ML Applications in Marketing: Optimizing Household Expenditure Predictions

Charl van Schoor*
Supervisor: Schahin Tofangchi

February 15, 2018

jajajaja

*E-mail: charlvanschoor1@gmail.com. The submission of this paper is in fulfillment with the requirements for the subject Crucial Topics in Information Management: Machine Learning Applications in the Context of Digital Transformation. The source files for this paper can be downloaded from the following GitHub repository:

Contents

1 Introduction 4

2 Theoretical Background 5

3 Data Structure 8
- 3.1 Product Definitions . . . 8
- 3.2 Final Dataset . . . 9

4 Methodology 10
- 4.1 LDA Model . . . 10
- 4.2 Logistic regression . . . 12
- 4.3 Neural Network . . . 12

5 Results 13
- 5.1 LDA . . . 13
- 5.2 Logistic Regression . . . 16
- 5.3 Neural Network . . . 17

6 Conclusion 21

References 21

7 Appendix 23
- 7.1 Tables . . . 23
- 7.2 Figures . . . 24

List of Figures

1 LDA Input-Output Layout . . . 11
2 Topics to Final Dataset Layout . . . 12
3 Product-Topic Distribution . . . 14
4 Product-Topic Mixture . . . 14
5 Top 6 Words Over 30 Topics . . . 15
6 Top 20 Predictors for 100 Topics . . . 17
7 Confusion Matrix: Top 20 Predictors for 100 Topics . . . 18
8 Auxiliary Statistics: Top 20 Predictors for 100 Topics . . . 18
9 Test Accuracy . . . 19
10 Test Loss . . . 19
11 Validation Accuracy . . . 20
12 Validation Loss . . . 20
13 Top 10 Words Over 7 Topics . . . 24
14 Word-Count Per Product . . . 25
15 CPP Matrix Visualization . . . 26
16 Distribution of Household Age for Gift and Non-Gift . . . 27
17 Distribution of Household Sex for Gift and Non-Gift . . . 27
18 Distribution of Household Education for Gift and Non-Gift . . . 28
19 Distribution of Household Income Class for Gift and Non-Gift . . . 29
20 Distribution of State for Gift and Non-Gift . . . 30

List of Tables

1 Variable Descriptions . . . 23

1 Introduction

Measuring hedonic and utilitarian consumption patterns has proven to be a difficult undertaking in marketing literature. Researchers argue over the scale of how to define utilitarian and hedonic consumption Batra & Ahtola (1991). However, the advent of machine learning and big data have brought about a new dimension of viewing marketing questions, such as the measurement and prediction of hedonic versus utilitarian consumer behaviour. Therefore, the aim of this paper is to contribute to the literature that utilizes machine learning methods to measure different types of consumer behaviour.

Marketing literature usually makes use of different sets of products and their associated attributes to define whether a product is of hedonic or utilitarian nature. Researchers then use these products and consumers' purchases of these products to assign a consumer as being either hedonic or utilitarian (Spangenberg et al., 1997). However, this method poses a few problems. Firstly, it is labour intensive, meaning that a researcher has to manually define each product. Secondly, although the assignment of a product is based on a set of literature, it is still subjective to the particular researcher. Thirdly, the definitions of the products are inconsistent when applied to different sets of products; a set of products can contain both hedoinc and utilitarian products. Therefore, generalizing a scale of hedonic and utilitarian products based on a certain set of products appears to be inefficient.

The purpose of this paper is thus to assist in this field of marketing literature by utilizing machine learning algorithms to automate the assignment of products to product categories, specifically hedonic and utilitarian categories. Moreover, the objective of the paper is to predict whether a consumer is either of hedonic or utilitarian nature. This is however a difficult challenge due to the nature of the definition process. How does one know whether a product is either hedonic or utilitarian? How can one assign a measurement to a product category for each consumer to get a distribution of a consumers preference of hedonic and utilitarian products? This is the innovation, or practical contribution, of this paper. It utilizes a machine learning algorithm, specifically a topic modeling method, called Latent Dirichlet Allocation (LDA from hereon out) to assign products into different topics, based on the description of the particular products.

The result of using this method is that each product has a probability of it belonging to a certain topic. In other words, the model gives a distribution of each product over the set of specified topics. However, this only answers one of the purposes of this paper. The second purpose is to model a consumer over these topics. To do this, one can multiply the amount of times a consumer purchases a product with the probability of the product being in each topic. This results in a unique distribution for each individual consumer over each topic, which in turn can be used to predict whether they are of hedonic or utilitarian nature.

Another feature of this paper is the combination of descriptive, explanatory and predictive statistics. The use of these different statistical methods is based on the intuition of Shmueli (2010) which argues that these methods can be used for the purposes of theory building. However, the author suggests that each method must be used in their respective manner. It is thus

also the aim of this paper to act as a theoretical case study whereby all of these methods are used to explain a particular topic. This is the theoretical contribution of this paper; to describe consumer behavior by reducing the dimensions of product descriptions, by explaining which topics contribute the most to hedonic and utilitarian behaviour and to predict whether a consumer is of hedonic or utilitarian nature. This paper thus serves as a case study to distinguish between the use of statistics to explain and predict consumer behaviour.

To accomplish the above mentioned secondary task, this paper makes use of both a diverse set of statistical methods. Firstly, the descriptive part of the paper can be considered as the dimensionality reduction section, or the LDA. Using LDA on product descriptions reduces the amount of information about products, and thus better describes the dataset. Secondly, this paper makes use of logistic regression functions to model the topics over a binary dependent variable. The result is a set of topics describing the binary outcome, which proxies for hedonic and utilitarian behaviour. The model also includes control variables measuring the demographics of consumers. Finally, the paper contains prediction results from a neural network and a random forest. The intuition is that there exists a non linear relationship between products purchased and hedonic behaviour.

The ultimate goal of this paper is to serve as an example of how modern statistics can be used in combination with consumer data to model consumer behaviour. It aims to contribute to the practical implementation of machine learning methods in the field of marketing, and to serve as a case study for the theoretical argument of which type of statistic is useful for analyzing consumer behaviour. The remainder of this paper is structured as follows: The following section gives a brief overview of literature related to this paper, followed by the data section which describes the data used in this analysis. Thereafter section 4 defines the methods used in the analysis of this paper. Section 5 presents the results and section 6 concludes.

2 Theoretical Background

Defining utilitarian and hedonic consumption is historically been done by considering certain products as hedonic products (Crowley et al., 1992). This approach, however, is speculated on as the intrinsic value of products are subject to change. This fluctuation is a result of market driving forces. Spangenberg et al. (1997) argues that market analysis of hedonic versus utilitarian consumption scales are subject to the products being observed. These scales of consumer behaviour are based on the specific product categories analyzed, and are problematic when applied to different baskets of goods, or product categories. Although these scales are useful for case studies, they lack general applicability to products, services and other non-shopping activities (Spangenberg et al., 1997). A different approach to this problem would be to model consumer purchases over a set of topics which is based on product descriptions, rather than predefined scales. Thus, having a distribution of consumer purchases might be a useful scale to measure hedonic and utilitarian consumption patterns.

It is however important to fundamentally understand hedonic and utilitarian consumption. Hirschman & Holbrook (1982) theoretically defines hedonic consumption in the following way: "Hedonic consumption designates those facets of consumer behavior that relate the multi-sensory, fantasy and emotive aspects of one's experience with products". In other words, hedonic consumption can be seen as consumption based on sensory satisfaction, or excessive pleasure, which based on previous experiences. Moreover, Okada (2005) argues that hedonic consumption is related to a sensation of guilt. The author argues that hedonic behavior is associated with impulse purchases as those purchases are related to satisfying pleasurable sensations. In other words, hedonic consumption can be viewed as consumption based on pleasure. The notion is that consumers can purchase pleasure, and having experienced that pleasure one wants more of it.

Measuring utilitarian consumption is considered to be the stark contrast of hedonic consumption. Batra & Ahtola (1991) argues that utilitarian consumption is associated with the "expectation of consequences". The authors argue that utilitarian behaviour is associated with the outcome of purchasing products. With this, consumers are more diligent in their purchasing behaviour and consider the impact of their purchases (Batra & Ahtola, 1991). In other words, utilitarian consumption is associated with a rational thought process, where purchasing is associated with functionality of the products rather than the sensational experience of the products. For this reason, utilitarian consumption can also, in part, be considered as altruistic consumption. The reasoning for this is that consumers consider the impact of their purchases rather than just considering the pleasure derived from those purchases.

The theoretical definitions of hedonic and utilitarian consumers are thus a sound definitions. However, empirically predicting whether consumers are hedonic or ultriustic poses a different challenge. As mentioned above, using product categories is subject to scrutiny when other products are considered, and thus empirical papers are case sensitive. The advent of big data might however bring about a new frontier when it comes to understanding the consumption patterns of individual consumers (Stroie) (2014). Having a customer specific distributions over the products purchased might yield fruitful measurements, in the form of topics, for hedonic and utilitarian consumption. For this reason, machine learning methods can be applied to model these distributions. This paper thus aims at creating a distribution of consumers by using text mining algorithm called Latent Dirichlet Allocation.

Most literature pertaining to marketing and LDA focus mainly on text rich sources such as social media and product reviews for the purposes of topic modeling. Ma et al. (2013) uses LDA and synonym lexicon methods to extract product features from online costumer reviews of certain products. Melville et al. (2009) discuss techniques related to clustering social media discussions on products, and highlight the role of LDA in discovering topics of products based on blogers discussing products. Moreover, Jacobs et al. (2016) compares LDA and mixtures of Dirichlet-Multinomials (MDM) on online consumer purchasing data to predict which product a consumer would buy next. The authors note that LDA is more scalable relative to MDM, and would be useful on broader datasets containing purchasing information.

The use of LDA in modeling consumer behaviour is nothing new. However, it's usage on purchase datasets is scarce. This is due to the nature of product descriptions. Christidis et al. (2010) explore this topic by using LDA to model consumer preferences as well as to effectively recommend products to consumers. In particular, the authors attempt to identify latent baskets and consumers from purchase data. They find intuitive baskets, modeled by their LDA structure, for different sets of products. However, a point to note is the lack of document information which is needed to distinguish topics from each other, and the authors suggest using social media and product hierarchies to further distinguish their topics (Christidis et al., 2010). This highlights the problem of using LDA on consumer purchases: Products have short descriptions. This leads to weak allocation of products to topics, as the LDA cannot distinguish between products. For this reason, the dataset used in this paper contains product descriptions augmented with Wikipedia definitions and descriptions of these products[1].

As mentioned in the introduction, this paper has a twofold empirical purpose. The first is the constructing of consumer purchase distributions, and the second is to use these distributions to predict whether a consumer is of hedonic or utilitarian nature. Moreover, the second purpose has two components: Explaining and predicting hedonic behaviour by using the newly defined topics and other consumer demographic variables. The paper thus includes a logistic regression model for explanatory purposes, and a neural network and random forest for prediction. This means that a suitable predicted variable is necessary. The solution is partly based on marketing and economic literature, as well as intuition. However, this proved to be challenging given the scope of the dataset used in this analysis. Therefore, the predicted variable used for the models is whether a consumer purchased a gift or not; where purchasing a gift can be considered as a proxy for utilitarian behaviour and not purchasing a gift is considered as hedonic behaviour. The variable is subject to scrutiny, but it fits the scope of this analysis and is therefore used as the predicted variable.

The intuition behind using gifts purchased as the predicted variable is quite simple. Sherry (1983) states that gift giving is vehicle of social obligation, and is considered as charitable behaviour. The author explains that gift-giving behaviour is associated with a social contract, that of trust. The argument is that consumer give gifts to build social contracts, and that for that reason purchasing a particular gift is mostly premeditated. It is for this reason that the variable is a suitable proxy for utilitarian behaviour, as it not based on random excessive purchases (or hedonic consumption). Moreover, Andreoni & Miller (2002) explains that economists view gift giving as rational, altruistic, behaviour. This view is not based on opinion, but on theoretical economic principles such as game theory, whereby giving a gift is associated with trust building in sequential games. It is a form of utility maximization, outside of short term gain (Andreoni & Miller, 2002). This is contrary to hedonic behaviour, and thus establishes a bases for the predicted variable used in this analysis.

To summarize, this paper aims to predict whether a consumer is likely to purchase a gift by using

[1]See the data section of this paper for a description of the augmentation process

the consumer's purchase history modeled over a set of topics and other demographic variables as predictor variables. The practical implication of the paper is a that of a recommendation to retailers and governments. It would be useful for producers, retailers and governments to know the distribution of their costumers' and constituents consumption behaviour. This method can be used to model customer preferences and decision behaviour. Costumers have different preferences and organization can utilize the tools presented in this paper to better understand their consumer base, which can be used to either implement social policies or individual marketing strategies. The recommendation is this: Datasets containing purchasing behaviour, such as the one used in this paper, should contain multidimensional product description. Possible ideas for variables defining the product are: A category variable for the product, a variable for compliment and substitute products, ingredients, production methods used describing the general recourses used in the industry producing the product. Such features describing the products would give a new dimension for researcher to use text mining algorithms to model consumer distributions among topics. Although ambitious, this recommendation is that organizations merely updated the text describing products, which will afford them a new avenue to model consumer behaviour.

3 Data Structure

The following section describes the dataset used in this paper and is comprised of two components: The first subsection explains the dataset containing the product definitions which is augmented with wikipedia definitions and descriptions of the products. The second subsection describes the final dataset used for the explanatory and predictive components of this paper.

However, it is important to first cover the source and dimensions of the full dataset used for the analysis. The sample dataset contains consumer expenditure survey data sourced for the US Department of Labor Statistics, and is known as the Consumer Expenditure Survey data (CES from hereon out). Effectively, the dataset is a cross sectional dataset consisting of a sample of US households' purchase expenditures[2]. Moreover, it contains demographic variables describing the households and a categorical variable indicating whether a consumer purchased a given product as a gift for a person outside of the household. The dataset used includes survey entries from 2013 up till 2016, in total yielding 57195 unique households and 548 unique products.

3.1 Product Definitions

Each product within the dataset comes with a description of the product. For instance, product 20110 is described as White bread. However, these descriptions are not suitable for a text mining technique such as LDA. The reason is that the descriptions are not suitable documents for the LDA, meaning that it cannot reasonably allocate a probability of each document belonging to topic. This is possibly the reason why text mining techniques are not used on purchase datasets, as the descriptions do not contain enough information to allocate documents to topics. Initial

[2]Each household in the sample was asked to keep a diary describing their product purchases for a duration of two weeks.

analyses suggested that the allocation of products to topics are random[3]. In other words, the products are randomly mixed among the topics and the result is that one cannot clearly see a topic representing a specific set of goods.

Therefore, either another dataset containing text of the products should be used, or the current dataset should be augmented with a dataset that contains more information on the products. The method of this paper is that of the latter; augmenting the dataset is the only way to go as there is not another dataset containing more information on the products used in this paper. The question thus becomes: Where can one obtain text information on all of these products? The solution was to manually crawl the website Wikipedia and augment each product description with a Wikipedia description of the product. This is done by searching the main words within the description on Wikipeida, and using the most relevant paragraphs as augmenters. In the above mentioned white bread example, one can search Wikipedia for the term "White Bread" and obtain the following description: "White bread typically refers to breads made from wheat flour from which the bran and the germ layers have been removed (and set aside) from the whole wheatberry as part of the flour grinding or milling process, producing a light-colored flour." (contributors, 2018). The result of augment each product with such a description, and even longer descriptions, is larger and more text rich documents to model into topics. Thus, each product has it's original description and the augmented descriptions (descriptions from hereon out) as a single description variable.

3.2 Final Dataset

The final dataset contains a identification variable for each individual household, the predicted gift variable, variables constructed from the CES data representing the demographics of each household and the individual distribution of each household over the topics[4]. Firstly, each household is, for the purposes of this paper, considered as a consumer. Each consumer has a categorical variable describing whether a product is purchased as a gift. However, the purpose is to predict whether a household would buy a gift or not. Therefore, each consumer is assigned a binary value of 1 for whether the consumer purchased any product as a gift, and 0 if the consumer did not purchase any gifts. This results in the predicted variable GIFT.

Moreover, the demographic variables used as control variables are as follows: AGE describes the mean age of the household and SEX describes whether the majority of a household was male, female or balanced. EDUCA represents the highest level of education attained by any member of the household. It is coded from 1 to 9 and acts as a proxy for the level of education within the household[5]. STATE is a categorical variable that reports the US state in which the household is surveyed, and can, to some extent, be viewed as a proxy for culture. INCLASS is a

[3]Figure 1 in the appendix represents the allocation of the top words (given by their respective beta values) in each product document over 7 topics. The graph shows that the word allocation is random, which makes it difficult to infer whether a certain topic is capturing a certain grouping of products.

[4]Table 1 in the appendix summarizes the dataset described in this subsection.

[5]A value of 1 represents no schooling and 9 represents a masters degree or higher.

categorical variable representing the income class of the household[6]. The final dataset includes all of the above mentioned variables as well as the topics modeled by the LDA[7]. The following section describes how the topic variables are constructed.

4 Methodology

4.1 LDA Model

Constructing the topic variables requires two components: First is a probability that each product belongs to each specific topic. Second is a measurement for each consumer that links a consumer's personal purchase profile with each topic. Constructing the first proves to be quite simple; the topic modeling library Grün & Hornik (2011) from CRAN's library of packages is used for the purposes of this paper[8]. The model requires a document-term matrix to return the beta and gamma coefficients. The topic modeling library provides useful functions to tokenize and remove any stop-words from the documents. by casting the documents, one obtains the document terms matrices required for the model[9]. The beta coefficients, or the word-topic probabilities, represents the probabilities for each word belonging to each topic. The gamma coefficients, or the document-topic probabilities, represent the probabilities of a document belonging to each of the topics. The gamma matrix is used for the purpose of this paper. Each product's description thus has a probability of belonging to a topic. However, this does not yield a particular distribution for each consumer. To get a distribution for each consumer over the topics, one needs to consider the link between the gamma matrix and a measurement for household "interest" in a product.

A simple way to to connect the gamma matrix with each individual is to have a normalized unit root value per topic per consumer. To do this, I counted the number of times a consumer bought a particular product. However, the surveys were conducted randomly and thus some households repeated in the sample. To account for this, I averaged the amount of times of those households purchased products over the amount of times they repeated in the sample. This resulted in the following consumer product propensity matrix:

$$CPP = \begin{bmatrix} NEWID & product & n \\ 1292531 & 10120 & 1.0 \\ 1292531 & 200110 & 1.5 \\ \vdots & \vdots & \vdots \end{bmatrix} \tag{1}$$

Where *NEWID* is a vector of the consumers, *product* is the UCC code for each product and n the average amount of times a consumer purchased a product[10]. Thereafter, I then multiplied the

[6] As before, the higher the category, the higher the level of income.

[7] Figures 3 to 5 show the distributions of the demographic variables stacked by whether the households purchased gifts or not. The graphs show that gift purchasing behaviour is relatively uniform over the demogrhapics of hte households.

[8] It's similar to the topic modeling packages like Scikit Learn for Python.

[9] Figure 14 in the Appendix shows the word-count per product description, after all the documents are tokenized and the stop words removed

[10] Figure 15 in the Appendix shows a visual representation of the CPP matrix, where households are on the y

amount of times, n, each product was purchased by each consumer by the probability (gamma value for that product) that the product belonged to a certain topic. Thereafter I took the unit root of each topic with respect to each topic and particular household to normalize the value:

$$Topic_i(k) = \frac{\sum_{j=1}^{509} (\gamma_j(k) * n_{ij})}{\sqrt{\sum_{k=1}^{K} \left(\sum_{j=1}^{548} (\gamma_j(k) * n_{ij}) \right)^2}} \tag{2}$$

where i is each individual household, j is each individual product and k is each individual topic. The n represents the average amount of times a consumer purchased a particular product. Figures 1 and 2 visually represent this process.

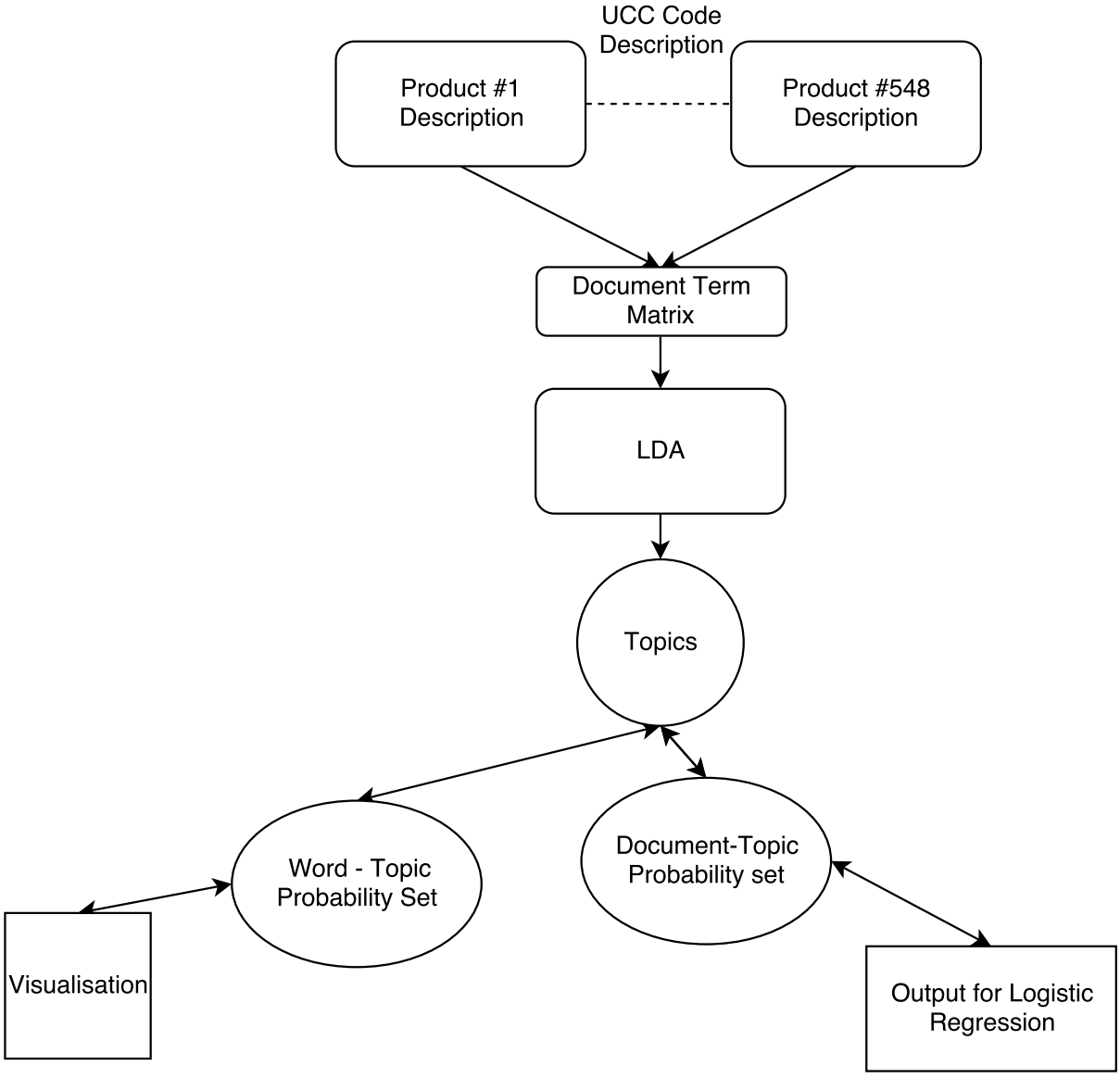


Figure 1: LDA Input-Output Layout

axis and products on the x axis; it simply shows that some products are frequently purchased.

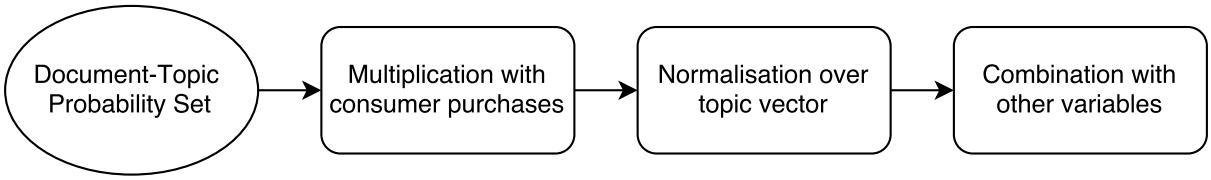


Figure 2: Topics to Final Dataset Layout

4.2 Logistic regression

The second goal of the paper is to identify which factors explain the most variation in gift purchasing behaviour. To do this I make use of a binomial logistic regression function which uses a logistic distribution log-likelihood function[11]. The purpose of the logistic regression is to act as a baseline for predicting gift purchasing behaviour. The regression structure is as follows:

$$GIFT_i = \sum_{k=1}^{K} Topic_i(k) + AGE_i + INCLASS_i + EDUCA_i + STATE_i + SEX_i \qquad (3)$$

where $i = 1....N$ and $k = 1....L + K$

Each variable follows intuitively on the variables specified in the data section of this paper. Notice that the variable "Topic" is a summation over the topic vectors. The reason for this is that various amounts of topics are modeled for the purpose of this paper; the aim of which is to determine what set of topics is optimal for the highest prediction accuracy. The regressions thus models the log-likelihood, or intuitively the probability, that a consumer would purchase a gift or not. The decision boundary is set a 0.5, meaning that a probability below 0.5 is categorized as a 0, whereas a value equal to or above 0.5 is categorized as a 1.

4.3 Neural Network

The purpose of the neural network is to improve the prediction accuracy of the model, as this method is known to have a higher predictive capability than linear regression models. It has the ability to model non-linear relationships in a black box manner. This is however also the disadvantage of using this technique, as it does not allow for possible causal inference of the variables in the model.

The neural network is set op to train and test the same model as that of the logistic regression model. The total dataset is constructed such that there are the same amount of households that have a value of 1 and 0 for GIFT. In other words, the training and testing datasets have reduced dimensions to account for possible pattern forming. If this is not done, the neural network learns to predict whether a household will not purchase a gift very well, but it falls short of predicting whether a household will purchase a gift.

[11]The programming software R comes standard with Generalized Linear Models, of which the binomial GLM is used for this paper.

Moreover, the neural network is constructed with the Keras library that utilizes a TensorFlow backend. The model is set to have one hidden layer with 128 neurons as the nature of the dataset does not require the complexity of more hidden layers. The model includes a 30 percent dropout during training, as this helps the model from overfitting(Srivastava et al., 2014). The model also includes cross-validation during the training set to help detect over-fitting and to assist in hyper-parameter optimization. Furthermore, the hidden layer is fitted with a standard sigmoid activation function and the output layer is fitted with a softmax activation function[12]. The model is fitted with adam as an optimization function, and a binary cross entropy loss function. Th model was set to run a smaller set of epochs with smaller bach sizes in order for it to train at a slower rate. Finally, the metrics used to validate the model is the accuracy of prediction in the test dataset and the optimization of the loss during training.

5 Results

5.1 LDA

The results from the LDA depends on the amount of topics specified. This requires consideration on two parts: First is whether the amount of topics specified makes intuitive sense. The second is what amount of topics produces the highest gift prediction accuracy. Therefore, I created an infrastructure to model a wide set of topics ranging from 2 to 10, and thereafter different intervals ranging from 10 up till a 100 topics[13]. These are useful for computing different regression and neural network predictions. However, I stick with 30 topics for the purpose of this section to show how the model performed[14].

Figure 3 below represents the product distributions over each of the 30 topics. In other words, the figure shows the total documents likely to be within each topic; each product's, or document's, maximum gamma probability value is taken, thereby allocating it to only a specific topic, thereafter each topic is counted for the amount of products within each topic. The figure shows a good distribution of products over the topics, meaning that the topics do capture a variety of products. This is the aim of using the Wikipedia descriptions of the products, to allocate them into distinct topics without having a single topic, or a set of topics, dominate the entire product space.

Moreover, Figure 4 shows a similar story as that of the one above; it shows the gamma probability distributions for each of the products, or documents. In other words, the figure shows the probabilities for each document belonging to each of the 30 topic[15]. As visible from the figure, most products are allocated to specific topics; this indicates that most products fall within a certain topic, which is a good result given the origin of the text data. However, some products

[12]The decision for which activation function is made on intuition, as both activation functions yields better results.

[13]the source files provided for this paper shows all of the topics modeled as well as figures representing all of the topics. Each of the figures presented in this paper has counterparts for all of the topics in the sources files.

[14]Although there exist test to check the optimal amount of topics within a dataset (such as those developed by Arun et al. (2010), Griffiths & Steyvers (2004) and more) the amount of topics used in this section is based on intuition and is meant to represent the results from the modeling.

[15]The sources files provided for this paper contains figures representing this for all of the topic sets.

Figure 3: Product-Topic Distribution

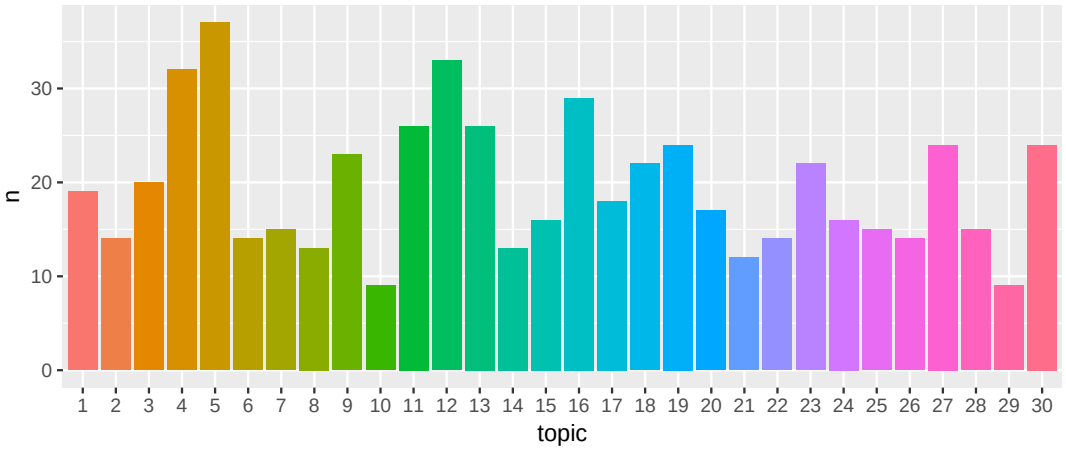
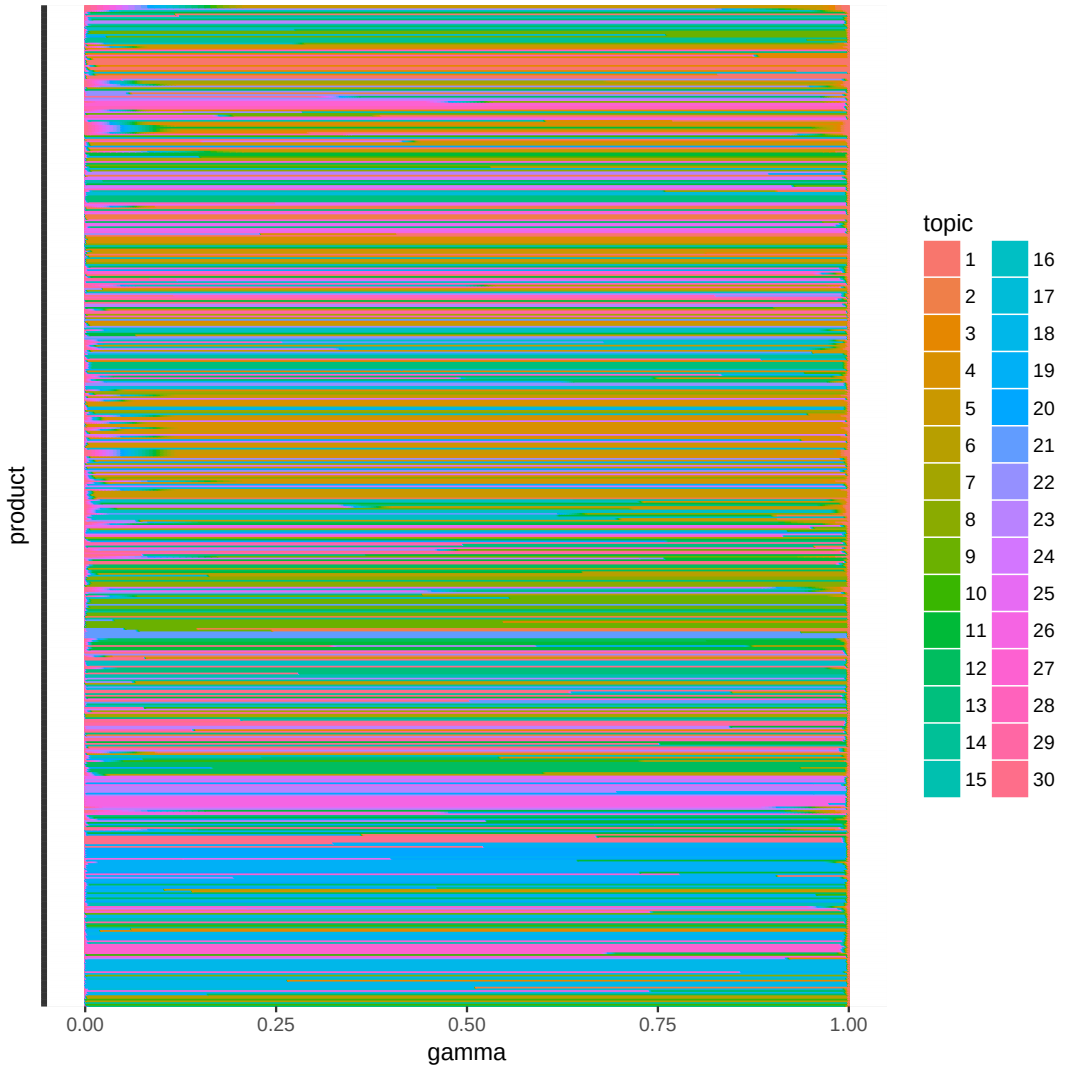


Figure 4: Product-Topic Mixture



belong to various topics. Once again, this is not a bad result as it gives some variance to the model. This means that the LDA possibly allocated a probability for complimentary products into the same topics, and probabilities for substitute products into other topics. In other

words, some products belong to each other, like vehicles and fuel, and the LDA allocated higher probabilities to those products in the same topic. Some products are substitutes products, like sugar and artificial sweeteners for example, and the LDA possibbly allocated probabilities for those products in the same topic. This means that the LDA might be picking up relationships between the products that are not available when considering the products on their own; it would be tedious for a person to allocate more than 500 products into topics.

Figure 5: Top 6 Words Over 30 Topics

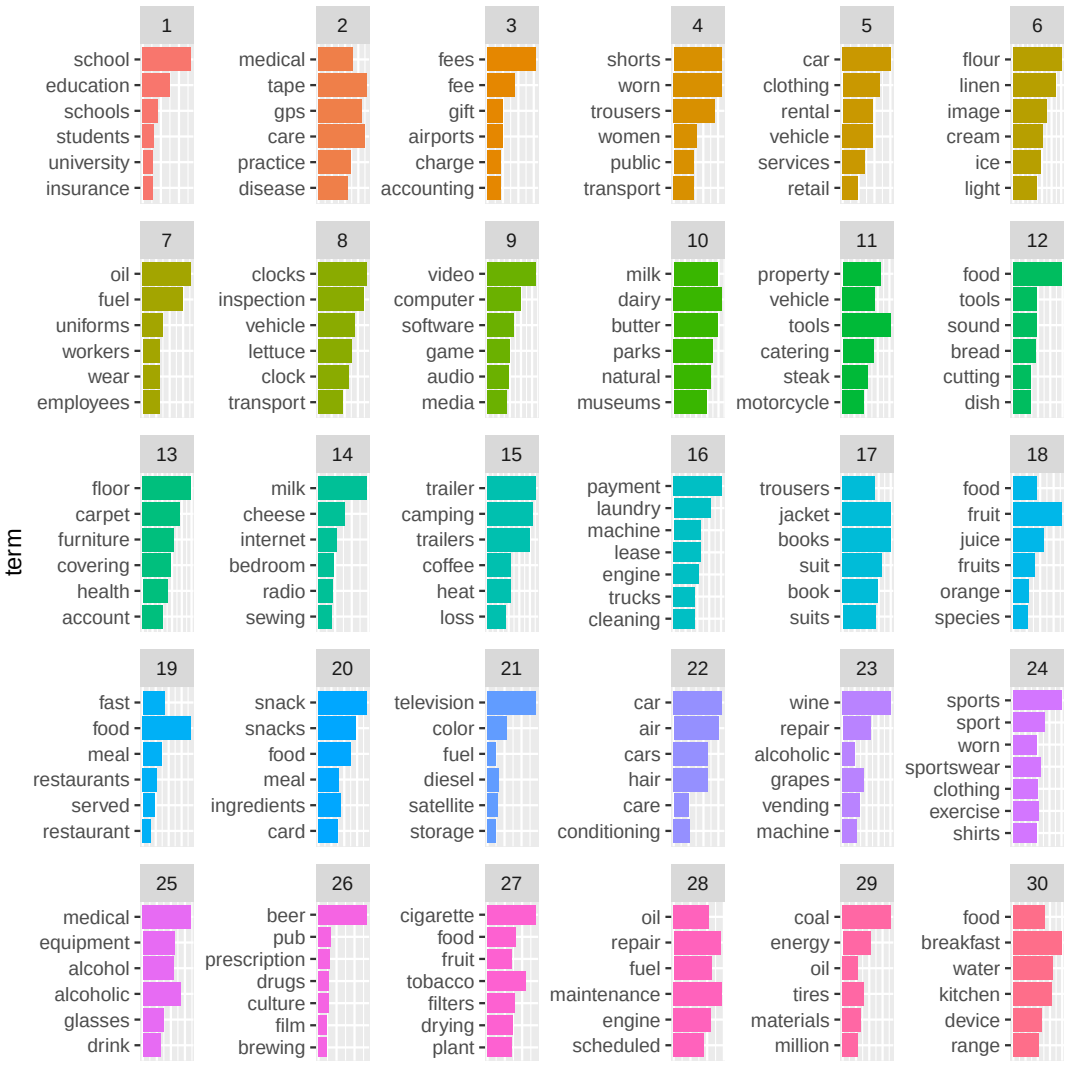


Figure 5 emphasizes the point made above. The figure shows top 6 words (the maximum beta probabilities for each word) in the 30 topics. There are a few notable topics: Topic 1 represents electronic devices, topic 2 consists of food products, topic 7 represents camping related products and services, topic 12 consists of fast food and restaurant products, topic 23 represents clothing items, topic 29 connects cultural activities etc. The figure shows that the model was able to allocate products that are related to each other. This emphasizes the point that a LDA model requires more information in the documents to allocate products into topics. The result is a mixture of product in each topic that are related in some way, by either cultural or economic

principles[16].

5.2 Logistic Regression

The results from the logistic regression show a compelling story for the inclusion of topics into the model. However, the models seems to show endogeneity, as the significance of the topics far outstretch the significance of the control variables. This is due to the prediction variable used for the model; it is intuitive that the relationship between product and gift purchases is strong, as the one is a direct result of the other. I controlled for this, to some extent, by allocating a binary value to whether a consumer purchased a gift or not, as this eliminates some of the direct relationship between the values. The results do however tell a story, thus we proceed.

Figure 6 shows the top 20 coefficients in terms of significance (sorted by p-value) for the logistic regression that included a 100 topics. The reasoning behind using 100 topics as a representation of the model is due to a few statistics indicating the accuracy and the explanatory power of the model[17]. The top 20 predictors are sorted from lowest to the highest p-value. It is clear from the figure that the most significant predictors are the topics[18]. Topic 73 is the most significant in explaining and predicting gift purchasing behaviour, where topic 99 is the 20th most significant predictor. It is interesting, and intuitive, to note that AGE is the 7th most significant predictor, indicating that generally older households tend to purchase more gifts. The top predictor topics, after inspecting the dataset, contained words such as vehicle, property, truck, agreement, school, bus, students, united, shorts, trousers, sports, electronic etc. Interpreting these results is however difficult, as one does not know exactly what the relationships the topics capture.

It is clear thought that purchases of these types of products are strongly related to gift purchasing behaviour. Note that all of the top predictors are positive, meaning that all increase the likelihood of gift purchasing behaviour. One can speculate that households with property, vehicles, social agreements, children as students etc. have a higher probability of buying gifts. This follows on the intuition of the theoretical background of this paper: Individuals that consider the outcome of their purchases are more probable to exhibit gift giving behaviour.

Figures 7 and 8 show the predictive capability of the logistic regression model. Figure 7 is the confusion matrix from using the 100 topics. As visible from the graph, the model is able to predict non-gift purchasing behaviour well, 86 percent accuracy, but not well with predicting gift purchasing behaviour, 55 percent accuracy.

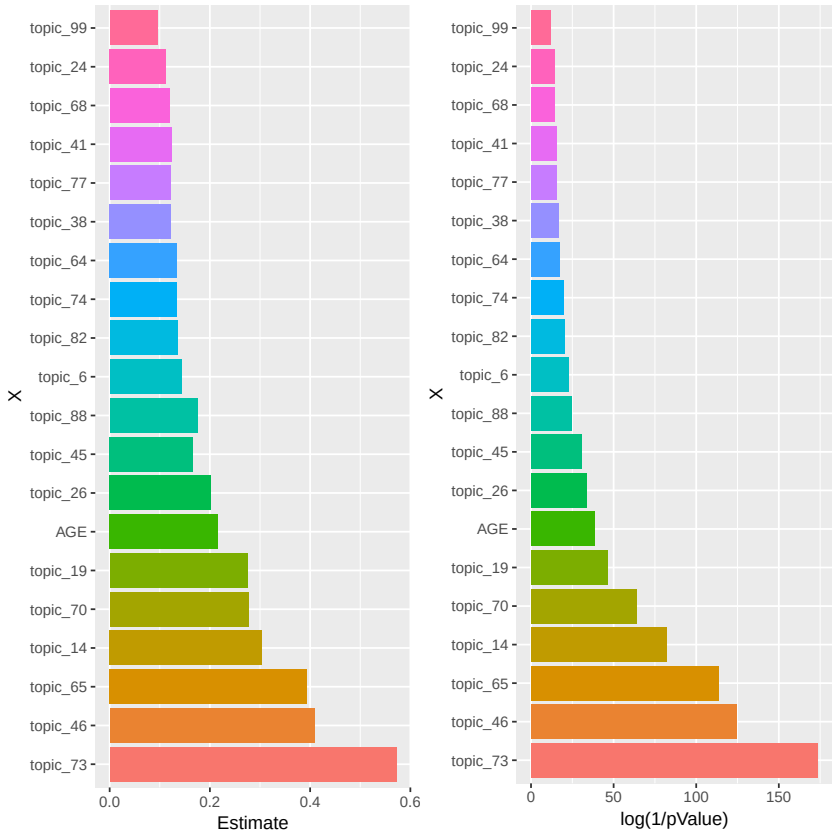
Moreover, Figure 8 shows various statistics for the model. Some important elements are the Mcfadden R-square value and the balanced accuracy. The Mcfadden R-square value is only

[16] As an economic example, topic 21 relates education with public transport, which makes sense as students use public transport.

[17] The source files provided for this paper contains the results for all of the models with a different set of topics.

[18] The source files show that with less topics, around 40, more of the demographic variables are of importance; an intuitive result.

Figure 6: Top 20 Predictors for 100 Topics



25 percent, meaning that the model as a whole has low explanatory power[19]. Moreover, the balanced accuracy, which is a measurement for the predictive power, is 70 percent. In other words, the model only predicts false positives and false negatives 30 percent of the time. Thus, the model does have some predictive capability, but lacks in the prediction of gift purchasing behaviour.

5.3 Neural Network

Figures 9 to 12 show the results for the neural network. Each figure reports the results for the models with 10, 50 and a 100 topics. Figure 9 and 10 show the accuracy and loss values for the test dataset. As visible from Figure 9, the neural network has a higher level of accuracy than the logistic regression; around 79 percent for a 100 topics.

Furthermore, the loss function in Figure 10 shows that there is a steady decrease in the loss from the model error, indicating that the model trained well[20]. Figures 11 and 12 show the validation set accuracy and loss. The errors, or losses in Figure 12, are spread around a mean and can be considered homoscedastic, indicating that the model is not overfitting the data. Thus the neural network improved upon the linear model's predictive power. It is also clear that a model with a 100 topics out-predicts models with a smaller set of topics. This is likely due to the extra

[19]The R-square value for the models with less topics are less than the R-square value for the 100 topic model.
[20]The model can be set to learn even slower with smaller batch sizes and more epochs, but this is left for future work.

Figure 7: Confusion Matrix: Top 20 Predictors for 100 Topics

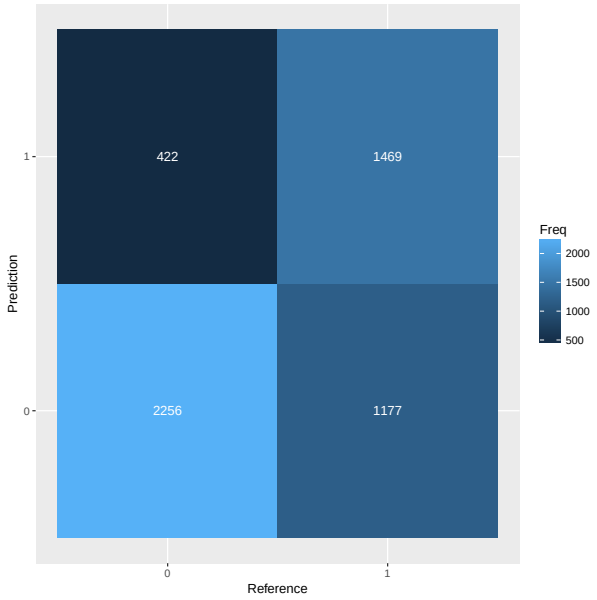
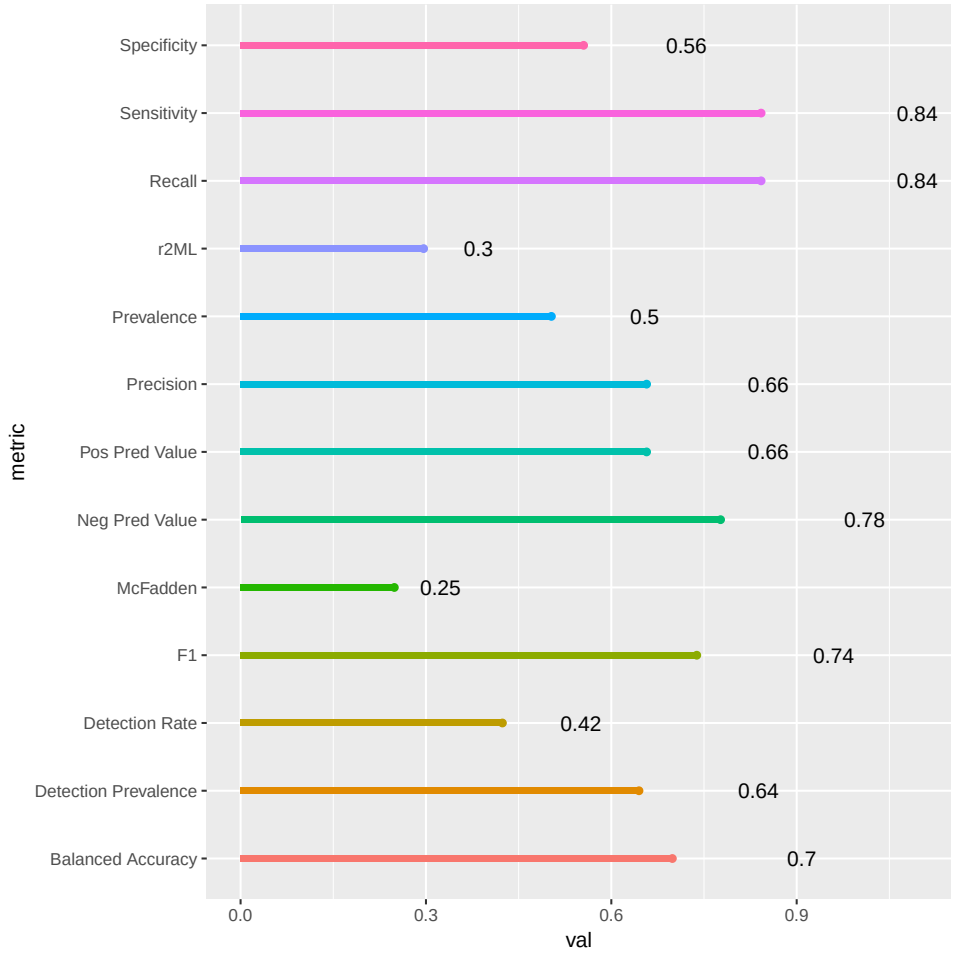


Figure 8: Auxiliary Statistics: Top 20 Predictors for 100 Topics



information given to the network. It also opens up the question of which amount of topics is optimal, however this is a question for future work.

Figure 9: Test Accuracy

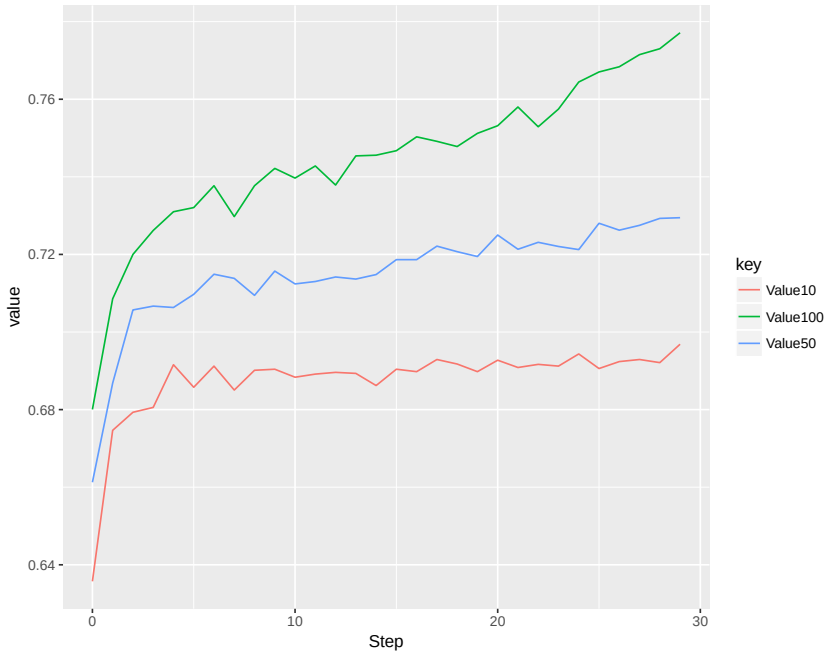


Figure 10: Test Loss

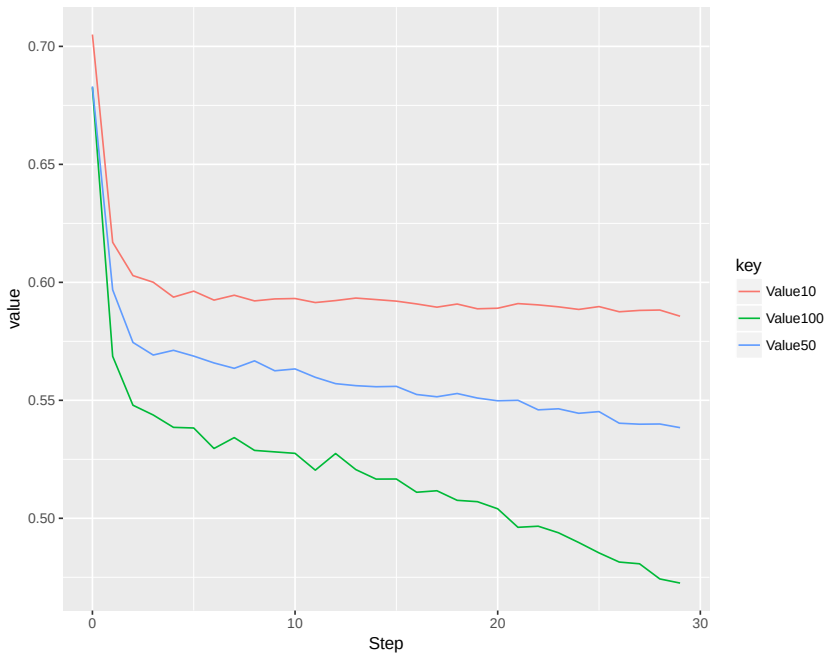


Figure 11: Validation Accuracy

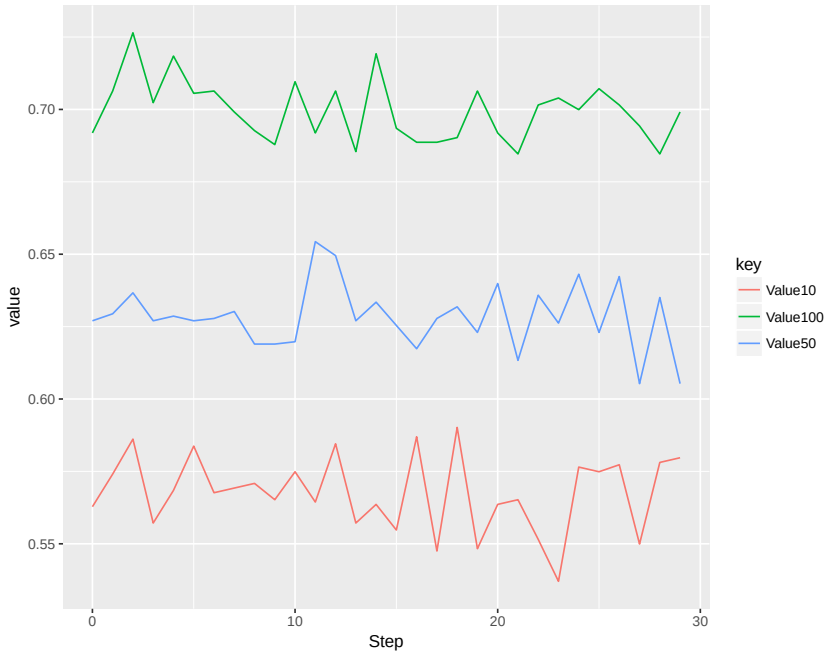
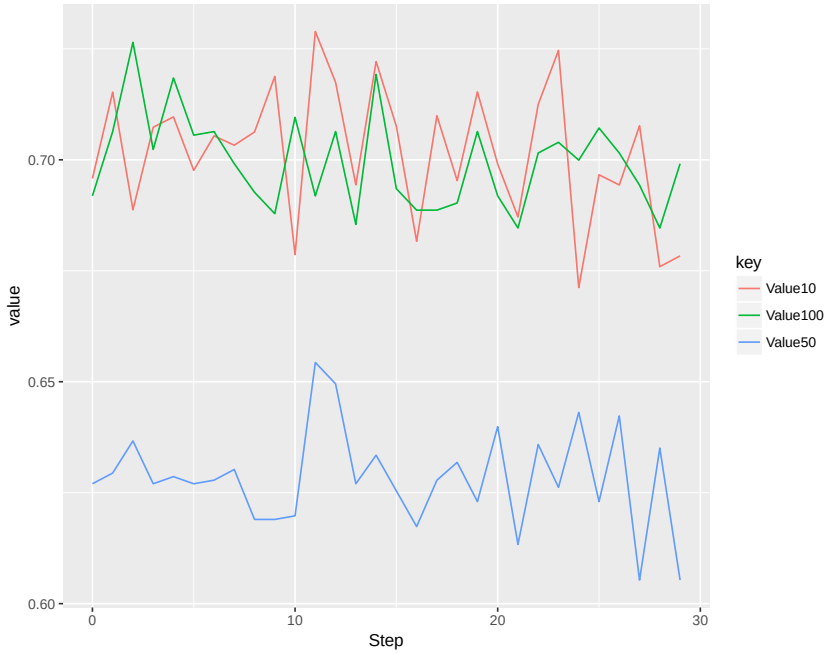


Figure 12: Validation Loss



6 Conclusion

Future work: find a different, preferably a sequential dataset to have RNN. Use the wiki situation but spend more time removing stop words specify a different predicted variable to detect whether consumer is utilitarian or hedonic Try to use a different measurement to relate consumer purchases with topics Use validation processes by having more computational power Test other models than just NN for prediction more research into neural networks for optimal prediction, as the increases in prediction accuracy was not much

References

Andreoni, J., & Miller, J. (2002). Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica*, *70*(2), 737–753.

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 391–402).

Batra, R., & Ahtola, O. T. (1991). Measuring the hedonic and utilitarian sources of consumer attitudes. *Marketing letters*, *2*(2), 159–170.

Christidis, K., Apostolou, D., & Mentzas, G. (2010). Exploring customer preferences with probabilistic topic models. In *European conference on machine learning and principles and practice of knowledge discovery in databases.*

contributors, W. (2018). *White bread.* Wikimedia Foundation. Retrieved from https://en.wikipedia.org/wiki/White_bread ([Online; accessed 9-February-2018])

Crowley, A. E., Spangenberg, E. R., & Hughes, K. R. (1992). Measuring the hedonic and utilitarian dimensions of attitudes toward product categories. *Marketing letters*, *3*(3), 239–249.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228–5235.

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30. doi: 10.18637/jss.v040.i13

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

Jacobs, B. J., Donkers, B., & Fok, D. (2016). Model-based purchase predictions for large assortments. *Marketing Science*, *35*(3), 389–404.

Ma, B., Zhang, D., Yan, Z., & Kim, T. (2013). An lda and synonym lexicon based approach to product feature extraction from online consumer product reviews. *Journal of Electronic Commerce Research*, *14*(4), 304.

Melville, P., Sindhwani, V., & Lawrence, R. (2009). Social media analytics: Channeling the power of the blogosphere for marketing insight. *Proc. of the WIN*, *1*(1), 1–5.

Okada, E. M. (2005). Justification effects on consumer choice of hedonic and utilitarian goods. *Journal of marketing research*, *42*(1), 43–53.

Sherry, J. F., Jr. (1983). Gift giving in anthropological perspective. *Journal of Consumer Research*, *10*(2), 157-168. Retrieved from +http://dx.doi.org/10.1086/208956 doi: 10 .1086/208956

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

Spangenberg, E. R., Voss, K. E., & Crowley, A. E. (1997). Measuring the hedonic and utilitarian dimensions of attitude: A generally applicable scale. *ACR North American Advances*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

(Stroie), L. M. B. (2014). Predicting consumer behavior with artificial neural networks. *Procedia Economics and Finance*, *15*, 238 - 246. Retrieved from http://www.sciencedirect.com/science/article/pii/S2212567114004924 (Emerging Markets Queries in Finance and Business (EMQ 2013)) doi: https://doi.org/10.1016/S2212-5671(14)00492-4

7 Appendix

7.1 Tables

Table 1: Variable Descriptions

Full dataset			
Variable Name	Variable Description	Example	Dimension
NEWID	NEWID is the variable inditcating each individual houseld survey	1292531	44746*1
UCC	UCC represents each products unique universal classification code	10120	548*1
Description	Description describes each individual product	Savings acc.	
N	The amount of times a product was purchsed by a particular household	1	1195829*1
Gift	Gifts represents whether a household purchased a gift or not	1,0	44746*1
AGE	The average age of a household	67	44746*1
EDUCA	The maximum level of education within a household	8	44746*1
STATE	The state in which the survey was conducted	51	44746*1
SEX	Whether the household is male or female dominated, or balanced.	M	44746*1
INCLASS	The inclome class of the household	>75000	44746*1

7.2 Figures

Figure 13: Top 10 Words Over 7 Topics

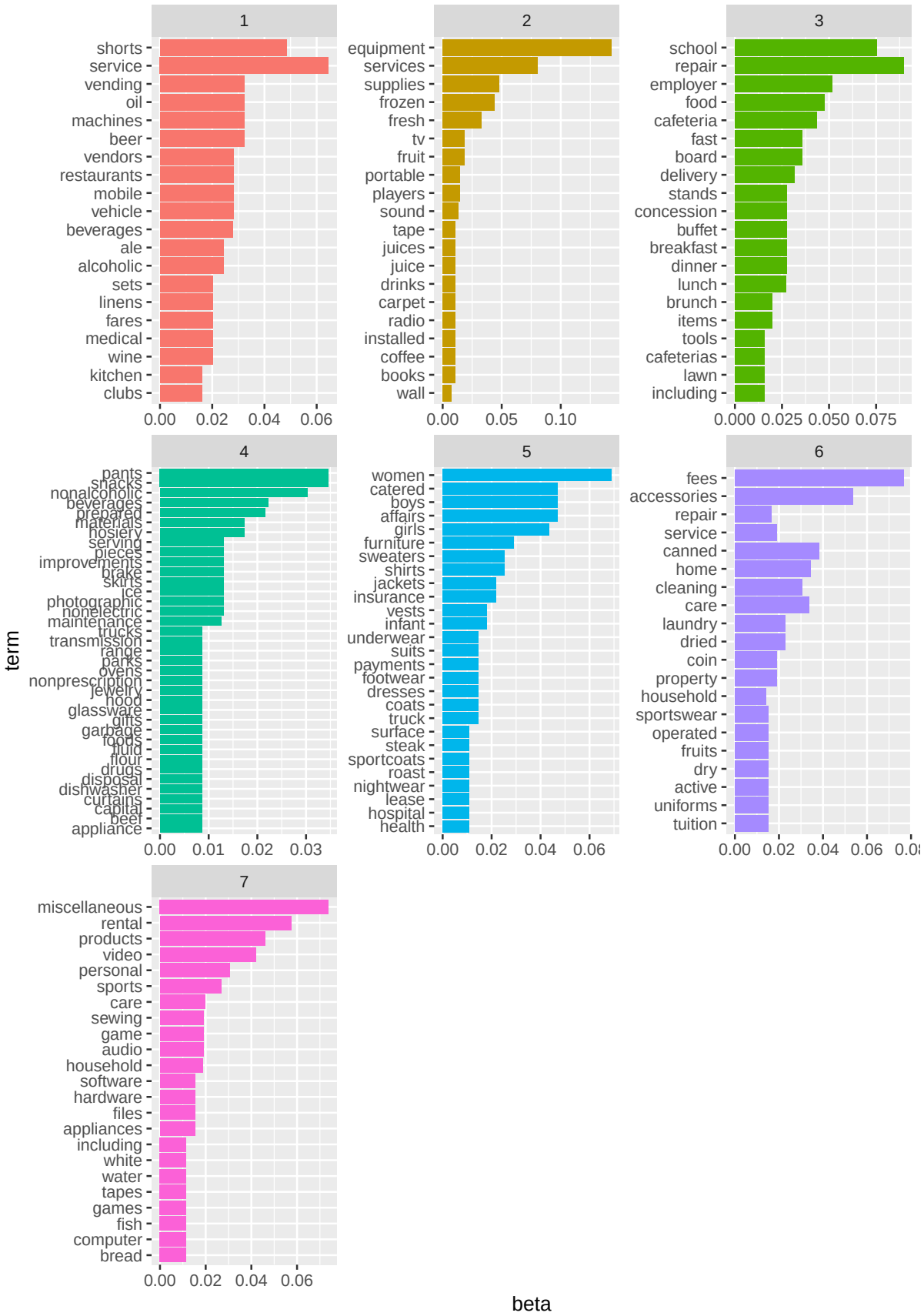


Figure 14: Word-Count Per Product

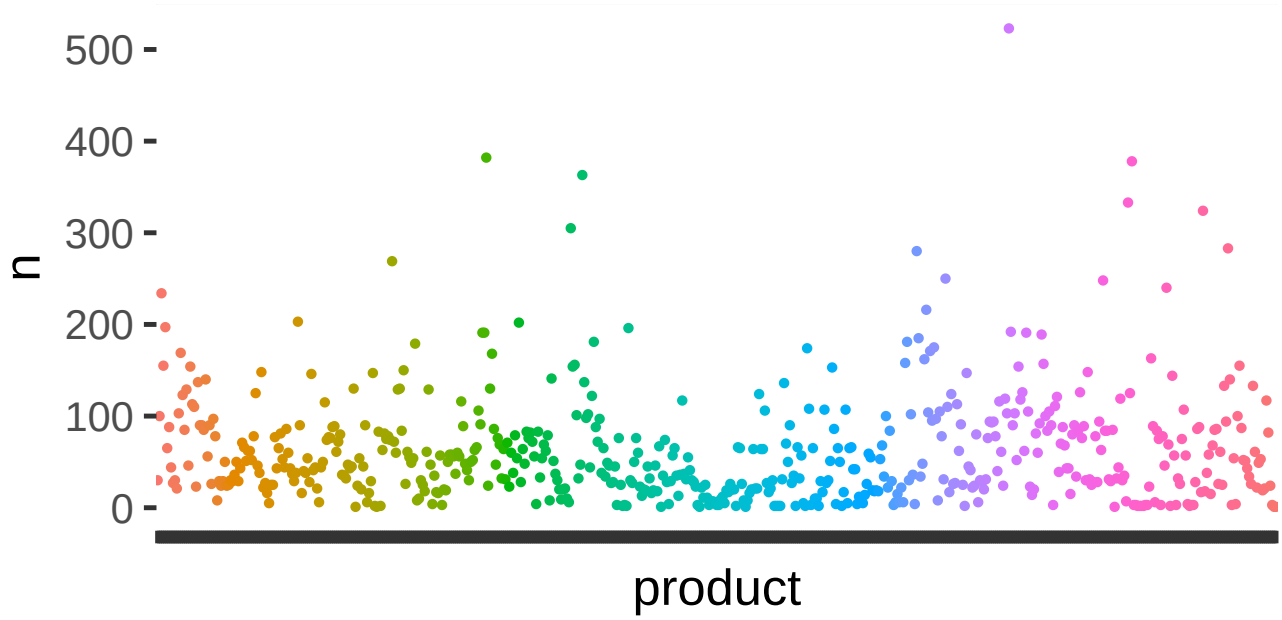


Figure 15: CPP Matrix Visualization

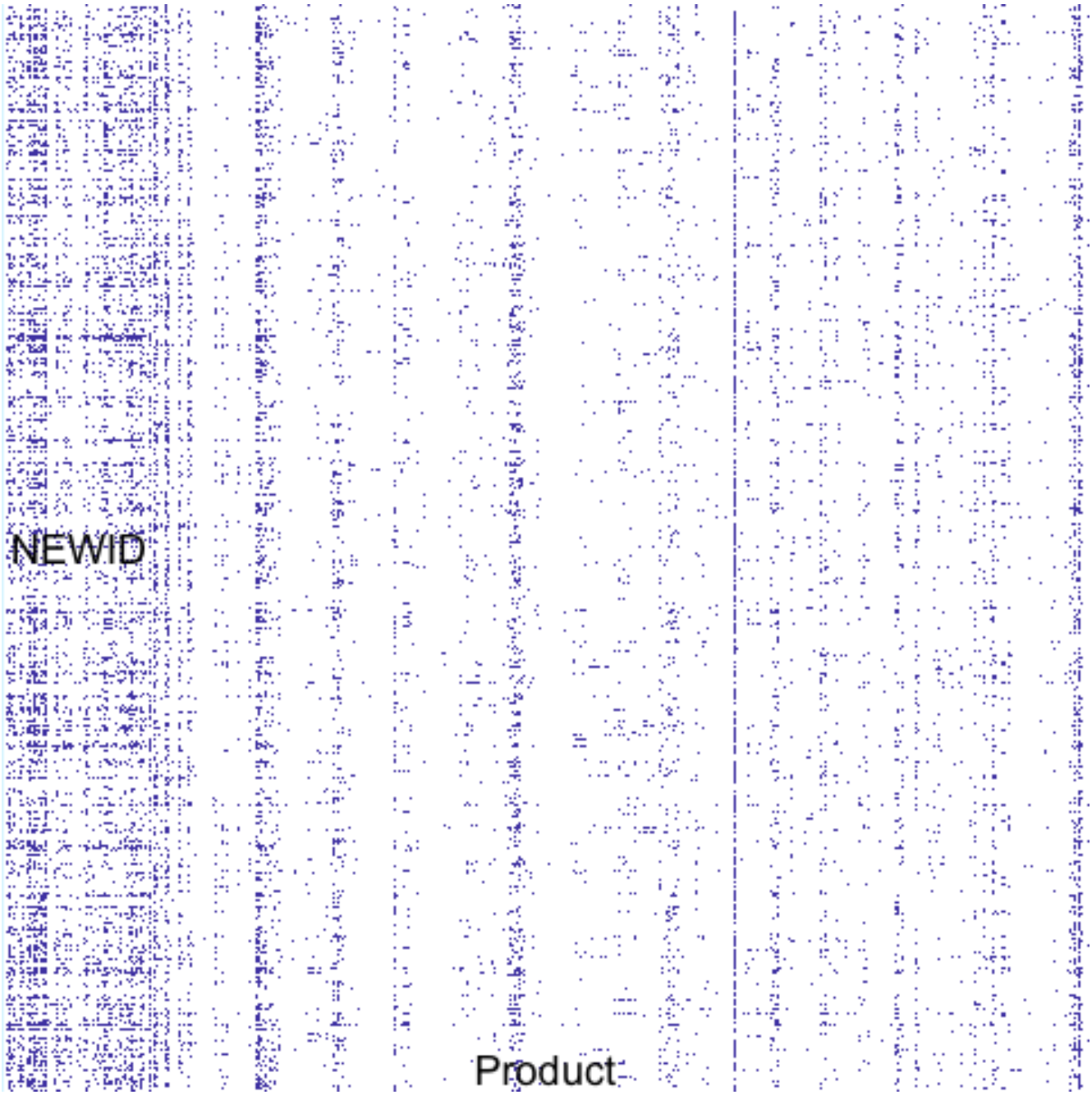


Figure 16: Distribution of Household Age for Gift and Non-Gift

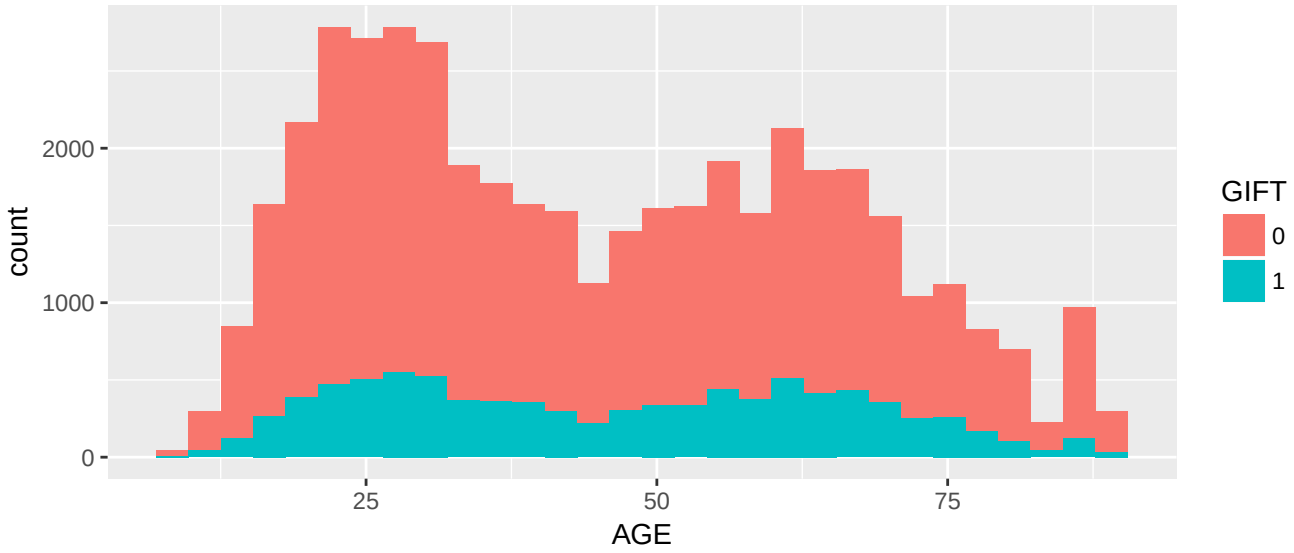


Figure 17: Distribution of Household Sex for Gift and Non-Gift

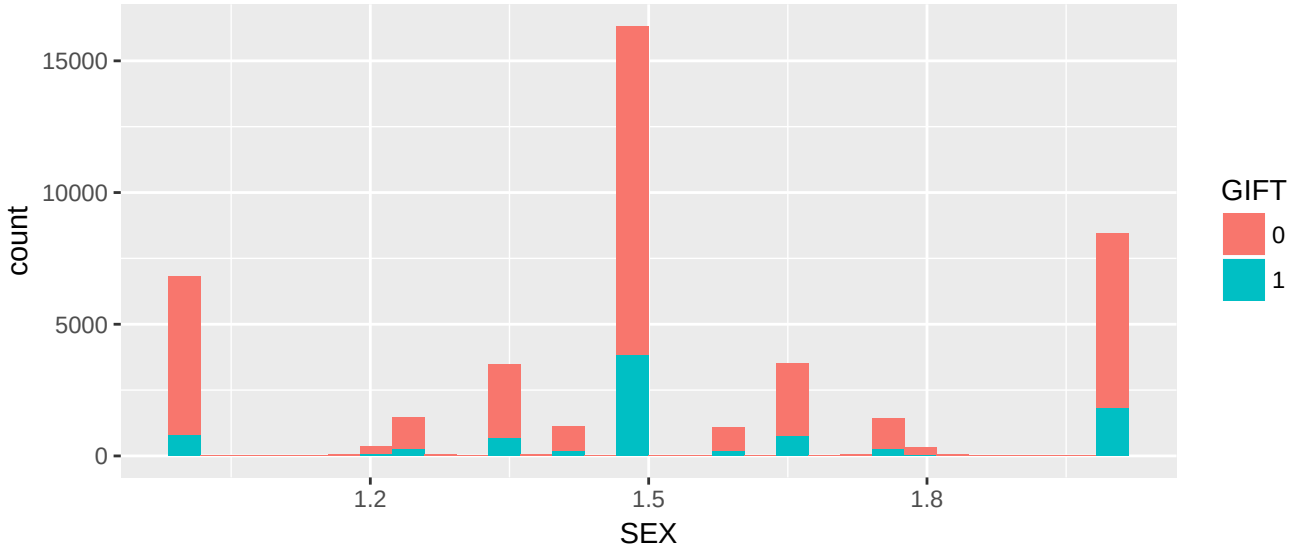


Figure 18: Distribution of Household Education for Gift and Non-Gift

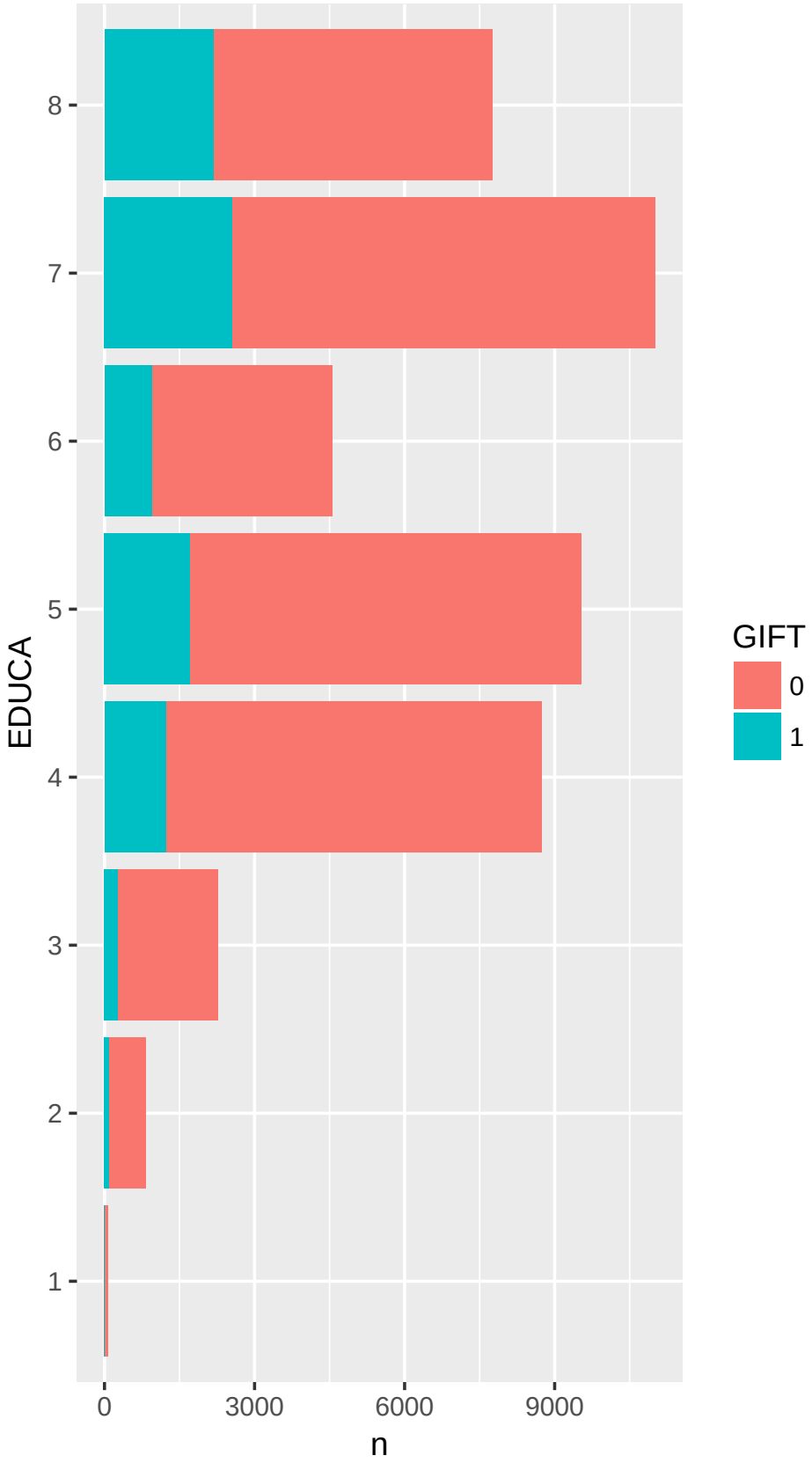


Figure 19: Distribution of Household Income Class for Gift and Non-Gift

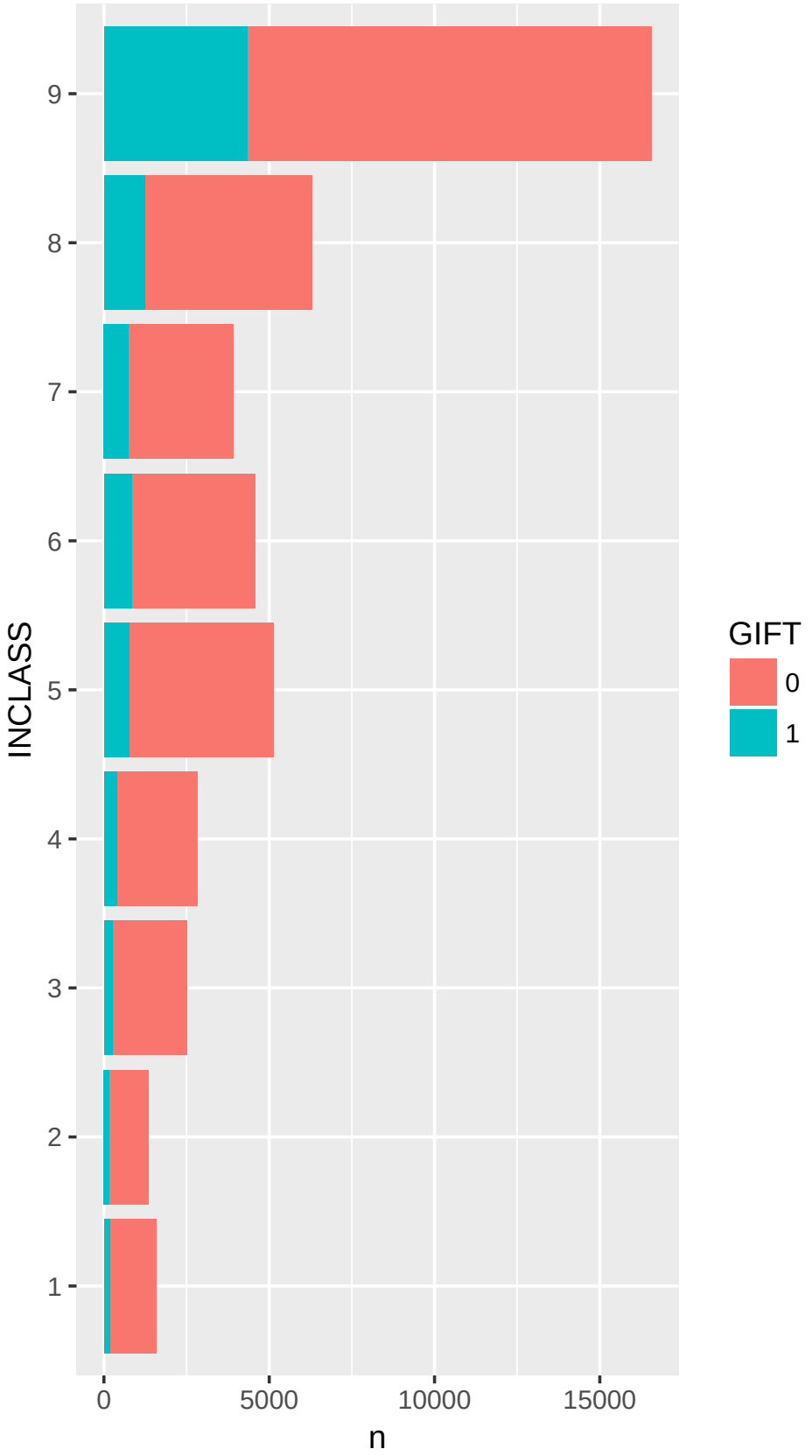


Figure 20: Distribution of State for Gift and Non-Gift