Giant Eagle

Curbside Express Customer Retention

Project Documentation

Team Members:

Aditi Mahajan, Ankush Batra, Arushi Vyas, Darshit Doshi,

Dhara Shah, Kartik Joshi, Yamuna Bhaskar Surendra Babu

Table of Contents

1. Dataset discussion
2. Clustering model for customer segmentation
3. Linear Regression model for curbside likelihood
4. **Dataset discussion**
5. ‘Curbside\_consolidated\_household.cvs’ is the dataset used for the K-means Clustering model to segment curbside customers. Below table indicates the attributes contained in the file and their definitions.

|  |  |  |
| --- | --- | --- |
| **No.** | **Attribute Name** | **Description** |
| 1 | HSHLD\_NO | Household number |
| 2 | TOTAL\_CURBSIDE\_SPEND | Amount spent on curbside transactions |
| 3 | TOTAL\_NON\_CURBSIDE\_SPEND | Amount spent on non-curbside transactions |
| 4 | TOTAL\_SPEND | Total amount spent on curbside and non-curbside transactions |
| 5 | NON\_CURBSIDE\_TRANS | Number of non-curbside transactions |
| 6 | Count\_Curbside\_Trans | Number of curbside transactions |
| 7 | CURBSIDE\_PICKUP\_% | Percentage curbside transactions that are pickup |
| 8 | Children\_0\_5 | Number of children up to 5 year old |
| 9 | Children\_6\_10 | Number of children between 6 to 10 years of age |
| 10 | Children\_11\_17 | Number of children between 11 to 17 years of age |
| 11 | No\_Of\_Children | Total number of children |
| 12 | Adult\_18\_34 | Number of adults between 18 to 34 years of age |
| 13 | Adult\_35\_64 | Number of adults between 35 to 64 years of age |
| 14 | Adult\_65\_75plus | Number of adults above 65 years |
| 15 | No\_Of\_Adults | Total number of children |
| 16 | Household\_Size | Size of household |
| 17 | Income\_Buckets | Income bucket the household falls under |
| 18 | Age | Age of first individual |
| 19 | EDUCATION | Education of first individual |
| 20 | CURBSIDE\_SPEND\_ALCOHOL\_TOBACCO | Amount spent on curbside for alcohol and tobacco |
| 21 | NON\_CURBSIDE\_SPEND\_ALCOHOL\_TOBACCO | Amount spent on non-curbside for alcohol and tobacco |
| 22 | CURBSIDE\_SPEND\_FRESHFOOD\_DESSERT | Amount spent on curbside for fresh food and dessert |
| 23 | NON\_CURBSIDE\_SPEND\_FRESHFOOD\_DESSERT | Amount spent on non-curbside for fresh food and dessert |
| 24 | CURBSIDE\_SPEND\_MEAT\_DAIRY | Amount spent on curbside for meat and dairy |
| 25 | NON\_CURBSIDE\_SPEND\_MEAT\_DAIRY | Amount spent on non-curbside for meat and dairy |
| 26 | CURBSIDE\_SPEND\_HOUSEHOLD\_GROCERIES | Amount spent on curbside for household groceries |
| 27 | NON\_CURBSIDE\_SPEND\_HOUSEHOLD\_GROCERIES | Amount spent on non-curbside for household groceries |
| 28 | CURBSIDE\_SPEND\_OTHER | Amount spent on curbside for other products |
| 29 | NON\_CURBSIDE\_SPEND\_OTHER | Amount spent on non-curbside for other products |
| 30 | CURBSIDE\_SPEND\_GROCERY\_EDIBLE | Amount spent on curbside for edible grocery |
| 31 | NON\_CURBSIDE\_SPEND\_GROCERY\_EDIBLE | Amount spent on non-curbside for edible grocery |
| 32 | S\_NumCompStoresIn05Miles | Number of competition stores within 5 miles radius |
| 33 | S\_NumNationalStoresIn05Miles | Number of national stores within 5 miles radius |
| 34 | S\_NumLocalStoresIn05Miles | Number of local stores in 5 miles radius |
| 35 | S\_isCORP | Number of corporate stores |
| 36 | S\_isGiantEagle | Number of Giant Eagle stores |
| 37 | curbside\_Pickup\_fee | Fee paid for curbside pickup |
| 38 | curbside\_Delivery\_fee | Fee paid for curbside delivery |

1. ‘Consolidated\_household.csv’ is the dataset used for the linear regression model to predict the likelihood of non-curbside customers starting curbside. Below table indicates the attributes that are used for modeling and their definitions.

|  |  |  |
| --- | --- | --- |
| **No.** | **Attributes** | **Description** |
| 1 | TOTAL\_CURBSIDE\_SPEND | Amount spent on curbside transactions |
| 2 | TOTAL\_NON\_CURBSIDE\_SPEND | Amount spent on non-curbside transactions |
| 3 | NON\_CURBSIDE\_TRANS | Number of non-curbside transactions |
| 4 | CURBSIDE\_TRANS | Number of curbside transactions |
| 5 | DIM\_2\_CD | Value segmentation score |
| 6 | Children\_0\_5 | Number of children up to 5 year old |
| 7 | Children\_6\_10 | Number of children between 6 to 10 years of age |
| 8 | Children\_11\_17 | Number of children between 11 to 17 years of age |
| 9 | Adult\_18\_34 | Number of adults between 18 to 34 years of age |
| 10 | Adult\_35\_64 | Number of adults between 35 to 64 years of age |
| 11 | Adult\_65\_75plus | Number of adults above 65 years |
| 12 | Household\_Size | Size of household |
| 13 | Income\_Buckets | Income bucket the household falls under |
| 14 | EDUCATION | Education of first individual |
| 15 | Age | Age of first individual |
| 16 | Vehicle\_Owned | Boolean to indicate if a household owns a vehicle |
| 17 | Value\_Score | Value score of a household |
| 18 | Promo\_Sensitive? | Promotional Sensitivity |

1. **Clustering model for customer segmentation**
2. Identify the number of clusters using Ward’s Method:

* Load the file.
* Go to Analyze -> Classify -> Hierarchical Clusters.
* Go to Method -> Use ‘Ward’s Method’ as the clustering method.
* Go to Plots -> Select ‘Dendrogram’ as an output.
* Using the dendrogram, identify the number of clusters(4 for our segmentation).

1. Perform K-means clustering:

* Go to Analyze -> Classify -> K-means clustering
* Input the number of clusters as ‘4’.
* Under ‘Options’, select ‘ANOVA Table’.
* Under ‘Save’, select ‘Cluster Membership’.
* Review the ANOVA Table to identify which factors are significant (i.e., factors that have a significance level of <0.05).

1. Profile clusters:

* Identify differentiating factors from ANOVA table
* Review demographic information for each cluster
* Verify that demographic information aligns with values of differentiating factors
* Test clusters for business sense

1. Iterate from step a. if needed