ML Applications in Marketing: Optimizing Household Expenditure Predictions

Charl van Schoor
Supervisor: Schahin Tofangchi

January 30, 2018

The story

Consumer behavior

- Hedonistic versus utilitarian consumption.
- Can we empirically test for, predict and extract information about, hedonistic consumption behaviour using ML.
- Utilizing predictive and explanatory statistics.

Literature

- Hirschman & Holbrook (1982)
- Shmueli (2010)
- Babin et al. (1994)

The Story

Purpose

Can we predict whether a consumer is utilitarian or hedonistic? In this case, can we predict whether a consumer will purchase a gift or not? We want to see what factors, or topics, contribute the most to consumer behaviour related to giving. Thus, LDA with Logistic regression.

What is the contribution?

- Practically: The method can be employed by retailers to model the distribution of individual consumers over specific topics.
- Theoretically: Serve as an example between the use of both predictive and explanatory modeling.

Data

Data

- The dataset is sourced from the U.S. Department of Labor: Bureau of Labor Statistics (2018) and contains household expenditure data. Put differently, households were asked to keep a diary of frequently purchased items over a period of two weeks.
- The dataset includes all purchases over the period and it also includes descriptions of the products purchased. It also contains other variables indicating the demographics of the household and an indicator for whether they bought a gift.

Data

Data: The Dimensions

- 57195 unique households (NEWID)
- 548 unique products and descriptions (UCC)
- Time dimension removed over the period

NEWID	COST	GIFT	UCC
3281521	1.49	2	120410
3281521	3.28	2	190112
3281521	13.75	2	190212
3281521	6.74498	2	190212

Figure 1: Basic Data

Methodology

Topic model: LDA

- The Topic modeling method used is Latent Dirichlet Allocation.
- The problem: the descriptions of the products don't contain enough information about the product to model different topics.

690117	Portable memory
690118	Digital book readers

LDA

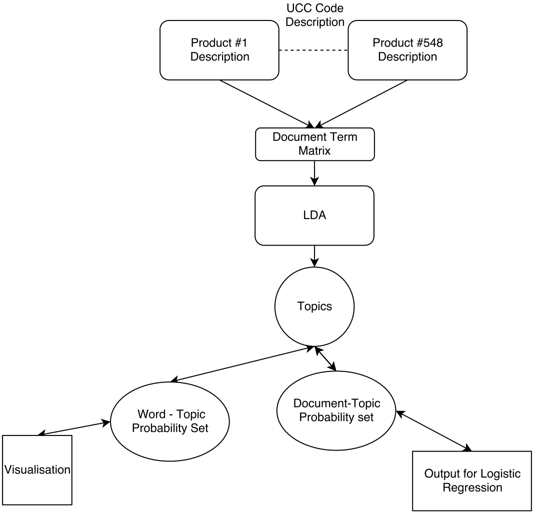
- The solution: Wikipedia

690117	Portable memory	Portable memory A portable storage device (PSD) is a small hard drive designed to hold any kind of digital data.[1] This is slightly different from a portable media player, which stores and plays music and movies. Some are fixed size hard drives of 256GB, 320GB, etc. Newer units are expandable using 2.5" laptop hard drives, allowing for an unlimited storage capacity, which is useful for video and images. When travelling, a portable storage device may be a useful alternative to backing up or purging memory cards if a computer is unavailable for downloading.
690118	Digital book readers	Digital book readers An e-reader, also called an e-book reader or e-book device, is a mobile electronic device that is designed primarily for the purpose of reading digital e-books and periodicals.[1] Any device that can display text on a screen may act as an e-reader, but specialized e-reader devices may optimize portability, readability (especially in sunlight), and battery life for this purpose. Their main advantage over printed books is portability: an e-reader is capable of holding thousands of books while weighing less than one[2] and the convenience provided due to add on features in these devices.

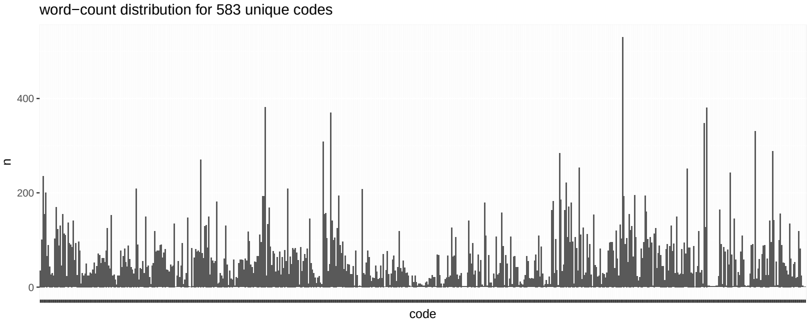
LDA - The Intuition

- The idea is to model each product description into a set of topics. Intuitively, we want each product description to have a value that represents the probability that the given description belongs to a certain topic.
- By doing this we can reduce the dimensionality of the data and use it for further analysis.
- Each product thus has a probability that it belongs to a certain topic. This means that each product is distributed along the topics, but for some more than others.

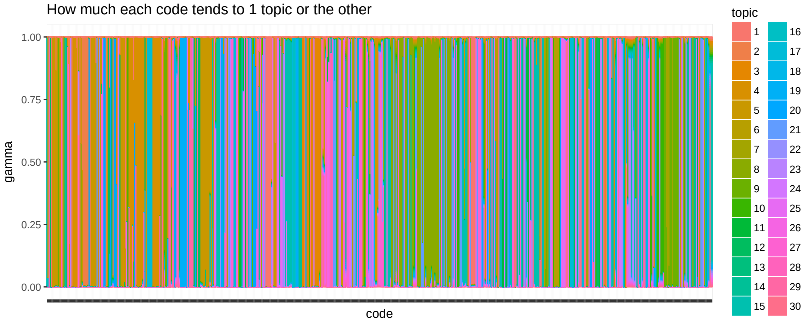
LDA - Layout


UCC Code
Description
Product #1
Description
Product #548
Description
Document Term
Matrix
LDA
Topics
Word - Topic
Probability Set
Document-Topic
Probability set
Visualisation
Output for Logistic
Regression

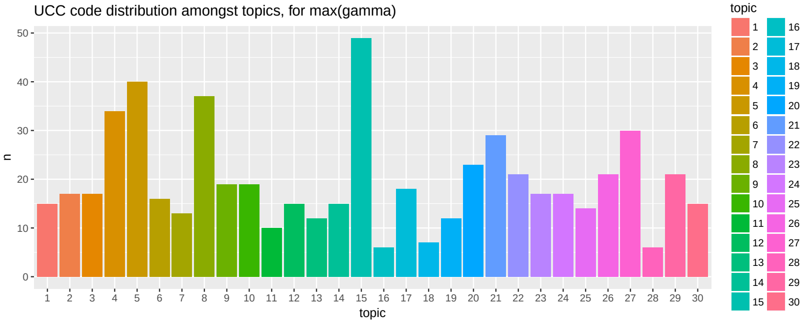
LDA - Some Figures



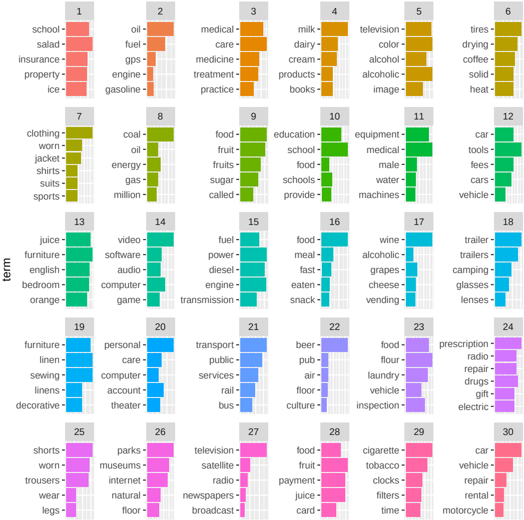
LDA - Some Figures



LDA - Some Figures



LDA - Some Figures



Methodology

Logistic Regressions: Topic Data

- We now have the topics and the probability of each product belonging to a topic. But we want to model the households over the topics (i.e. get a distribution of the topics representing each household)
- I do this by multiplying the probability of each product, given the topic, with the amount of times the household bought each product.
- Thereafter I normalize the topics for each household by taking the unit root of all the products.
- Thus, a distribution for each household over each topic.

Logistic Regressions: Topic Data in Math

$$Topic_i(k) = \frac{\sum_{j=1}^{548} (\gamma_j(k) * n_{ij})}{\sqrt{\sum_{k=1}^{K} \left(\sum_{j=1}^{548} (\gamma_j(k) * n_{ij}) \right)^2}}$$

where i is each individual household, j is each individual product and k is each individual topic. The n represents the amount of times a consumer purchased a particular product.

Logistic Regressions

First Level

$$Gift_i = Topic_i(1) + ... + Topic_i(K)$$

Second Level

$$Gift_i = Topic_i(1) + ... + Topic_i(K) + Inc_i + Educ_i + Sex_i + Age_i + \sum_{s=1}^{55} State_{si}$$

Results

GLM: Topics

Estimate	StdError	zVal	pValue	UCC
-1.98403628	0.06481621	-30.610188	8.957826e-206	topic_30
-0.76305252	0.16647092	-4.583699	4.568226e-06	topic_8
-0.51525338	0.03589368	-14.354989	9.914127e-47	(Intercept)
-0.45379639	0.06205698	-7.312576	2.620686e-13	topic_28
-0.39259318	0.03234708	-12.136896	6.733220e-34	topic_27
-0.21977243	0.04728590	-4.647737	3.355972e-06	topic_13
-0.19290288	0.03618258	-5.331375	9.747210e-08	topic_4
-0.17620187	0.04894867	-3.599728	3.185505e-04	topic_6
-0.17139210	0.04430960	-3.868058	1.097056e-04	topic_2
-0.05901857	0.02987430	-1.975563	4.820425e-02	topic_11

Results

GLM: Topics

Estimate	StdError	zVal	pValue	UCC
3.0079207	0.07407675	40.605462	0.000000e+00	topic_29
1.7140584	0.05708732	30.025205	4.601857e-198	topic_21
1.3344784	0.06599509	20.220874	6.413728e-91	topic_15
1.1377417	0.05567997	20.433591	8.408531e-93	topic_22
0.9859608	0.06560108	15.029644	4.695246e-51	topic_20
0.9142519	0.14288213	6.398644	1.567626e-10	topic_9
0.7352492	0.07498528	9.805246	1.068857e-22	topic_24
0.7122457	0.08460668	8.418315	3.819366e-17	topic_19
0.7112459	0.04807274	14.795203	1.573082e-49	topic_17
0.5213028	0.03717793	14.021837	1.146092e-44	topic_7

Results

GLM: All Predictors

Estimate	StdError	zVal	pValue	UCC
-3.324823e+00	6.224493e-01	-5.3415168	9.217208e-08	(Intercept)
-2.949371e-01	1.804484e-01	-1.6344683	1.021605e-01	topic_6
-1.965635e-01	3.909390e-02	-5.0279834	4.956647e-07	SEXM
-1.918539e-01	1.613237e-01	-1.1892478	2.343422e-01	topic_25
-1.494439e-01	5.486478e-01	-0.2723858	7.853254e-01	topic_8
-1.165106e-01	2.164560e-01	-0.5382648	5.903943e-01	topic_30
6.359636e-07	2.481376e-07	2.5629479	1.037876e-02	FINCBEFX
6.536862e-03	9.300551e-04	7.0284675	2.088142e-12	AGE
3.228125e-02	1.177750e-01	0.2740926	7.840135e-01	topic_12
5.614647e-02	1.582909e-01	0.3547043	7.228111e-01	topic_2

Results

GLM: All Predictors

Estimate	StdError	zVal	pValue	UCC
3.2362392	0.19135170	16.912518	3.638222e-64	topic_22
3.2280386	0.26211160	12.315512	7.473935e-35	topic_29
2.3999099	0.26669634	8.998661	2.284879e-19	topic_24
1.6691431	0.28787153	5.798222	6.702161e-09	topic_19
1.6484076	0.20720863	7.955304	1.786924e-15	topic_21
1.2273696	0.17175022	7.146247	8.918223e-13	topic_17
1.1799147	0.24036415	4.908863	9.160592e-07	topic_15
1.0238406	0.09854285	10.389802	2.759267e-25	topic_1
0.7934677	0.09851200	8.054528	7.978596e-16	topic_26
0.7742876	0.59699200	1.296982	1.946376e-01	EDUCA8

Results

GLM: Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1246  848
         1  383  748

               Accuracy : 0.6183
                 95% CI : (0.6013, 0.6351)
    No Information Rate : 0.5051
    P-Value [Acc > NIR] : < 2.2e-16
```

In other words..

- Using topic modeling seems to be an interesting approach to reduce the dimensionality of the data. However, the results are subject to the amount of topics specified.

In other words..

- Using topic modeling seems to be an interesting approach to reduce the dimensionality of the data. However, the results are subject to the amount of topics specified.
- Moreover, the predictive capability of the GLM does not seem to fit the data. Another method should be used for prediction.

In other words..

- Using topic modeling seems to be an interesting approach to reduce the dimensionality of the data. However, the results are subject to the amount of topics specified.
- Moreover, the predictive capability of the GLM does not seem to fit the data. Another method should be used for prediction.
- GLM does offer some explanatory power. The top hedonic predictors contained descriptions like: Tires, wear, coal, oil, vehicle etc.. Another top predictor is a male dominated household.

In other words..

- Using topic modeling seems to be an interesting approach to reduce the dimensionality of the data. However, the results are subject to the amount of topics specified.
- Moreover, the predictive capability of the GLM does not seem to fit the data. Another method should be used for prediction.
- GLM does offer some explanatory power. The top hedonic predictors contained descriptions like: Tires, wear, coal, oil, vehicle etc.. Another top predictor is a male dominated household.
- The top utilitarian predictors contained descriptions like: Pub, culture, clocks, decorative, public, transport etc. Another top predictor is a high level of education within the household.

Still to come

- NN for prediction
- Different dependent variables
- More or less topics

Discussion

Bibliography

Babin, B. J., Darden, W. R., & Griffin, M. (1994). Work and/or fun: Measuring hedonic and utilitarian shopping value. *Journal of Consumer Research*, *20*(4), 644-656. Retrieved from +http://dx.doi.org/10.1086/209376

Hirschman, E. C., & Holbrook, M. B. (1982). Hedonic consumption: emerging concepts, methods and propositions. *The Journal of Marketing*, 92–101.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, 289–310.

U.S. Department of Labor: Bureau of Labor Statistics. (2018). *Consumer expenditure surveys : Diary interview survey.* Retrieved from https://www.bls.gov/cex/pumd_data.htm