Capstone project summary

Customer segmentation analysis

Lisa Lang
Student ID 1902224
June 16, 2020

General Background

The client is a musical production company which owns and operates three historic theater venues and produces several musicals that are hosted all over the world.

As part of their digital marketing strategy, VBW creates email campaigns to raise interest for their upcoming shows and boost ticket sales. The client provided me with three data files that had records from their customers, from their historic transaction back to 2007 and event data and asked me how they could segment their customers in order to refine their marketing strategies.

The goal of this analysis is to distinguish groups of customers from the mass in terms of their value to the company and in terms of their purchasing behaviour/preferences.

Methodology

I used three approaches to define the customers in more detail.

Descriptive Analytics

I used **descriptive analysis** to find general patterns in the data and find metrics and thresholds to form larger groups, for example business customers versus individual customers, local guests versus guests from abroad, etc.

RFM Model

Then I used an **RFM model** to assign an RFM score to each customer. The RFM score is used to rank and categorize each customer and thus to identify the most valuable customers.

The RFM score is computed by concatenating three individual scores: the **r**ecency score, the **f**requency score and the **m**onetary score. These three scores are assigned based on the distribution of the most recent purchase date, the frequency of purchases by each customer and the total revenue generated by each customer. Customers who purchased very often receive a high frequency score, one-time-customers receive a very low score.

Based on the RFM score the customers are grouped into segments, that are labelled in a way to understand what status they have within the company (for example “Champions”, “Need Attention”, “Lost”, …).

RFM App

In order to be able to reproduce the scoring and segmentation results with new customer data, or even subgroups of the customer data, I created an interactive web application which allows the user to upload a dataset with transaction data and compute an RFM score for each customer ID. Furthermore based on the computed RFM score, segment labels are assigned to each customer. The results can be downloaded as a .csv file.

I created this app using 'Shiny', which is an R package for building interactive web applications straight from R.

The app is currently hosted as a web page on the RStudio server shinyapps.io and can be accessed through the weblink.

Clustering Model

Then I used an unsupervised machine learning model to cluster the customer data based on their R, F an M scores, demographic and product features. Clustering algorithms are designed to identify groups in data and my aim was to refine the “customer archetypes” that were previously generated.

The basic premise of clustering is the similarity/dissimilarity between the data points. The algorithm iterates over and over until it attains a state wherein all points of a cluster are similar to each other, and points belonging to different clusters are dissimilar to each other. This similarity/dissimilarity is defined by the distance between the points.

A popular choice for the metric to measure this distance is the Euclidean distance, but

this distance metric is applicable only for continuous data, not for categorical data I used for this analysis.

In addition I needed to use an algorithm that was feasible for large data sets.

For this analysis I used the R package clustMixType (Szepannek, 2018), based on the idea of Huang's k-prototypes algorithm (Huang, 1998), which was able to compute the distance on a large data set with mix-type data.

Results

The results presented to the client were an extensive **description of their data**, including graphical representations that were able to visualise the **changes in purchase behaviour over time**.

The customer database was returned including an **RFM score** for each customer, a **segment label** and a **cluster association** from which the marketing team is able to select subgroups for specific marketing actions.

The segments and the cluster memberships were described in detail and **recommendations** were provided on **how to address the individual subgroups** as part of a global marketing strategy.

The **RFM app** that is currently hosted on an R shiny server will be transferred to the clients server in the near future.

Lessons Learned

During all the steps conducting the analysis I would like to point out the following lessons I learned:

- Get the data as early as possible to have more time available to get familiar with the data during the cleaning and descriptive analytics process
- Involve experts early to discuss your Analysis plan. They can point out weaknesses early or spark ideas to improve the project
- Regularly schedule short meetings with the client to evaluate intermediate results and to be alerted early when going into the wrong direction with regards to expectation management and/or interpretation of results
- Set seeds