Purchase Likelihood Prediction for Targeted Organic Food Marketing Campaigns in China

Beau Giannini[1], Song Chen[1], Pavel Paramonov[2], Ying Yu Wu[1]
[1]Tongji University, School of Economics and Management, Shanghai 200092, P. R. China
[2]Multi-disciplinary Insights LLC, San Francisco, CA 94108 USA

Abstract--**The demand for organic food products in China is growing in response to both increased spending power and food safety concerns. However, identifying likely buyers of organic products proves challenging due to their relatively small fraction in the overall population. Our study explores applications of machine learning algorithms for effective management of organic food marketing campaigns in China. Based on the data we collected through an online choice-experiment type questionnaire of Chinese consumers, a purchase likelihood estimation framework has been developed that utilizes customer profile traits such as age group, family status, education level, and geographic location. In addition, we apply clustering techniques to perform data-driven organic market segmentation and identify consumer profiles ready to pay more for high quality, certified organic products. The resulting market segments are compared to various types of organic consumers discussed in the literature. Our algorithms provide a useful framework for online retailers who are seeking lean strategies of market entry in China with their health food brands.**

I. INTRODUCTION

Food quality and safety considerations have been gaining considerable importance among Chinese consumers [1], [2]. Together with rapid socioeconomic development and continuously growing spending power, this triggered an upsurge in the number of food offerings [3] and stimulated the expansion of the organic food market [4]. While a major part of the organic food industry in China is focusing on export to developed countries, organic products are increasingly being sold domestically and imported to China [5].

Identifying likely buyers of organic food products is particularly helpful for organic/health food businesses that seek market entry or try to expand their current market share in China. A general survey conducted in three Chinese cities in 2010 suggests that consumer factors such as the income, trust in organic certification, and concern regarding self-health are essential for the intent to purchase organic products, while the age, education level, and environmental considerations are less significant in determining food choices [6]. It is important to note that purchase intention is also influenced by specific characteristics of a given product, and such factors are particularly influential in the case of food purchases.

Organic attributes of a given food product have an inherently high uncertainty from a consumer perspective since it is not easy to verify whether organic principles were followed in the course of production and packaging processes. The information is unevenly distributed between the supplier and the consumer, leaving considerable room for fraud. This places a very heavy weight on organic certifications as a way to gain consumer trust [7], with governmental and third-party certifications being generally the most effective tools for this purpose [8].

In European countries several organic labeling schemes are present in the market. A recent study [8] explored through interviews and choice experiments whether consumers give preference to some of those schemes over others. Considerable differences in the willingness to pay were found between different product labels/logos. The authors also concluded that consumer perception was mainly subjective in its nature, not relying on objective knowledge [8]. A conceptually similar study was conducted earlier in the United States, where the United States Department of Agriculture (USDA) seal was compared with a general "organic" label on meat products [9]. USDA certification was found to enjoy a significantly higher level of consumer trust, which translated into consumers' acceptance of higher prices. The study [9] is also particularly useful in pointing out variations between the types of consumers with respect to their frequency of organic purchasing, generally suggesting higher willingness to pay in the case of frequent organic shoppers.

In contrast to European countries and the United States, a number of consumers in China still do not have significant knowledge around organic food [1], and current organic certification procedures [5] have not yet achieved high levels of consumer trust. In spite of these factors, Chinese consumers hold generally positive attitudes toward safe food and organic products in particular, and may be willing to pay more to purchase such products [1], [6]. However, the majority of Chinese consumers are only occasional buyers and show a strong inconsistency between their attitudes and actual behavior [1], which somewhat devalues generalized conclusions derived from broad surveys and calls for more individualized analyses of consumer choices.

In this application study we explore approaches to predict the likelihood of organic food purchases for a given individual provided a set of consumer traits available in a common marketing campaign setting along with key product attributes such as the price and a specific labeling scheme. The primary goal is to illustrate the corresponding methodology with a particular case study and stimulate future developments applicable to a wide range of products and

markets. We use an online product choice experiment questionnaire to generate a dataset containing consumer traits and recorded product choices, and develop machine learning tools for predicting individual choices from consumer traits. Our development is driven by the hypothesis that significant marketing campaign revenues may be achieved using appropriate data management techniques even if consumer information in the dataset is fairly scarce.

Several previous studies reported large-scale applications of machine learning methods to direct marketing [10], brand choice prediction [11], and online advertising [12]. In the current work, we seek to build predictive algorithms for purchase probability that are based on a *limited number of consumer characteristics* and work with relatively *small datasets*. This way, rather than trying to drill down to the details of customer segments as frequently done in broad market surveys, we emulate a situation commonly experienced by an entry-level company in direct marketing and online advertising when only a few things are known about the target customer/website user and that data must be put to use for effective campaign management. This falls within the scope of the technology management discipline as it relates to utilizing online data collection technologies and statistical learning approaches to optimize marketing strategy decisions and gain competitive advantages.

In addition to the development of purchase likelihood prediction algorithms, the objectives of our application study include a data-based assessment of the value of such predictions for guiding targeted organic food marketing efforts in China, with a particular emphasis on paid online advertising in mind. We demonstrate multiple-fold projected revenue gains and quantitatively show predictive value of even a fairly limited set of consumer traits in estimating purchase probabilities for organic produce and dairy products.

II METHODOLOGY

The basic design of our application study involves (i) data collection through an online choice experiment style questionnaire, (ii) machine learning algorithm development, optimization, performance evaluation using the acquired dataset, and (iii) estimation of projected marketing campaign revenue gains corresponding to our specific predictive algorithm implementation. In this section we outline each of these steps, attempting to provide a sufficient level of pertinent details for our approach to be readily adopted in the relevant application domains of consumer data management.

A. Choice experiment questionnaire

The participants of our choice experiment were offered to select a product among several options, which, as opposed to e.g. open-ended survey questions, closely resembles a real situation that they face when shopping online or in a physical store setting. In order to facilitate a focused study, the respondents were presented with product options that differ only in label information and price. Such an approach provides an opportunity to study consumer preferences in direct relation to the product nature (organic or not) and a specific certification labeling scheme separately from other product attributes, and is a generally well-established tool for predicting consumer trade-offs [8], [9]. On the other hand, consumer traits such as the age, gender, education level, and family status varied between the study participants allowing for representation of different market segments. Additional questions about organic shopping habits were also included but not used as predictive algorithm input variables. An established "cheap talk" script was embedded into product choice screens in order to mitigate hypothetical bias in consumer preferences [9]. The full set of questions is presented in the Appendix along with product selection options in the choice experiments. The questionnaire was distributed in China online via email invitations, and fully anonymous responses were used to form the product choice dataset.

Choice experiments were conducted with two different kinds of products, namely apples and milk. These products are widely distributed in China from both domestic and import sources, regularly bought by many consumers, and available in more than one variety of organic quality. Each selection set contained a domestic as well as imported organic product along with a conventional (non-organic) counterpart. Well-recognized brands were chosen for both domestic and imported products with their typical retail market prices at the time when the questionnaire was administered.

B. Machine learning algorithms

Two types of machine learning methods, namely (i) unsupervised clustering and (ii) supervised classification, were applied to the collected dataset to perform market segmentation and purchase probability prediction, respectively. While unsupervised methods (i) utilize input variables to group consumers by similarity, supervised learning approaches (ii) rely on a set of provided examples with their outcomes (purchase decisions) to predict expected outcomes of new cases.

Feature vectors used by both types of learning methods were encoded from the consumer traits as described in Table 1. A feature vector for a given consumer is composed of discrete and continuous numerical variables bound between 0 and 1. The age variable was based on pre-defined age intervals taken as a difference ($Age - Age_{min}$) and normalized by the full range ($Age_{max} - Age_{min}$). In order to capture the effect of consumer location, city average gross domestic product (GDP) values were used (as of 2012 [13]), taken relative to the maximum city GDP in China (i.e. the GDP of Shanghai).

TABLE 1. NUMERICAL ENCODING OF CONSUMER CHARACTERISTICS FOR MACHINE LEARNING INPUT. BOTH CONTINUOUS AND DISCRETE SCALES WERE USED WITH THE RANGES BOUNDED BETWEEN 0 AND 1.

Consumer trait	Options	Feature encoding	Type of variable
Age	Numerical	$(Age - Age_{min})/(Age_{max} - Age_{min})$	Continuous
Gender	Male, Female	0, 1	Discrete
Education	Groups ranging from "some high school" to "doctorate"	From 0 to 1 with equal increments between groups	Continuous
Children	Yes, No	1, 0	Discrete
Traveled abroad	Yes, No	1, 0	Discrete
Health problems	Yes, No	1, 0	Discrete
City	City names	(City GDP) / (Shanghai GDP)	Continuous

Unsupervised learning algorithms for market segmentation were based on a hierarchical clustering (HC) technique [14]. Kernel-based principal component analysis (kPCA) [15] with a Gaussian kernel function was performed prior to HC in order to provide an efficient similarity measure between feature-vectors and achieve more informative clustering. We also explored K-means clustering as an alternative market segmentation technique but found HC to produce better groupings in terms of a qualitative understanding of predominant consumer traits.

1) Supervised learning with imbalanced classes

Marketing campaign management frequently encounters a situation when the number of actual buyers is significantly smaller than the number of marketing attempts. This is an objective feature of many direct marketing campaigns inherent to the nature of the marketing task, including online advertising [12]. From a supervised learning perspective, the corresponding binary classification problem (predicting buyers vs non-buyers) may encounter a severe class imbalance in the training data. The class imbalance problem significantly impacts the effectiveness of commonly used supervised learning methods and calls for special care in algorithm training as well as performance evaluation [16]. When imbalanced data are used for model fitting (learning algorithm training), a classifier generally becomes biased towards the majority class unless special care is taken to compensate for such a bias.

Various approaches to remedy the class imbalance problem involve either data splitting and balancing prior to the model fitting, or algorithm-specific tuning designed to compensate for the imbalanced training set [16]. Recent developments in response modeling for direct marketing offer solutions that handle class imbalance and work well when a considerable amount of data is available [17]. However, such methods are not readily applicable when the number of respondents in the dataset is relatively small.

In the current application study we explored data balancing prior to model fitting as well as cost function weighted learning methods that compensate for the training set imbalance. For data balancing, we evaluated two approaches, namely a relatively well-established synthetic minority oversampling technique (SMOTE) [18] and a more recently developed random over sampling examples (ROSE) method [19]. In the case of cost function weighted methods, cost-sensitive support vector machine (SVM) and classification and regression trees (CART) were used [16], which are powerful and flexible classification methods that rely on non-linear decision boundaries and recursive partitioning, respectively [14], [20]. SVM hyperparameter tuning was performed using 3-fold cross-validation to ensure a sufficient number of minority class points in each fold of imbalanced training sets. All learning models were implemented using *caret* package [21] in the R statistical computing environment [22].

2) Performance evaluation for predictive algorithms

Two main challenges for predictive model performance evaluation in our study are class imbalance and the small size of the dataset. Class imbalance prevents using performance measures such as the overall classification accuracy (or misclassification rate) [16], while a relatively small number of respondents does not allow for setting aside a separate testing set [11] and complicates the use of resampling techniques [23], [24]. To address both of these challenges, we used receiver operating characteristics (ROC) as a classification performance measure in combination with a leave-pair-out (LPO) cross-validation procedure [25]. In essence, ROC (more specifically, the area under the ROC curve) measures the likelihood (as probability) of a classifier to correctly distinguish a randomly selected pair of examples belonging to different classes and works well for imbalanced cases [16], while LPO resampling minimizes evaluation bias in small datasets and has been recently shown to outperform a number of other measures including leave-one-out cross-validation [25]. It is important to note that LPO cross-validation evaluates predictive performance on *unseen data exclusively*, i.e. the evaluation pair is always excluded at the model fitting stage.

C. Targeted marketing campaign

In order to utilize purchase likelihood predictions in marketing campaign management, a selection mechanism for *Yes/No* marketing decisions has to be defined. Here we

implement a relatively straightforward selection based on a single threshold probability parameter p^* that can be chosen by a campaign manager. Specifically, marketing is attempted with those consumers that fall within a region of purchase likelihood bound by p^*.

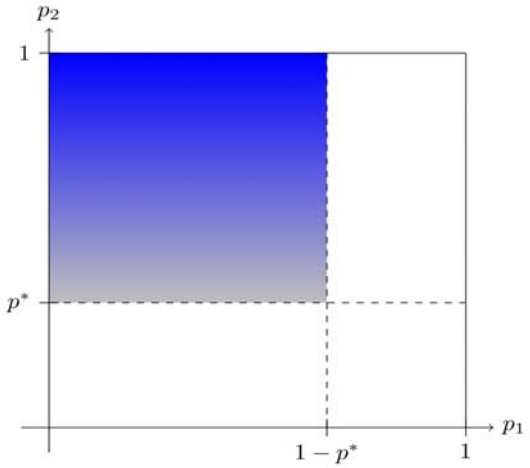

Fig. 1. Consumer selection approach adopted in the current study; p_1 and p_2 are the no-purchase and purchase probabilities, respectively, obtained from different classifiers. A single threshold parameter p^* is chosen by a campaign manager, which may also be optimized to maximize the campaign revenue.

When more than one predictive algorithm is used to achieve better results, a similar selection is done for each classifier. The selection region is illustrated in Fig. 1 for the case of combined purchase and no purchase classifiers: marketing is attempted for consumers with predicted purchase probability exceeding p^* and no-purchase probability below 1-p^*. Note that in general, more complex and adaptive approaches may be implemented for consumer selection based on purchase likelihood predictions, which is outside the scope of our application paper.

III. RESULTS AND DISCUSSION

Consumer information and product selections were collected from 193 valid responses submitted through our self-administered online choice experiment questionnaire. As already mentioned in the introduction, we worked with a relatively small sample in order to develop information management tools and suggest marketing campaign optimization strategies for small businesses entering the growing organic food market in China. Such companies may have minimal number of records in their customer relation management databases and yet could benefit from using that data from the very beginning of their market entry. Therefore, our dataset illustrates a small but realistic sample of potential consumers within the reach of an entry-stage online retailer.

Age distribution and gender composition of the questionnaire respondents is presented in Fig. 2. The dominant age group is 25-34 years, followed by 35-44 and 18-24 years, which appears to be fairly typical for Internet users and online shoppers in China [26]. The largest number of respondents hold bachelor-level degrees (Fig. 3), also in agreement with other studies that profile e-commerce and mobile application users [26], [27]. Shanghai and Hangzhou are the most represented cities in the dataset as evident from Fig. 4, where the distribution is shown with respect to location and experience traveling outside the country. Importantly, our overview of respondent demographics is presented for descriptive purposes and although the results are in agreement with other studies of online shoppers, we make no claims about our relatively small dataset adequately sampling the entire population of China.

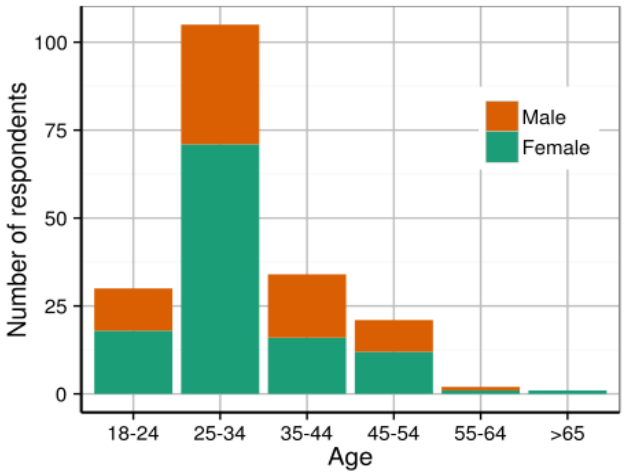

Fig. 2. Respondent age and gender distribution in the collected choice experiment questionnaire dataset.

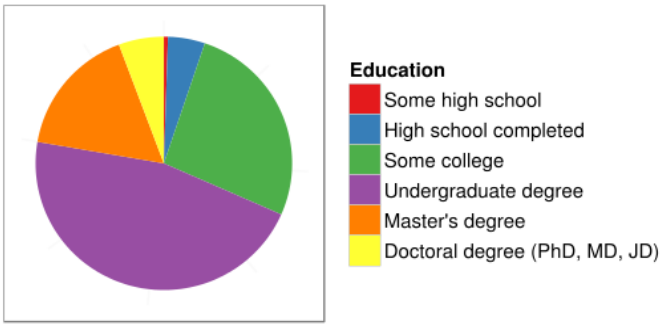

Fig. 3. Education level distribution in the collected choice experiment dataset.

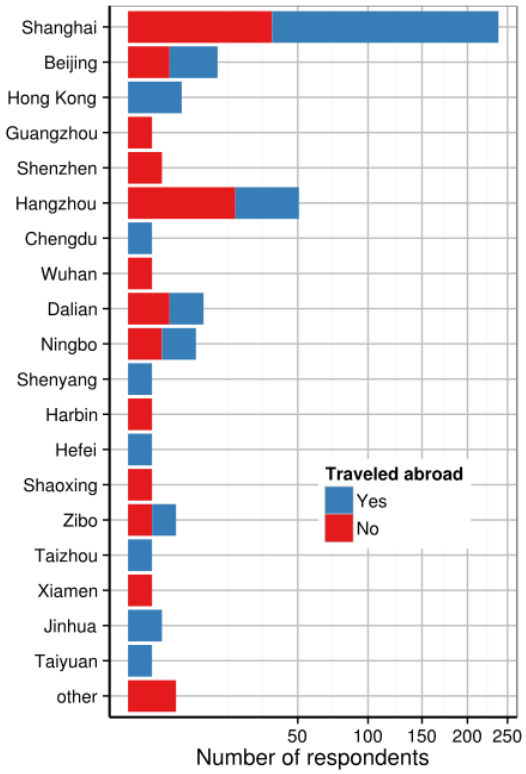

Fig. 4. Chinese cities represented in the collected questionnaire dataset with the corresponding number of respondents who had and had not traveled outside China. The cities are listed in the order of decreasing GDP.

A. Product selection counts and shopping frequency

The total number of times conventional, organic domestic, and organic imported varieties of apples and milk were chosen in our online experiment are compared in Fig. 5. While the fraction of respondents indicating their willingness to pay for organically certified (domestic or imported) milk was over 80%, it was much lower for organic apple varieties with more than half of the respondents selecting conventional apples. One possible reason behind this considerable disparity may be related to higher trust in organic certification when the product is presented in a packaged form. These results also highlight the importance of various product attributes for organic shopping decisions and suggest that purchase likelihood prediction models should be built separately for different product categories.

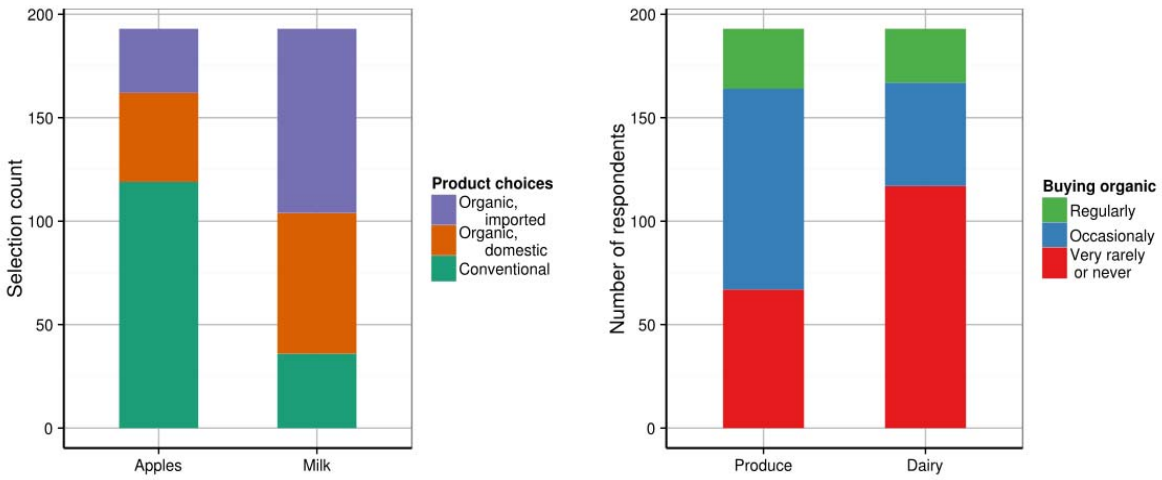

Fig. 5. Product selection counts for conventional and organic varieties of apples and milk compared to organic shopping frequencies in the produce and dairy categories, respectively. Note a significantly larger fraction of questionnaire respondents willing to buy organic milk than the corresponding fraction of periodic organic dairy shoppers.

We have investigated the relation between organic shopper habits and product selections within the scope of the collected dataset. While regular organic shoppers are typically more likely to pay extra for organic products [9], this observation does not necessarily represent the organic purchase behavior of Chinese consumers [1]. To gain further insights, we included consumer shopping frequency for organic product categories among the questions (the full set of questions is listed in the Appendix). Note however that, as already mentioned in the methodology section, the answers were not used among learning algorithm features. Instead, we used this information to compare purchase intent with the currently established shopper habits and explore potential opportunities for organic product sales.

We conclude the absence of any significant correlations between organic shopping frequency and product selections in the collected dataset, which agrees with the previously discussed apparent inconsistency in the organic purchasing habits of Chinese consumers [1]. In addition, a particularly interesting finding has emerged from the comparison of product choices and shopping frequencies within the same product categories presented in Fig. 5. Less than half of the respondents qualified themselves as regular or at least occasional organic dairy shoppers, while over 80% indicated their willingness to pay for organic milk. Moreover, over 75% of those consumers who were not even occasional organic dairy shoppers selected some variety of organic milk. In the case of apples, close to 25% of those who do not currently purchase any organic produce were willing to buy organic apples.

While the reported fractions could be somewhat inflated due to a hypothetical bias in consumer preferences (i.e. the fact that the hypothetical choices they indicated may not fully translate into real online purchases), the presented comparisons suggest that there may still be a considerable business-to-consumer e-commerce opportunity for selling organic food and in particular dairy to that fraction of consumers who are willing to purchase it but had not done so on a regular basis. Our findings in this respect are however yet to be validated by larger-scale consumer surveys and more detailed market analysis, which is one possible extension of the current application study.

B. Data-driven market segmentation

A combination of kernel-based principal component analysis (kPCA) and hierarchical clustering (HC) was used to group the respondents by consumer trait similarity and explore representative consumer profiles of organic shoppers. The results of HC applied to the collected questionnaire dataset are presented in Figs. 6 and 7. Four major clusters stand out on the dendrogram and are highlighted in Fig. 6. The same clusters are shown in Fig. 7, where feature-vector points of individual consumers are plotted in the coordinate system of the two largest kPCA principal components. We note that unsupervised learning with HC was based on basic consumer traits listed in Table 1 and did not involve any information about product choices nor previous shopping frequency for organic food categories.

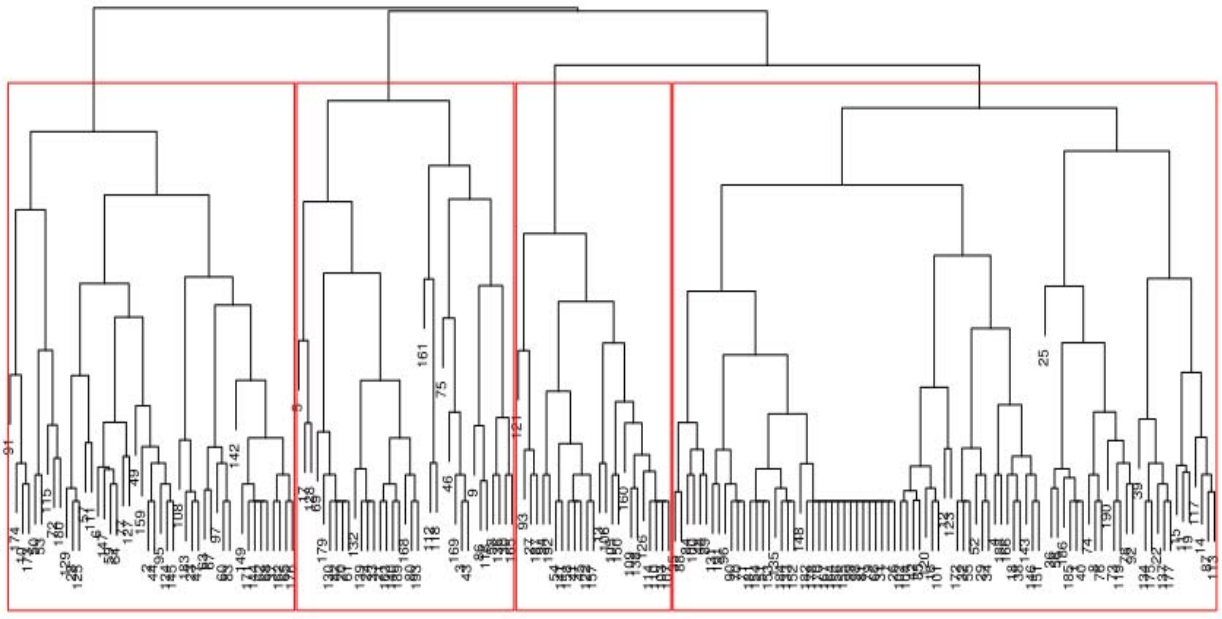

Fig. 6. HC dendrogram of questionnaire respondents with the major consumer clusters highlighted in red. HC was performed in the kPCA space as described in the methodology section.

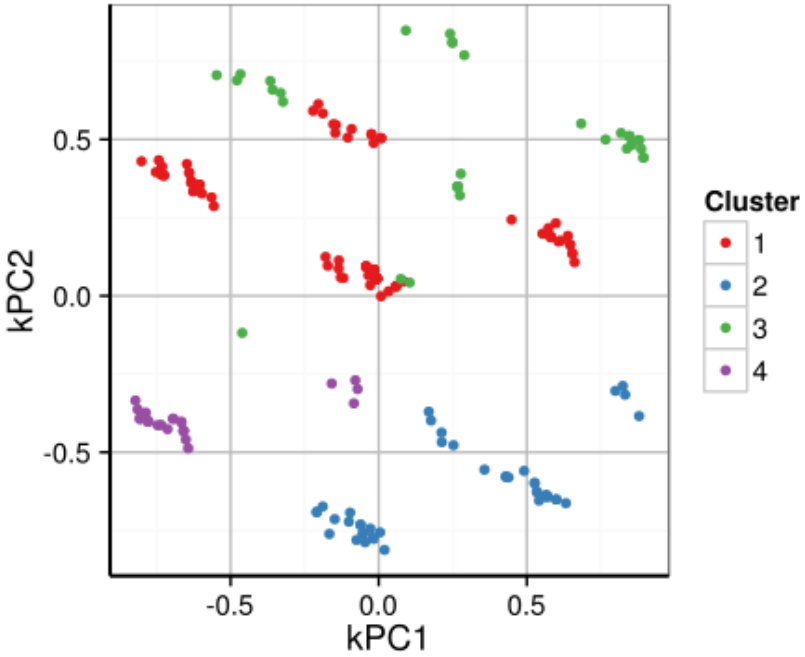

Fig. 7. Major consumer clusters (representing market segments) plotted in the space of the two largest kPCA principal components.

Further analysis of consumer traits performed for each of the identified clusters suggests the following consumer profiles:

(1) Younger female shoppers without major health problems;
(2) Male shoppers with somewhat higher education levels who had traveled abroad;
(3) Female consumers with health problems;
(4) Male consumers without major health problems who had not traveled abroad.

A classification system of Chinese consumers proposed in the previous organic market studies includes several categories with the major ones being "white collar professionals", "families with dietary concerns", "oversees returnees", "young people", and "families with young children" [4]. In terms of consumer clusters identified in our work, cluster (1) corresponds to young people tending towards a western lifestyle, cluster (2) most closely resembles white collar professionals as well as oversees returnees, while cluster (3) represents families with dietary concerns.

Our analysis of product choices suggests that the more prominent market segments for organic retail purchases are distributed throughout clusters (1), (2) and (3), while consumer profiles corresponding to cluster (4) describe respondents who are least likely to buy organic products and may therefore be excluded from cost-effective marketing efforts. From the shopping frequency perspective, more current organic shoppers are found among better educated professionals in cluster (2) and female consumers with dietary concerns in their families represented by cluster (3), which is in agreement with previous market studies [4], [6].

C. Purchase likelihood prediction

Profitability of a marketing campaign generally depends on the ability to predict purchase outcomes for the advertised product. Every piece of information about a prospective buyer may be useful in some way for a campaign manager to target their direct-to-consumer advertisements. We now proceed from descriptive analysis and unsupervised learning approaches to predictive modeling of organic product purchase decisions. It is important to reiterate that the specific objectives of our application study included the development of algorithms that extract some predictive value from very limited datasets and a small number of consumer traits, thus opening up opportunities for emerging small retailers to increase marketing campaign revenues.

We developed classification algorithms to identify consumers choosing higher-priced imported organic products as well as those who are unlikely to buy any organic products at all. In this way we are building complementary predictors useful for choosing a target set of consumers for a given marketing campaign as discussed in the methodology section.

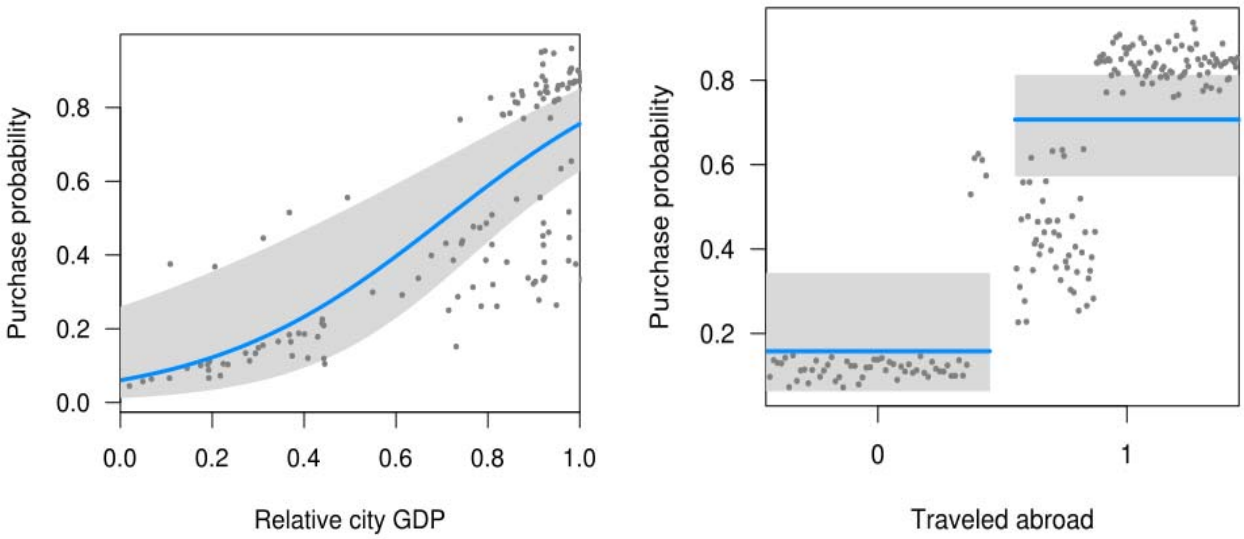

Fig. 8. Essential consumer characteristics for purchase likelihood prediction: city GDP (left), experience traveling outside of China (right). Visualization of the individual predictor effect on the organic product purchase probability is based on partial residuals in the logistic regression fit of the ROSE-balanced dataset. Shaded areas are confidence intervals for partial residuals.

Among the learning methods discussed in the methodology section, the best performing algorithms were based on the combined ROSE and SMOTE data balancing with logistic regression, as well as on cost-sensitive SVM. In particular, the performance of ROSE balancing followed by logistic regression in terms of the area under the receiver operating characteristic (ROC, see the methodology section for details) was 0.69 for classifying imported organic product buyers vs. the rest of the consumers, and SVM achieved a very similar ROC of 0.68. The highest ROC for predicting unlikely buyers was 0.61 in the case of SMOTE balancing followed by logistic regression. For comparison, a hypothetical perfect classification algorithm would have ROC of 1.0, while random class choices (no predictive value) result in the ROC value of 0.5.

Identifying consumer characteristics essential for the ability to predict purchase likelihood brings useful insights for organizing the data management campaign. To this end, we analyzed partial residuals [28] in the logistic regression fit of the ROSE-balanced dataset to evaluate individual predictors of the organic product purchase probability. The most relevant consumer characteristics were found to be city GDP and previous travel outside of China. The corresponding partial residual plots are presented in Fig. 8.

It should be noted that city average gross domestic product (GDP) was anticipated to play a significant role in product choices since residents of big and mid-sized cities in economically more developed regions of China such as the eastern coastal areas enjoy notably higher purchasing power as compared to other regions of the country [6], [29]. A much larger number of organic buyers among those respondents that had traveled outside China suggests travel-related marketing strategies such as travel website advertisements. The fact that the age and education level of Chinese consumers individually play less important roles in the organic food purchase intent within the scope of our dataset is consistent with previously reported findings [6]. Note however that while these consumer traits individually were not strong predictors of the purchase likelihood, they still play an essential role *in combination with other features* for achieving the reported performance values of consumer classification algorithms.

D. Marketing campaign revenue projections

In order to demonstrate potential benefits of the outlined data management framework, we now study how and under what conditions the predictive ability of machine learning classifiers discussed in the previous section translates into increased marketing campaign revenues. Let m be the cost of a marketing attempt per consumer and r be the revenue gained from a successful sale (which can also be thought of as a cumulative or some other adjusted revenue received over multiple sales to the same consumer that resulted from a successful marketing attempt). If N potential buyers within the reach of the marketing campaign are all included in the marketing effort, the resulting net revenue would be

$$G_0 = rN_s - mN, \tag{1}$$

where $N_s < N$ is the number of successful marketing attempts that resulted in a purchase.

Based on the target consumer selection discussed in the methodology section and depicted in Fig. 1, let M and M_s be the number of total and successful marketing attempts, respectively, when p^*-based selection takes place. The relative gain in the net campaign revenue resulting from the target consumer pre-selection can then be calculated as

$$g = \frac{G(M, M_s)}{G_0} = \frac{RM_s - M}{RN_s - N} \tag{2}$$

Here we have used a relative return per sale value $R = r/m$ describing the revenue from a successful sale relative to the cost of the corresponding individual marketing attempt. Using our collected dataset with specific product choices and a hypothetical marketing campaign where the consumer selection is guided by purchase likelihood predictions, we evaluated the set of quantities $\{N, N_s, M, M_s\}$ for different values of p^* and R by means of leave-one-out cross-validation. The resulting relative revenue gain g is presented in Fig. 9 as a function of R for organic milk and apple marketing campaigns.

Our results suggest the projected revenue gains reaching a factor of 10 or more for organic milk, and somewhat smaller but still significant gains in the case of apples. Note that these estimates are based on the completely realistic dataset of product choices collected in the current study. Moreover, as the values of R become smaller (beyond the scale in Fig. 9), a situation can take place when a campaign that looses money without consumer pre-selection (i.e. $G_0 < 0$) would still be profitable when purchase likelihood predictions are used to guide the marketing effort.

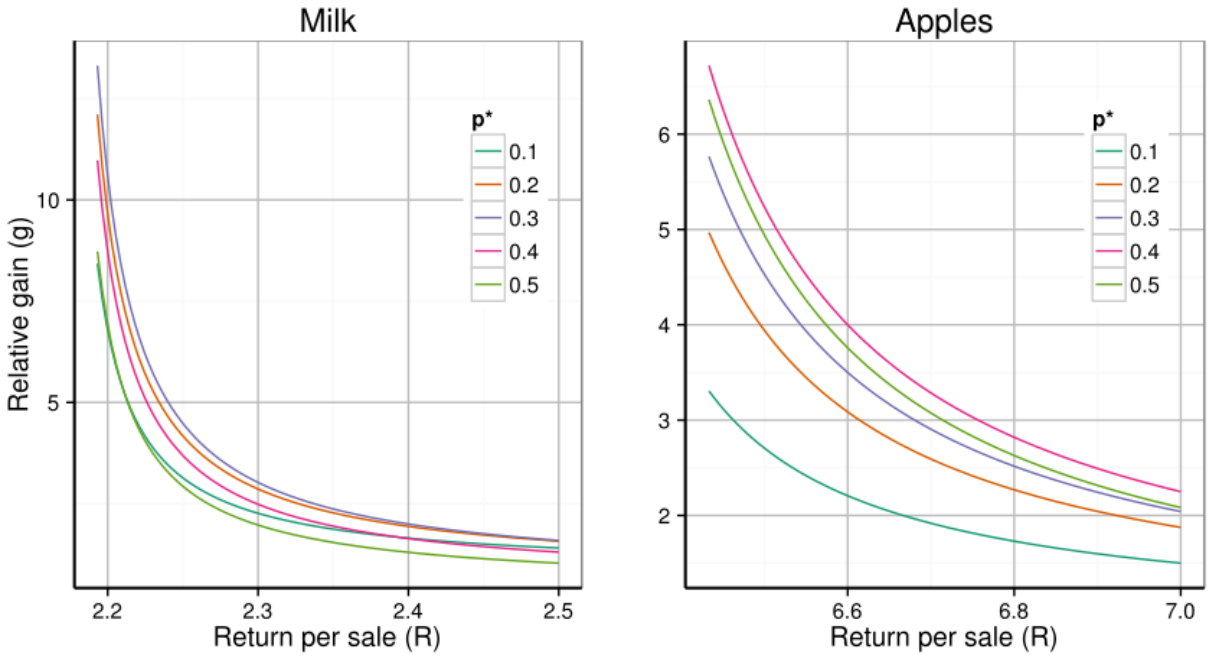

Fig. 9. Relative revenue gain (2) in organic milk and apple marketing campaigns resulting from the use of purchase likelihood prediction algorithms. The variation of g with respect to the return per sale R is presented for different values of the threshold probability parameter p^*. Minimal values of R are bound by limitations set on r and m values for a given campaign to be profitable without the use of predictive algorithms (thus allowing for comparison of the revenues with and without target consumer pre-selection). Note the projected revenue gains of the factor of 10 or more in the case of milk.

Revenue gains in general depend on the return per sale R which is specific to particular product and marketing channels, and the relative advantage of using selective marketing based on purchase likelihood predictions decreases with R. In the extreme case of a single-sale revenue r significantly exceeding the marketing cost m, it may be preferable to attempt marketing with any potential consumers within the reach of the given retailer. Such an extreme situation is however expected to be fairly uncommon in organic food retail, and the use of consumer data management for purchase likelihood predictions would be beneficial in the vast majority of organic food marketing cases.

IV. CONCLUSIONS

A growing number of Chinese consumers are willing to pay for safe and nutritious organic food products [1]. In the current application study we have developed machine learning algorithms for identifying those consumers and demonstrated the utility of those algorithms in purchase intent predictions using a dataset collected through a self-administered online product choice experiment questionnaire. Using a market segmentation strategy that relies on data clustering we also developed an algorithm for exploring representative consumer characteristics in various segments of the organic market.

Through the analysis of individual purchase likelihood predictor variables we have found that consumer location quantified in terms of the average city gross domestic product is one of the main factors in determining shopping preferences for organic food. In addition, we have found consumers who had traveled outside of China significantly more likely to purchase organic products, suggesting marketing strategies that utilize travel website advertisements and other relevant digital media.

Our results show that the net revenue in organic food marketing campaigns can considerably grow if consumer data are used for predictive algorithm training and every marketing attempt is based on the output of those algorithms. One possible extension of the current study may include incorporation of revenue optimization approaches to explore advanced strategies for consumer targeting, particularly in the online advertisement setting.

Another possible direction for future research is paved by our results obtained through the comparison of previous consumer shopping frequency and their current product choices. We have found a considerable fraction of shoppers who are willing to purchase organic products but had not done so on a regular basis yet, suggesting an e-commerce opportunity particularly with organic dairy products. More extensive consumer surveys are needed to validate the results on a larger scale and further explore potential online retail opportunities associated with our findings.

The methodology developed in the current study applies to organic brands and retailers in China as well as foreign brands seeking to increase their market shares. In addition, our approach can be readily adapted to other markets and product categories, enabling managers to obtain efficiencies utilizing online data collection technologies along with predictive machine learning algorithms. Importantly, we have shown that predictive value can be extracted even from a relatively small dataset with a limited number of available consumer characteristics, thus opening up data management strategies not only for "big data" owners but also for small, lean companies seeking to enter the expanding organic food market and establish their own online retail channels.

REFERENCES

[1] R. Liu, Z. Pieniak, and W. Verbeke, "Consumers' attitudes and behaviour towards safe food in China: A review," *Food Control*, vol. 33, no. 1, pp. 93–104, Sep. 2013.

[2] L. Qiang, L. Wen, W. Jing, and D. Yue, "Application of content analysis in food safety reports on the Internet in China," *Food Control*, vol. 22, no. 2, pp. 252–256, Feb. 2011.

[3] A. Veeck and A. C. Burns, "Changing tastes: the adoption of new food choices in post-reform China," *J. Bus. Res.*, vol. 58, no. 5, pp. 644–652, May 2005.

[4] International Trade Centre, "Organic Food Products in China: Market Overview," Geneva, Switzerland, 2011.

[5] B. Xie, L. Tingyou, and Q. Yi, "Organic certification and the market: organic exports from and imports to China," *Br. Food J.*, vol. 113, no. 10, pp. 1200–1216, 2011.

[6] S. Yin, L. Wu, L. Du, and M. Chen, "Consumers' purchase intention of organic food in China.," *J. Sci. Food Agric.*, vol. 90, no. 8, pp. 1361–7, Jun. 2010.

[7] A. Krystallis and G. Chryssohoidis, "Consumers' willingness to pay for organic food: Factors that affect it and variation per organic product type," *Br. Food J.*, vol. 107, no. 5, pp. 320–343, 2005.

[8] M. Janssen and U. Hamm, "Product labelling in the market for organic food: Consumer preferences and willingness-to-pay for different organic certification logos," *Food Qual. Prefer.*, vol. 25, no. 1, pp. 9–22, Jul. 2012.

[9] E. J. Van Loo, V. Caputo, R. M. Nayga, J.-F. Meullenet, and S. C. Ricke, "Consumers' willingness to pay for organic chicken breast: Evidence from choice experiment," *Food Qual. Prefer.*, vol. 22, no. 7, pp. 603–613, Oct. 2011.

[10] G. Guido, M. I. Prete, S. Miraglia, and I. De Mare, "Targeting direct marketing campaigns by neural networks," *J. Mark. Manag.*, vol. 27, no. 9–10, pp. 992–1006, Aug. 2011.

[11] T. Kaya, E. Aktas, I. Topçu, and B. Ülengin, "Modeling Toothpaste Brand Choice: An Empirical Comparison of Artificial Neural Networks and Multinomial Probit Model," *Int. J. Comput. Intell. Syst.*, vol. 3, no. 5, pp. 674–687, Oct. 2010.

[12] F. Wang, P. Zhang, Y. Shang, and Y. Shi, "The Application of Multiple Criteria Linear Programming in Advertisement Clicking Events Prediction," *Procedia Comput. Sci.*, vol. 18, pp. 1720–1729, Jan. 2013.

[13] J. Ma, Ed., *China Statistical Yearbook 2012*. China Statistics Press, 2012.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[15] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.

[16] M. Kuhn and K. Johnson, "Remedies for Severe Class Imbalance," in *Applied Predictive Modeling*, New York, NY: Springer New York, 2013, pp. 419–443.

[17] P. Kang, S. Cho, and D. L. MacLachlan, "Improved response modeling based on clustering, under-sampling, and ensemble," *Expert Syst. Appl.*, vol. 39, no. 8, pp. 6738–6753, Jun. 2012.

[18] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[19] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data Min. Knowl. Discov.*, Oct. 2012.

[20] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

[21] M. Kuhn, "Building Predictive Models in R Using the caret Package," *J. Stat. Softw.*, vol. 28, no. 5, pp. 1–26, 2008.

[22] R Core Team, "A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria, 2013.

[23] B. Hanczar, J. Hua, C. Sima, J. Weinstein, M. Bittner, and E. R. Dougherty, "Small-sample precision of ROC-related estimates.," *Bioinformatics*, vol. 26, no. 6, pp. 822–30, Mar. 2010.

[24] B. J. Parker, S. Günter, and J. Bedo, "Stratification bias in low signal microarray studies.," *BMC Bioinformatics*, vol. 8, p. 326, Jan. 2007.

[25] A. Airola, T. Pahikkala, W. Waegeman, B. De Baets, and T. Salakoski, "An experimental comparison of cross-validation techniques for estimating the area under the ROC curve," *Comput. Stat. Data Anal.*, vol. 55, no. 4, pp. 1828–1844, 2011.

[26] M. D. Clemes, C. Gan, and J. Zhang, "An empirical analysis of online shopping adoption in Beijing, China," *J. Retail. Consum. Serv.*, p. in press, Sep. 2013.

[27] A. Y.-L. Chong, "Mobile commerce usage activities: The roles of demographic and motivation variables," *Technol. Forecast. Soc. Change*, vol. 80, no. 7, pp. 1350–1359, Sep. 2013.

[28] M. Kutner, C. Nachtsheim, J. Neter, and W. Li, *Applied Linear Regression Models*. McGraw-Hill, 2004.

[29] G. Guo, "Consumer Behavious in China," in *Marketing Management in Asia*, S. Paliwoda, T. Andrews, and J. Chen, Eds. Routledge, 2013.

APPENDIX: CHOICE EXPERIMENT QUESTIONNAIRE

Questions (English translation)

1. Have you purchased any organic food product during the last year? (Yes/No; No-answer skips the following two questions)
2. How often do you purchase organic fruit or vegetables? (Regularly/Occasionally/Very rarely or never)
3. How often do you purchase organic milk products? (Regularly/Occasionally/Very rarely or never)
4. Please select your age group (18-24, 25-34, 35-44, 45-54, 55-64, 65 and older)
5. Your gender (Male/Female)
6. What best describes your level of education? (Some high school, High school completed, Some college, Undergraduate degree, Master's degree, Doctoral degree (PhD, MD, JD))
7. Where do you live? (City name)
8. Have you ever traveled outside China? (Yes/No)
9. Do you have children less than 18 years old living with you? (Yes/No)
10. Does anybody in your household have health problems that sometimes restrict the range of your food choices? (Yes/No)

Product choice experiments (English translation)

We want you to make product selections in the same way that you would if you really had to pay for the product and take it home or order online. Please take into account how much you would really want the product, as opposed to other alternatives. Imagine you are facing the choices below as if you were really in a grocery store.

Which milk product would you buy?

- Conventional, domestic, Meng Niu brand, 1.92 RMB for 250 mL
- Organic, domestic, Meng Niu brand, 5.68 RMB for 250 mL
- Organic, imported (US), Organic Valley brand, 8.92 RMB for 250 mL

What type of apples would you choose?

- Conventional, domestic, no brand labels, 14.95 RMB for 1 kg
- Organic, domestic, Chinese organic certification label, 23.60 RMB for 1 kg
- Organic, imported (New Zealand), NZ QUEEN brand, 34.87 RMB for 1 kg