Markets for Consumer Data

Ayush Kanodia
Computer Science
Stanford University
akanodia@stanford.edu

Lina Lukyantseva
Graduate School of Business
Stanford University
linaluk@stanford.edu

1 Introduction

Information on buyers' purchases history helps the sellers to make better personalized offers and predictions and, hence, maximize profit. Because this information is valuable, it oftentimes gets sold by consumer companies in the form of datasets. While the data industry is booming, it is still very hard for both buyers and sellers of the data to price it correctly since it is not clear how to measure the value of such datasets for the companies. It is also an open question to which extent this information should be anonymized in order to protect privacy of the consumers. We look into the proposition to not allow to disclose customer ids but to allow to cluster customers into bigger groups and disclose what cluster a given customer belongs to (for example, in March 2021 Google announced the replacement of third-party cookies with a new clustering-based system called FLoC). In this project we pose two questions: 1) clearly, only knowing to which cluster a user belongs is less valuable than knowing the id exactly; we are interested in quantifying to which extent this is the case. 2) A particularly challenging task is to make predictions for a user's purchase in a given category if that user never bought anything from that category before. We compare the performance of two different models (one modeling the consumers behaviour explicitly based on economic theory, and another one using deep learning) under privacy restrictions (i.e. clustering).

We are using a dataset on purchases from the grocery stores (see detailed description in the Methods section). We start by splitting the data into train, validation, and test, and assuming that user ids are available. We are interested in comparing different models' predictive power in this setting. As a baseline model, we use a model of random choice: for a given user and a given category we predict user's choice as a randomly chosen item from that category. As main model 1 we use a structural economics model based on Bayesian Random Utility model which outputs for a given user and a given category the probabilities of every item in that category being chosen. As main model 2 we use a neural collaborative filtering model which also outputs for a given user and a given category the probabilities of every item in that category being chosen. We use loglikelihood as the measure of model performance, i.e. the sum of logarithms of predicted probabilities for the items that were actually chosen. We find that the deep model performs best in this setting, followed by the economics model which is better than the random model (with the corresponding loglikelihoods of ???).

We proceed to answering question (1) next. We assume that user ids are no longer available, and instead the seller can only observe to which cluster a given user belongs. We consider several restrictions on the number of clusters: 0 (which correspond to a regulation that does not allow to identify a given user in the dataset in any way), 2 (which corresponds to a regulation that allows to divide the dataset into halves and tell the seller which half a given user belongs to), 3, 4, 5, 10, 50,

CS229: Machine Learning, Spring 2021, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

250. We find that both for deep model and economics model the predictive power goes up as we allow for more and more granular clustering.

Finally, we answer question (2). In order to do so, we randomly pick 50 users and 2 categories, and we remove the information about purchases of those categories for those users from the train data, i.e. we make it look as if these users never purchased anything from these 2 categories before. We call these users and categories salient. We consider the same cluster sizes as in question (1) and analyze the predictive power of both deep and economics models for both salient and non-salient users and items under different clustering restrictions. Our main findings are that even though for non-salient users and non-salient items we continue to observe that the predictive power goes up as the granularity of the clustering increases, it is no longer the case if we consider salient users or salient categories. It turns out that for salient users/categories the predictive power is non-monotonic in the clusters size. This phenomenon needs to be studied further in order to understand better what exactly is driving this result. Our hypothesis is that the size of salient dataset is small which is why we see more variance in these results. It is also possible that because of that the model overfits when the clusters are too granular. A natural next step would be to verify this hypothesis by training the models on the full dataset instead of the dataset that we downsampled for this project.

2 Related Work

The literature on retail choice modeling is rich in both the Economics and Machine Learning/Computer Science disciplines. [1] is a standard reference to understand Discrete Choice Modeling, which is the most common class of methods used in Economics. [2][3] are state of the art Machine Learning models embellished with rich Structural Economic information, which return both embeddings as well as Economic indicators such as price elasticities. More recently, [4] is a foray into using more powerful non linearities using Deep Neural Networks for the same problem, and they are shown to acheive higher predictive accuracy although at the cost of losing clear Economic interpretation of models. These incorporate ideas from [5], seminal ideas in modern recommendation systems.

3 Data

A novel contribution of our paper is that we segregate train and test sets in systematic ways rather than by pure random sampling. We describe this further:

3.1 Data Cleaning and Downsampling

We refer the reader to our milestone report for details.

3.2 Salient Users and Categories

We pick randomly 50 users and 2 categories which we define as salient users and salient categories. The problem setup is that we do not observe interactions of these salient users over these salient categories, and these are the new categories and new users who we would like to predict choices over given our dataset, as in the motivation. We end up with two types of users, salient and non salient, and we call them **SU** and **NU** respectively. Likewise, we have two sets of items **SI, NI** and two sets of categories **SC, NC**. As described, there are 2 **SC** categories, and 8 **NC** categories, and **SI, NI** are the respective item sets in these categories.

3.3 Non Salient Categories

Our non salient categories are *Chana Dal* (a lentil),*Guards*, *Instant Noodles (vegetarian)*, *Leafy Vegetables*, *Melon*, *Other soft vegetables*, *Misc Popular Items*, Sugar

3.4 Salient Categories

Our salient categories are *Potato*, *Detergent Bar*

3.5 Data Splits

We have the following data splits:

- Train: We take the following data **NU** over **SC, NC** and **SU** over **NC**. We randomly split this data into train and validation, and test, 60, 20, 20 percent. Importantly, we hold out interactions between salient **SU** and salient **SC** categories.
- Validation: We take all validation data points as above.
- Test-NU-SC: From the test split defined above, we take only the interactions of non salient **NU** users over salient **SC** categories. This is a reference set to compare our predictive accuracy between salient and non salient users
- Test-SU-SC: We take all interactions between salient **SU** and salient **SC** categories. This is our main target set.

4 Methods

We model the conditional probability of purchase given that a consumer purchases a product in a category. As such, we write a modified loss function which takes into account only those instances where a customer purchased within a category. We note that was a proposed improvement in the milestone as the full problem runs into the problem of sparsity. This is a standard formulation of the problem, as for eg in the lower level of the logit choice model in [1]. We try two different classes of models for our dataset.

4.0.1 Structural Economics Models

This is a model that borrows heavily from Economic Theory to setup a modified Logistic Regression style Multinomial Logit formulation. It assumes that consumers are making decisions with the goal of maximizing their utility, and that utility is observed with some errors. The model is broadly similar to the model in [3], but with significant differences. We refer the reader to [3] and our milestone report for a description as detailing it further would make this paper too long.

4.0.2 Deep Learning Multitask Models

We learn from [4] that one of the best ways to model our choice problem using purely predictive methods is to use deep learning recommendation systems, in the same vein as [5]; except that we should encode our assumption in such a model that the consumer shall purchase at most one item from a category. We note that this is a data tested assumption (true 99% of the time), and this makes the predictive problem much simpler. This leads to a multi task deep learning embeddings model; it is very similar to the set of final models proposed in [4] which we refer the reader to; we skip model definition here in the interest of space. Our models differ from the models referred in model architecture, embeddings dimensions and regularization paramaters.

4.0.3 K Nearest Neighbours for clustering

We used the K Nearest Neighbours Algorithm for clustering users; for this purpose, we used the use embedding vectors given to use by our Structural Economic Model. This is an important design choice - we could use other methods of clustering such as topic modeling (LDA for eg [6]), models such as GLoVe [7] and so on alternatively. These would require alternative setups for the user item purchase history information and are left for future work

5 Experiments and Results

5.1 Structural Economic Models compared to Deep Learning Models

We first compare predictive accuracy between our structural economic model and our deep learning models

- trained on our dataset

- tested on our salient user/salient category dataset, to check the gains to predictive accuracy by using data on unrelated stores.

We use the following abbreviations for our models:

- **SER**: Structural Economic Model with Random Parameters
- **SE**: Structural Economic Model
- **MTL3**: Multitask learning model with 3 hidden layers

We show our results in Table 1

Table 1: Multi-Task Learning with Embeddings

	Train LLH	Validation LLH	Test LLH
SER	-0.15	-0.15	-0.15
SE	-0.048	-0.062	-0.083
MTL3	-0.055	-0.100	-0.3

We note that our **SE** model is outperformed by our **MTL** model as evidenced in [4] in terms of predictive accuracy. Both models also far outperform a random model (**SER**) which gives us a baseline sanity measure.

We conclude from this analysis that our MTL models are well calibrated; further, the structural economic models do not retain their structure with the analysis under cluster indicators that we intend to do. The following analysis works with tuned MTL models, based on different user data specifications such as full information, no information or cluster information. Our tuned MTL models have 2 hidden layers, with 1000 and 800 neurons each, and like [5] we interact user and item embeddings and also concatenate them into a giant input vector.

5.2 Performance under different information specifications

We now compare predictive performance under full information, partial information and no information. We note that clusters equals 0 is the case where we provide no consumer information at all - so the predictions will be equivalent for everybody; this is equivalent to a popularity model. We also note that clusters equals 500 is the case where we provide full information as we have 500 customers.

Table 2: Loss (negative LLH) vs number of clusters and splits for salient category **Potato**

clusters	Test-SU-SC	Test-NU-SC	Validation	Train
0	0.1423	0.1383	0.1453	0.1758
2	0.1478	0.1355	0.1432	0.1705
3	0.1440	0.1341	0.1408	0.1662
4	0.1414	0.1333	0.1390	0.1673
5	0.1391	0.1293	0.1369	0.1647
10	0.1842	0.1252	0.1311	0.1545
50	0.1626	0.1170	0.1241	0.1347
250	0.1807	0.1051	0.1103	0.1216
500	0.1815	0.1063	0.1116	0.1166

We note the following important points:

- The losses have very different absolute values for Potatoes and Detergents; in general this is true across categories. This is because each category has a different number of items and different frequencies of purchases for different items. A higher loss means more choices within the category, typically. This makes sense as there are typically far more varieties of detergent than potato.

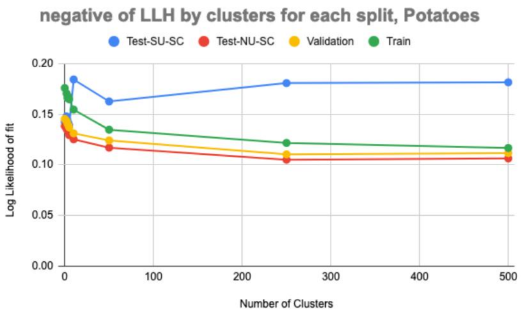

Figure 1: Potato

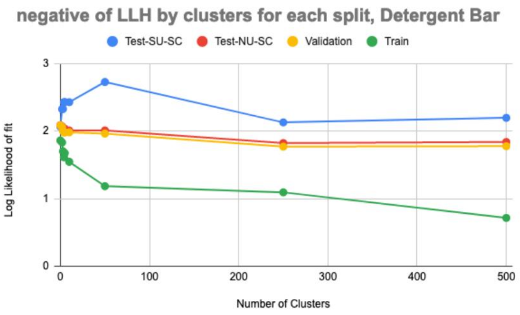

Figure 2: Detergent Bar

Table 3: Loss (negative LLH) vs number of clusters and splits for salient category **Detergent Bar**

clusters	test	test_non_salient_users	validation	train
0	2.0670	2.0874	2.0854	1.8609
2	2.3251	2.0691	2.0578	1.8318
3	2.3271	2.0343	2.0007	1.7004
4	2.4270	2.0248	1.9815	1.6130
5	2.4299	2.0347	1.9787	1.6739
10	2.4277	2.0088	1.9821	1.5457
50	2.7264	2.0099	1.9626	1.1843
250	2.1286	1.8226	1.7700	1.0932
500	2.1959	1.8387	1.7763	0.7147

- Loss on the Train split falls with more clusters - this is expected as this means we are providing more user information to the model
- Loss on both the Validation and the Test-NU-SC categories also falls. These sets are qualitatively similar as no interactions from these sets are systematically held out in train. This tells us that we perform better on held out data too when we show more information to the model
- The returns are diminishing with increased number of latents. For eg. we get approximately the same Loss with 250 clusters as with 500 (full information), and get very close with even 50 clusters
- For potato, train loss is marginally higher than validation loss - we note that this is due to random sampling and the difference is not statistically significant (we analyze this separately, do not go into details here)
- Most interestingly, we see that the performance on Test-SC-SU is not crystal clear. Performance falls, and then rises for potato, while it rises, then falls before rising again for detergent bars. For the latter, showing the highest amount of information leads to the highest predictive accuracy. For the former, showing an inbetween number of clusters is optimal. We hypothesize that these maybe either due a> data sparsity - more data would help or b> we use neural networks for our modeling which are not designed to predict on systematically different datasets than models informed with economic structure as in [3] designed for counterfactual inference.

6 Conclusion and future research

In this project we compared the predictive power of two different models – a structural econometric model and a deep learning model under different privacy regulations that require data sellers to not reveal user ids and only allow to disclose cluster assignments. We compared the performance of these models in the setting when we have to make predictions for a new user-category pair that has not been observed in the dataset previously. In this project we estimated the value of the information in terms of loglikelihood of the predictions. Of course, in order to price the datasets in the real world we need to be able to translate that value into dollars. This is an important direction for our future research. This would be a large extension of this work. Further, we note in the results that it might be promising to use Structural Economic Models especially because we have systematic held out data which resembles a counterfactual setting.

7 Contributions

As true students of Economics, we demarcated our individual comparative advantages before we started the project and settled down on individual experiments. It turned out that our contributions were statistically indistinguishable from equal.

References

[1] Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

[2] Susan Athey, David Blei, Robert Donnelly, Francisco Ruiz, and Tobias Schmidt. Estimating heterogeneous consumer preferences for restaurants and travel time using mobile location data. In *AEA Papers and Proceedings*, volume 108, pages 64–67, 2018.

[3] Robert Donnelly, Francisco J. R. Ruiz, David M. Blei, and Susan Athey. Counterfactual inference for consumer choice across many product categories. *CoRR*, abs/1906.02635, 2019.

[4] Ayush Kanodia and Sakshi Ganeriwal. Deep consumer choice models.

[5] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. *CoRR*, abs/1708.05031, 2017.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[7] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.