

Clustering Based Recommendation System

Team: BoxCox (Andrew, Dipesh, Layla, Radhika, Tushar)

Preliminary Analysis: More advertisement clicks and higher time spent on the website result in greater generation of leads.

Research Question: Can we segment (cluster) the visitors based on multiple parameters (demographics and psychographics) to better understand customer preferences and make real-time recommendations (even for first-time visitors)?

Outside Data: 1. Average income based on ZIP Code (www.irs.gov)
2. Average gas prices based on state (www.GasBuddy.com)

The income, together with the gas price for a region, gives an estimate of the **purchasing power** of a visitor from that region, and is a stronger parameter (when taken together), than just location, to influence visitor behavior on the Edmunds website.

Method: To segment the visitors, a **k-means clustering** approach was implemented. k-means clustering aims to partition the n observations into k ($\leq n$) sets $S = \{S_1, S_2, \dots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS). The formula is:

$$\arg \min_{S} \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2$$

where μ_i is the mean of points in S_i .

Since the initial center of the mean is random, the algorithm was performed 10 times over and averaged to get a stable clustering. The clustering was validated using ANOVA. The parameters used to create the clusters are

1. Platform used to access the Edmunds website (desktop, tablet, mobile)
2. First entry channel to the Edmunds website (free/paid/bookmarked)
3. Types of cars interested in (New/Used/Certified Pre-owned/Pre-production)
4. Income based on average income for the region [*Outside Data*]
5. Gas price based on state [*Outside Data*]

Result: Implementing the above method created 15 clusters. These clusters are based on both, the demographic and psychographic data, and can be utilized to better understand customer preferences and make real-time recommendations on **targeted advertising, targeted advice pages** etc. and also to find **possible problematic clusters**.

The following insights and recommendations are made:

1. Visitors within some clusters have a greater preference for luxury cars.
2. Visitors interested in luxury cars spend lesser time on the website.
3. Certain clusters have visitors spending very less time on the website and need more focusing to increase leads from those clusters

NOTE: This is just an example of the **immense possibilities** of our clustering algorithm in enhancing user experience on Edmunds.com.