marketing costs by differentiating between customers. Moreover, these predictions can be employed as a preliminary performance measure of marketing campaigns and help optimise demand forecasting. This research is not only valuable for online grocery retailers, but also for other online retailers aiming to predict customer quality based on early shopping behaviour.