Analyses of Market retail promotion for Kroger.

Introduction:

Kroger is an American retail company which is considered as one of the largest super markets of United states. The data we have is of around 102 weeks starting from march 2005 to February 2007, of around 800 households using loyalty cards to shop, our main goal is to analyze the data and help Kroger make maximum profit by the promotional campaigns and increase the number of customers with loyalty cards which would impact on the overall sales. So far we were finding the impact of R-F-M variables now we have a new metric- D.

Recency(R) – it is the duration from customers last purchase to current purchase, we want to minimize this as much as possible to increase overall sales.

Frequency(F) – it is the number of visits per household per months, we want to increase this to boost overall sales.

Monetary Value(M) – monetary value is the average spending per customer over a month, we need to maximize this as much as possible, I have used spending average which an average of current weekly spending and 3 previous transactions by the same household.

D (D)- log of number of weeks for each house id since the first campaign, I have calculated this based on the NumAll\_Now variable which is a summation of all the campaigns present within a given calendeweek a particular household id is exposed to.

After calculating and taking log of it, its histogram looks as below,

Chart

Description automatically generated with low confidence

I have filtered out all the zero/ missing values from the dataset.

Part 1: **Response Model with Fixed Effects**

**we will be checking the effect of R-F-M-D and other covariate variables along with fixed effect of every household and Campaigns A, B & C.**

I have built 2 models

**Model 1 –** consisting of RFMD, covariates and household ID as fixed effect.

**Model 2 –** consisting of RFMD, covariates and Campaign A, B&C as fixed effect.

|  |  |  |
| --- | --- | --- |
|  | *Dependent variable:* |  |
|  |  |  |
|  | weeklyspend |  |
|  | -1 | -2 |
|  |  |  |
| recency | -0.396\*\*\* (0.102) | -0.337\*\*\* (0.089) |
| freq | -12.847\*\*\* (0.227) | -12.524\*\*\* (0.198) |
| spendingavg | 0.993\*\*\* (0.006) | 0.988\*\*\* (0.004) |
| d | -0.354\* (0.193) | -0.385 (0.249) |
| t\_loydisc\_cat | 0.132\*\* (0.064) | 0.381\*\*\* (0.040) |
| t\_display | 0.788\*\*\* (0.045) | 0.668\*\*\* (0.041) |
| t\_favbrndcoup | 0.067\*\* (0.034) | 0.014 (0.052) |
| t\_mailer | 2.003\*\*\* (0.044) | 2.016\*\*\* (0.041) |
| hhsize | 5.968 (10.006) | -0.634 (0.436) |
| numkids | 2.775 (8.057) | 0.179 (0.496) |
| income | 0.0001 (0.0001) | 0.00000 (0.00000) |
| married | -20.464 (18.764) | 0.358 (0.430) |
| Camp\_A |  | 0.862 (0.916) |
| Camp\_B |  | -0.241 (0.646) |
| camp\_C |  | 0.354 (0.844) |
| age | 0.717 (0.658) | 0.029\*\* (0.014) |
| id7 | 12.954 (10.134) |  |
| id8 | -4.994 (6.074) |  |
| id2499 |  |  |
| Constant | -43.090 (45.260) | 6.112\*\*\* (1.271) |
|  |  |  |
| Observations | 73,863 | 73,863 |
| R2 | 0.628 | 0.625 |
| Adjusted R2 | 0.624 | 0.625 |
| Residual Std. Error | 43.191 (df = 73056) | 43.153 (df = 73846) |
| F Statistic | 153.257\*\*\* (df = 806; 73056) | 7,693.058\*\*\* (df = 16; 73846) |
|  |  |  |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 |  |

Looking at the model summary, R-F-M-D coefficients look consistent in both the models.

R square is also close to 62% meaning both models are equally good fit.

**Part 2. Exploring K-type of Heterogeneity**

**I have implemented 2 types of segmentation method:**

1. **3 segmentation : using the same code shared by professor I have created 3 clusters and implemented similar fixed effect models as in the above for each cluster.**

**Clusters(datasets) are of below dimensions:**

|  |  |
| --- | --- |
| **Seg3data1** | **23318** |
| **Seg3data2** | **16082** |
| **Seg3data3** | **34462** |

1. Below are the outputs of **fixed effect household id** model.

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Dependent variable:* |  |  |
|  | weeklyspend |  |  |
|  | -1 | -2 | -3 |
|  |  |  |  |
| recency | -0.580\*\*\* (0.112) | -1.488\*\*\* (0.498) | -2.757\*\*\* (0.345) |
| freq | 2.336\*\*\* (0.322) | 4.066\*\*\* (0.595) | 3.177\*\*\* (0.413) |
| d | -0.700\* (0.405) | -0.495 (0.487) | 0.383 (0.339) |
| t\_loydisc\_cat | 0.139 (0.107) | -0.065 (0.168) | 0.252\*\* (0.125) |
| t\_display | 0.738\*\*\* (0.084) | 0.660\*\*\* (0.113) | 0.821\*\*\* (0.083) |
| t\_favbrndcoup | 0.033 (0.071) | 0.283\*\*\* (0.081) | 0.052 (0.060) |
| t\_mailer | 2.144\*\*\* (0.081) | 2.074\*\*\* (0.110) | 2.070\*\*\* (0.083) |
| hhsize | 10.310 (9.726) | 111.329\*\*\* (15.028) | 34.541 (27.876) |
| spendingavg | 0.988\*\*\* (0.006) | 0.989\*\*\* (0.005) | 0.991\*\*\* (0.001) |
| numkids | -12.416 (13.082) | 50.259\*\*\* (12.258) | -67.243\* (35.023) |
| income | -0.0002 (0.0002) | 0.003\*\*\* (0.0001) | -0.001\*\*\* (0.0002) |
| married | -43.672 (32.020) | 170.633\*\*\* (10.519) | 47.485\*\*\* (17.567) |
| age | 1.168 (0.928) | 1.514\*\*\* (0.428) | 2.611\*\* (1.308) |
| as.factor(id)2 | 27.093 (22.188) |  |  |
| as.factor(id)5 | -27.264 (34.850) |  |  |
| as.factor(id)4 |  | -243.181\*\*\* (16.292) |  |
| as.factor(id)7 |  | 326.139\*\*\* (22.187) |  |
| as.factor(id)8 |  |  | 34.858\* (20.856) |
| as.factor(id)17 |  |  | 173.564\*\*\* (33.726) |
| as.factor(id)24 |  |  | 196.270\*\*\* (47.609) |
| Constant | -46.651 (39.874) | -508.656\*\*\* (52.738) | -97.737 (84.216) |
|  |  |  |  |
| Observations | 23,318 | 16,082 | 34,462 |
| R2 | 0.5784 | 0.601 | 0.598 |
| Adjusted R2 | 0.586 | 0.602 | 0.596 |
| Residual Std. Error | 40.838 (df = 23061) | 54.505 (df = 15899) | 55.670 (df = 34082) |
| F Statistic | 68.719\*\*\* (df = 256; 23061) | 70.943\*\*\* (df = 182; 15899) | 77.212\*\*\* (df = 379; 34082) |
|  |  |  |  |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 |  |  |
|  |  |  |  |

1. Below are the fixed effects considering all 3 Campaigns.

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Dependent variable:* |  |  |
|  | weeklyspend |  |  |
|  | -1 | -2 | -3 |
|  |  |  |  |
| recency | -0.076 (0.082) | -1.249\*\*\* (0.388) | -1.582\*\*\* (0.271) |
| freq | -12.038\*\*\* (0.268) | -12.182\*\*\* (0.484) | -13.032\*\*\* (0.345) |
| spendingavg | 0.970\*\*\* (0.008) | 0.992\*\*\* (0.008) | 0.992\*\*\* (0.006) |
| d | -0.513 (0.425) | -0.599 (0.530) | -0.116 (0.374) |
| t\_loydisc\_cat | 0.344\*\*\* (0.056) | 0.261\*\*\* (0.095) | 0.424\*\*\* (0.067) |
| t\_display | 0.710\*\*\* (0.065) | 0.473\*\*\* (0.085) | 0.732\*\*\* (0.065) |
| t\_favbrndcoup | -0.003 (0.081) | 0.065 (0.116) | -0.027 (0.080) |
| t\_mailer | 2.049\*\*\* (0.064) | 2.020\*\*\* (0.085) | 1.932\*\*\* (0.066) |
| hhsize | 0.002 (0.595) | -0.582 (0.756) | -0.770 (0.945) |
| numkids | -0.262 (0.698) | 0.251 (0.966) | 0.165 (1.014) |
| income | 0.00001 (0.00001) | 0.00001 (0.00001) | 0.00000 (0.00000) |
| age | 0.033\* (0.019) | 0.006 (0.037) | 0.023 (0.024) |
| married | -0.490 (0.655) | -0.234 (1.851) | 1.203 (0.839) |
| Camp\_A | 0.543 (1.385) | 2.332 (2.147) | 0.679 (1.407) |
| Camp\_B | 0.102 (1.126) | -0.048 (1.387) | -0.372 (0.961) |
| camp\_C | 0.091 (1.693) | 0.675 (1.784) | 0.330 (1.205) |
| Constant | 3.589\*\* (1.732) | 5.075 (3.415) | 8.633\*\*\* (2.251) |
|  |  |  |  |
| Observations | 23,318 | 16,082 | 34,462 |
| R2 | 0.594 | 0.614 | 0.609 |
| Adjusted R2 | 0.594 | 0.613 | 0.609 |
| Residual Std. Error | 34.362 (df = 23301) | 45.367 (df = 16065) | 47.209 (df = 34445) |
| F Statistic | 2,132.492\*\*\* (df = 16; 23301) | 1,595.132\*\*\* (df = 16; 16065) | 3,352.481\*\*\* (df = 16; 34445) |
|  |  |  |  |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 |  |  |

Model with fixed effect using household ID fits better for cluster 2 with an adj. R^2 ~60 which is close to the fixed effect model of the overall dataset.

Similarly model with fixed effect using campaigns as fixed effect have a adj. R^2 ~61 which is closed to the model using entire dataset .

1. 4 Segmentation : **I have created 4 clusters and implemented similar fixed effect models as in the above for each cluster.**
2. Below are the outputs of **fixed effect household id** model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  | *Dependent variable:* |  |  |
|  |  | weeklyspend |  |  |
|  | -1 | -2 | -3 | -4 |
|  |  |  |  |  |
| recency | -0.107 (0.087) | -2.427\*\*\* (0.790) | -47.288\*\*\* (17.228) | -2.076\*\*\* (0.264) |
| freq | -14.014\*\*\* (0.296) | -16.052\*\*\* (0.845) | -8.263 (5.294) | -11.027\*\*\* (0.284) |
| spendingavg | 0.980\*\*\* (0.008) | 1.002\*\*\* (0.011) | 1.083\*\*\* (0.155) | 0.988\*\*\* (0.005) |
| d | -0.119 (0.466) | -1.386\* (0.822) | 8.725 (5.915) | -0.193 (0.311) |
| t\_loydisc\_cat | 0.310\*\*\* (0.067) | 0.302\* (0.156) | -0.302 (1.517) | 0.363\*\*\* (0.053) |
| t\_display | 0.809\*\*\* (0.074) | 0.561\*\*\* (0.135) | 2.280\*\* (1.032) | 0.583\*\*\* (0.052) |
| t\_favbrndcoup | -0.041 (0.091) | 0.007 (0.180) | -1.275 (1.347) | 0.031 (0.065) |
| t\_mailer | 2.204\*\*\* (0.072) | 2.421\*\*\* (0.141) | -0.945 (0.969) | 1.722\*\*\* (0.053) |
| hhsize | -0.220 (0.654) | 0.645 (3.345) | -16.082 (15.477) | -0.731 (0.554) |
| numkids | 0.092 (0.758) | -0.153 (3.391) |  | -0.320 (0.806) |
| income | 0.00000 (0.00001) | -0.00001 (0.00001) |  | 0.00001 (0.00000) |
| age | 0.068\*\*\* (0.022) | -0.022 (0.068) |  | 0.011 (0.019) |
| married | -0.120 (0.642) | 0.957 (1.977) |  | 0.696 (0.578) |
| Camp\_A | 0.268 (1.594) | 4.174 (3.305) | 10.674 (30.139) | 0.360 (1.150) |
| Camp\_B | -0.644 (1.223) | 2.438 (2.072) | 3.615 (14.481) | -0.769 (0.812) |
| camp\_C | 0.607 (1.716) | 0.350 (2.492) | -24.066 (17.164) | 0.645 (1.064) |
| Constant | 2.506 (1.980) | 7.018 (8.300) | 39.725 (58.289) | 7.903\*\*\* (1.712) |
|  |  |  |  |  |
| Observations | 22,817 | 9,330 | 174 | 41,542 |
| R2 | 0.59 | 0.61 | 0.549 | 0.608 |
| Adjusted R2 | 0.589 | 0.61 | 0.516 | 0.608 |
| Residual Std. Error | 37.196 (df = 22800) | 56.858 (df = 9313) | 37.384 (df = 161) | 42.494 (df = 41525) |
| F Statistic | 2,047.365\*\*\* (df = 16; 22800) | 911.144\*\*\* (df = 16; 9313) | 16.343\*\*\* (df = 12; 161) | 4,025.446\*\*\* (df = 16; 41525) |
|  |  |  |  |  |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 |  |  |  |

1. Below are the fixed effects considering all 3 Campaigns.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  | *Dependent variable:* |  |  |
|  |  | weeklyspend |  |  |
|  | -1 | -2 | -3 | -4 |
|  |  |  |  |  |
| recency | -0.107 (0.087) | -2.427\*\*\* (0.790) | -47.288\*\*\* (17.228) | -2.076\*\*\* (0.264) |
| freq | -14.014\*\*\* (0.296) | -16.052\*\*\* (0.845) | -8.263 (5.294) | -11.027\*\*\* (0.284) |
| spendingavg | 0.980\*\*\* (0.008) | 1.002\*\*\* (0.011) | 1.083\*\*\* (0.155) | 0.988\*\*\* (0.005) |
| d | -0.119 (0.466) | -1.386\* (0.822) | 8.725 (5.915) | -0.193 (0.311) |
| t\_loydisc\_cat | 0.310\*\*\* (0.067) | 0.302\* (0.156) | -0.302 (1.517) | 0.363\*\*\* (0.053) |
| t\_display | 0.809\*\*\* (0.074) | 0.561\*\*\* (0.135) | 2.280\*\* (1.032) | 0.583\*\*\* (0.052) |
| t\_favbrndcoup | -0.041 (0.091) | 0.007 (0.180) | -1.275 (1.347) | 0.031 (0.065) |
| t\_mailer | 2.204\*\*\* (0.072) | 2.421\*\*\* (0.141) | -0.945 (0.969) | 1.722\*\*\* (0.053) |
| hhsize | -0.220 (0.654) | 0.645 (3.345) | -16.082 (15.477) | -0.731 (0.554) |
| numkids | 0.092 (0.758) | -0.153 (3.391) |  | -0.320 (0.806) |
| income | 0.00000 (0.00001) | -0.00001 (0.00001) |  | 0.00001 (0.00000) |
| age | 0.068\*\*\* (0.022) | -0.022 (0.068) |  | 0.011 (0.019) |
| married | -0.120 (0.642) | 0.957 (1.977) |  | 0.696 (0.578) |
| Camp\_A | 0.268 (1.594) | 4.174 (3.305) | 10.674 (30.139) | 0.360 (1.150) |
| Camp\_B | -0.644 (1.223) | 2.438 (2.072) | 3.615 (14.481) | -0.769 (0.812) |
| camp\_C | 0.607 (1.716) | 0.350 (2.492) | -24.066 (17.164) | 0.645 (1.064) |
| Constant | 2.506 (1.980) | 7.018 (8.300) | 39.725 (58.289) | 7.903\*\*\* (1.712) |
|  |  |  |  |  |
| Observations | 22,817 | 9,330 | 174 | 41,542 |
| R2 | 0.59 | 0.61 | 0.549 | 0.608 |
| Adjusted R2 | 0.589 | 0.61 | 0.516 | 0.608 |
| Residual Std. Error | 37.196 (df = 22800) | 56.858 (df = 9313) | 37.384 (df = 161) | 42.494 (df = 41525) |
| F Statistic | 2,047.365\*\*\* (df = 16; 22800) | 911.144\*\*\* (df = 16; 9313) | 16.343\*\*\* (df = 12; 161) | 4,025.446\*\*\* (df = 16; 41525) |
|  |  |  |  |  |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 |  |  |  |

**Part 3. Discussion:**

**Looking at Means of R-F-M-D**

**Original Dataset:**

|  |  |  |  |
| --- | --- | --- | --- |
| R | F | M | D |
| 0.607991 | 2.863139 | 60.503 | 0.887282 |

**3 Segmentation clusters**:

|  |  |  |  |
| --- | --- | --- | --- |
|  | C1 | C2 | C3 |
| R | 1.232867 | 0.306305 | 0.325982 |
| F | 2.274766 | 3.161236 | 3.122135 |
| M | 38.36077 | 69.06625 | 71.48919 |
| D | 0.629132 | 0.953417 | 1.031031 |

Difference between clusters and actual dataset.

|  |  |  |  |
| --- | --- | --- | --- |
|  | C1 | C2 | C3 |
| R | 0.624877 | 0.301685 | 0.282008 |
| F | 0.588373 | 0.298097 | 0.258996 |
| M | 22.14223 | 8.56325 | 10.98619 |
| D | 0.25815 | 0.066135 | 0.143749 |

**Insights:**

* Looking at the difference of means and variances of R-F-M-D, from 3 segmentation clusters, Cluster 2 & Cluster 3 closely represent the original cluster.
* Looking at Cluster 1, the customer base from this cluster spends less, have lesser frequency, and have more recency. Hence to increase their visits i.e to reduce their recency they should be exposed to more campaigns.
* Looking at Cluster 2 & 3, they have almost similar characteristics, both have low recency hence have a higher frequency. Both have a spending average higher than the then entire dataset.

**4 Segmentation Clusters**:

**Difference between actual dataset and clusters**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | C1 | C2 | C3 | C4 |
| R | 1.344699 | 0.275027 | 0.057471 | 0.280439 |
| F | 2.148442 | 3.189925 | 3.678161 | 3.178879 |
| M | 39.46532 | 92.56019 | 66.37206 | 64.83361 |
| D | 0.607895 | 1.114968 | 1.091235 | 0.988745 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | C1 | C2 | C3 | C4 |
| R | 0.736709 | 0.332964 | 0.550519 | 0.327551 |
| F | 0.714697 | 0.326786 | 0.815022 | 0.31574 |
| M | 22.14223 | 8.56325 | 10.98619 | 60.503 |
| D | 0.25815 | 0.066135 | 0.143749 | 0.887282 |

**Insights**:

* Looking at the difference of means and variances of R-F-M-D, from 4 segmentation clusters, Cluster 2 & Cluster 4 closely represent the original cluster.
* Cluster 1 have customer with higher recency hence they have lesser visits and lower weekly average spending.
* Cluster 2 has the highest average spending compared to other clusters, they are most exposed to any campaigns, they are more frequent coz there recency is lower compared to cluster 1 and 3.
* Cluster 3 has highest frequency, moderate exposure to campaigns, moderate spending average compared to other models.
* Cluster 4 has the lowest weekly spending average customers, they are very less exposed to campaigns, they have low recency hence have high frequency.

Looking at the analysis of both fixed effects and K-means + fixed effect, as a Data Analyst I would prefer to work with K-means + fixed effect, even though it is a clustering method i.e., an unsupervised method.It would be more appropriate as I will be dealing with high volumes of data, it is a better approach to combine similar kinds of data points in a cluster and analyze each cluster independently, this would minimize the overall error caused by outliers and extreme records. I would use the entire dataset as a starting point and a baseline to compare my analysis and find relevant Independent Variable for a particular Dependent variable.

**Pros and Cons of Sampling, IT Level**

|  |  |
| --- | --- |
| Pros | Cons |
| Comparitively easier to implement than supervised learning | Choosing the ideal number of segments manually |
| Easily scalable to large datasets | Computation need to be dependent on the initial centroid centres |
| Easily adaptable to new addition in the datatset | Accounting outliers would be a major problem |
| Can easily for any type of segments/cluster | Data is of different nature may include null values.  Tackling these issue could be difficult. |
|  | Choosing the ideal iterations is not easy. |

**Pros and Cons of Sampling, Managerial**

|  |  |
| --- | --- |
| Pros | Cons |
| Easy to focus on one particular audience | It is difficult to understand |
| Easy to the most profitable and most non profitable segments, and focus on them accordingly. | Focusing on multiple parts would be more costly |
|  |  |

Conclusion:

After analyzing given data, it is clear that Campaign C was the most successful followed by B and A was not of much advantage in fact it caused a considerable about of loss. Customers with more people living in their household tend to visit more. Store needs to do more improvement on campaign A to gain more profit out of it. Store should be focusing more on old, aged customer like providing the delivery services or pick up services.

Appendix:

I am omitted all the missing/Null values from the dataset.