

Contribution Statement

Although the Chinese ready-to-eat (RTE) food market has really taken off in the last decades, increasing its appeal towards researchers and practitioners, no study offers an attitudinal and behavioral analysis of the topic. Thus, some important features among consumers have not been recognized, as well as, the possible methodological frameworks to investigate it. The present work is designed to be the first to profiling the typical consumer of RTE salads, studying a never explored market, and provide foresight to investors and policymakers. Second, we test competing models in order to reduce prediction errors; bringing new perspectives and findings to the literature of the fresh-cut; contributing to the nascent literature on using ML methods to predict customers' behaviors.

.

Chinese consumers' attitude towards ready to eat salads by comparing traditional Logit with Machine Learning methods

October 6, 2021

Keywords: Ready to Eat Salads, Fresh-cut fruits & vegetables, Chinese consumers, Machine Learning

Abstract

In the last decade, the ready to eat (RTE) food market has experienced substantial growth in China. The consumption of fresh-cut fruits and vegetables is a common eating trend for the expanding urban middle-class that is now more healthy-oriented than in the past. The attributes of RTE meals make it an attractive food choice for many consumers. However, to the authors' knowledge, no study offers an attitudinal and behavioral analysis of the topic. Therefore, a survey among Chinese respondents was conducted to understand better the factors associated with RTE salad consumption. The goal of this paper is thus twofold. First, we aim at profiling the typical consumer of RTE salads. Second, we test different machine learning classification algorithms on primary consumer data. Results show that consumption is more common among young respondents, females, healthy-oriented individuals at the beginning of their career. The

role of the subjective norm is also positively associated with increased consumption of RTE salads.

Introduction

In recent years, the demand for ready-to-eat (hereafter RTE) food has increased noticeably in several world areas. According to current data¹, the growth of the fresh food packaging market is expected to go from USD 79.8 billion in 2020 to USD 95.2 billion in 2025, with a CAGR of 3.5%. The Asia-Pacific Countries, including China, are leading this growth due to the rising demand for fresh foods, particularly vegetables and meat products, from the emerging middle-class. RTE products include fresh-cut fruits and vegetables and other ready-to-eat meals like cereals, dairy, cold cuts, and meat. The nutritional benefits of RTE foods have been extensively acknowledged by experts Santeramo et al. (2018); Singla et al. (2020), although some others warn about the health risks associated with high sugar and fat content in some RTE foods Poti et al. (2016); Thike et al. (2020). Eating fruits and vegetables is associated with a robust immune system and risk reduction of many cardiovascular diseases and cancer Gibson et al. (2012); Gu et al. (2021). RTE fruits and vegetables are a practical way to increase their consumption as recommended by the health authorities of many countries. RTE meals can also provide a balanced diet in nutrient intake without giving up on taste and quality. This is also why they are often used in school menus Gibson&Partridge (2019). RTE salads are the most popular food among fresh food products, consisting of pre-cleaned leafy and fresh-cut mixed vegetables immediately ready for consumption or with few days of shelf-life. The consumption of pre-packaged salads is a well-established eating trend in many countries, and their popularity is exceptionally high among Western consumers. The success of RTE fresh meals in Western countries has

¹Fresh Food Packaging Market by Material (PE, PP, Paper, Aluminum, BOPET, PVC), Pack Type (Converted Roll Stock, Gusseted Bags, Flexible Paper, Corrugated Box), Application (Meat Products, Vegetables, Seafood, Fruits), and Region - Global Forecast to 2025, Markets and Markets, available at: <https://www.marketsandmarkets.com/Market-Reports/fresh-food-packaging-market-240678791.html>

been interpreted as a sign of how innovation and integration between traditional agriculture and the food industry can meet the needs of modern consumers (Brennan et al. (2013); Massaglia et al. (2019)). However, in other realities like China, the fresh food market is still expanding, and this circumstance offers an interesting investigation opportunity. To the authors' knowledge, no one ever proposed an empirical study based on Chinese consumers' data, despite the preeminent growing expectations on the specific market. To cover this gap, the present study has a twofold goal.

Obj.1. By using survey data, the first objective of this study is to understanding socio-demographic, socio-economic, and psycho-social factors that affect the consumption of ready-to-eat salads among urban consumers.

Consumption patterns in urban China have dramatically changed in the last few years. Most consumers, especially young and urban, pay attention to their health, diet, and physical appearance much more than in the past, and this is reflected in their dietary choices. At the same time, most of them conduct a busy and fast-paced life, inspired by the need to enjoy life pleasures and save time and money. Moreover, as market research data show Daxue Consulting (2021), health and future uncertainty have become more prominent preoccupations after the Covid-19 crisis. RTE fruits and vegetables are a consumption trend suitable to respond to these emerging consumer needs because do not require any time for preparation, are largely available, and offer a healthy meal substitution. However, more research is needed to understand the driving forces that favor its consumption. In addition, the intake of vegetables in China is still below the level recommended by the Chinese dietary guidelines Xiao et al. (2015) and the attributes of RTE salads, such as convenience and practicality, might help reach the expected consumption standard in the future.

Obj.2: The second objective of the paper is to apply an innovative methodological approach to consumer research. We propose a comparison among traditional Logit, random forests (RF), Deep Neural Network (DNN), Regression Tree (RT), and Support Vector Machine (SVM) in a food choice determinant analysis.

Consumers' behavior studies have most of the time referred to econometrics in order to predict choices (Storm et al. (2020)). However, limited data associated with a high number of variables/factors have always been an issue to have easy interpretation and good prediction ability of models for the inference process. In recent years, however, different modelling approaches have been brought to light, also because of the remarkable increase in computational power Hastie, T., Tibshirani, R. and Friedman, J. (2009). Machine learning (hereafter ML) approaches, in fact, could present advantages over traditional econometric models in certain situations. It is enormously useful in our context where a large number of factors may affect the fresh-cut customer's choice. Traditional techniques, such as the logistic regression, might be criticized because of rigid structure when compared to "distribution-free" approaches. However, they have the advantage of furnishing interpretability of coefficients, while the other approaches could be convenient for their ability to predict (Lecun et al. (2015)). To the best of our knowledge, there is a limited number of studies that use ML models for consumer profiling in relation to food choices. Summing up, the paper contributes to literature by offering manifold elements of innovation:

(i) studying a never explored market and provide foresights to investors and policy makers; (ii) testing competing models in order to reduce prediction errors; (iii) bringing new perspectives and findings to the literature of the fresh-cut; (iv) contributing to the nascent literature on using ML methods to predict customers' behaviors. The paper is structured as follows. Section 2 presents the research background. It discusses the literature review on RTE salads consumption and gives the policy and legal framework. Section 3 discusses the data and methods. Section 4 presents the results and Section 5 discusses the findings and Section 6 concludes.

Background

Literature review

Literature presents plenty of evidence about food consumers' preferences towards fresh-cuts, even though mostly analyzing Western countries. In general, Europeans and US consumers are oriented towards quality, safety, nutritional elements, freshness, richness of assortment, etc. (Santeramo et al. (2018)). In addition, other findings state that medium-high income consumer target, generally also high educated people (middle-class), is interested in purchasing products with a higher level of service and quality (Artes et al. (2019); Watada et al. (2019), Soliva Fortuny et al. (2000), Rico et al. (2007)). However, common outcomes regard income as a driver of purchase (Marshall et al. (1994); Cassady et al. (2007)) as well as time pressure, social environment and eating-out habits (Frewer et al. (2013); Buckley et al. (2007)). Other papers highlight that quality, safety and labeled information are also important (Pilone et al. (2017); Santeramo et al. (2018); Pollard et al. (2002); Baseline et al. (2017)). All these contributions, however, present empirical evidence of Western countries. Among the countries of South East Asia that are most active in the development of fresh-cut products consumption, there is undoubtedly Thailand, whose model has been analyzed by FAO in its research (James et al. (2010), FAO 2010) Regarding fresh food products, the great majority of studies target Western consumers. All are empirical investigations aimed at segmenting the market and individualizing the perceived benefits and barriers. The literature indicates that consumption is prevalent among females, middle-income earners, and individuals below forty years old (Massaglia et al. (2019); Sgroi et al. (2018); Van Loo et al. (2010). In terms of time, money, and practicality, convenience is also an essential purchasing factor (Stiletto et al. (2020); Vidal et al. (2013). Other aspects that play a crucial role in the purchasing decisions are freshness (Dinnella et al. (2014)), the understanding of food processing methods (Hicks et al. (2009); Liem et al. (2016)), and the use of experience, visual intrinsic, and extrinsic

attributes (Chonpracha et al. (2020); Pilone et al. (2017)). Furthermore, food safety and shelf-life information are critical factors affecting RTE consumption (Baselice et al. (2017)). Few studies target Asian consumers, and most of them do not focus on RTE fruits and vegetables precisely but also consider other types of food, like cereal-based meals. A common finding among these studies is that the preferences and attitudes towards RTE food are strongly affected by experience and intrinsic attributes like freshness, odor, and taste (Kwon & Ju (2018); Thienhirun & Chung (2018)). In Japan and South Korea, the market of fresh-cut products has been developed since the late 1990s. Fresh-cut meals are primarily used in schools and restaurants and also for personal consumption at home. Consumers in these regions perceived RTE products as safe and of high quality (Kim (2007)). Another study on Korean consumers found that consumption was higher among females, young, and healthy-oriented professionals. The only research targeting Chinese consumers investigated the demand for RTE imported meals. Results found that higher consumption of instant RTE foods is more frequent among commuters, and extrinsic attributes on food packages are often used as a cue to determine the product's quality (Lilavanichakul et al. (2018)). Among these attributes, appearance is a fundamental purchasing driver when suggests freshness. This attribute is essential for Chinese consumers to determine food quality, especially the elderly (Hansstein et al. (2017)).

Policy and consumption framework

Between 1960 and 2017, FAO statistics about food balance highlight that, while Western countries show decreasing consumption of fruits and vegetables, some Asian countries, including China, show the opposite. In fact, population growth, increasing trade, nutrition awareness, and migration are all factors having an incidence on those trends (F. A. O. (2017)). In the last decades, the consumption of more nutritious foods increased worldwide, however, different trends have been observed. For instance, fruit consumption rose in all areas, but vegetable consumption increased with different pace

especially in Asian countries, as shown in Figure 1 (FAO, 2017). When correlating consumption and income, per capita fruit intake tends to rise with income, whereas vegetable consumption declines (F. A. O. (2017)).

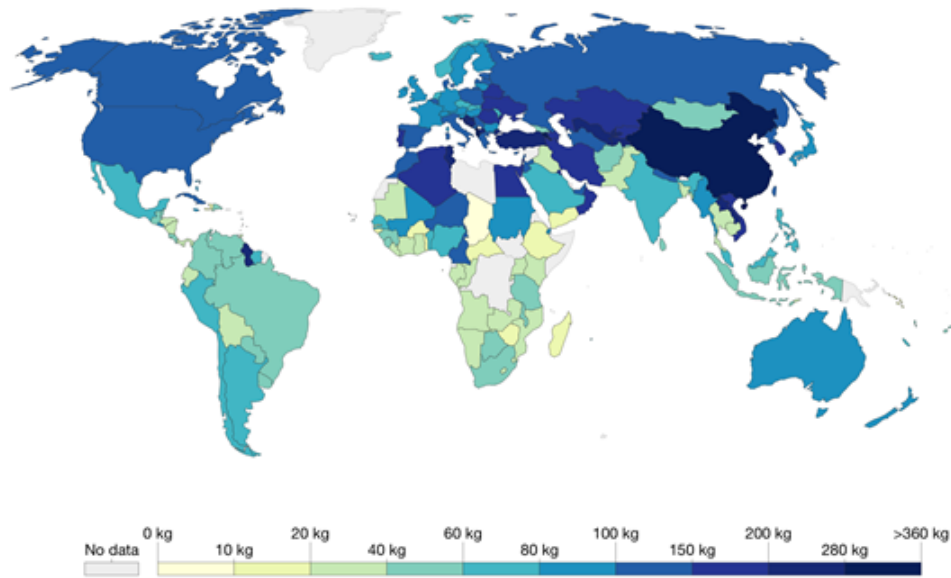


Figure 1: Map of evolving per capita consumption of vegetables (F. A. O. (2017)).

Given the great concern about contemporary dietary habits, governments in several countries have launched informational and educational initiatives aimed at increasing public awareness about the benefits of F&V. As a consequence, the effectiveness of these interventions have been evaluated by several authors (Seiders & Petty (2004); Gordon et al. (2006); Mazzocchi et al. (2009)). These studies lead to the development of guidelines and recommendations. Food policy and regulatory guidelines are prime tools to achieve a healthy diet in a population, with a remarkable impact on the health care systems and social and public health policies. China has also a long-standing commitment to policy guidance on food security-related nutrition. In the '90s it launched the "China Nutrition Improvement Action Plan (1996-2000) Healthy China Action Plan (2019-2030)", followed by the "Management Measures for Nutrition Improvement (2010)" and the "National Nutrition Plan (2017-2030)", which aims to improve nutrition and health on a national level and set national nutrition strategies for 2020 and

2030. Recently, China made great strides in food security and nutrition, implementing the "Healthy China Action Plan (2019–2030)" (HCAP). This regulation centralizes efforts to improve the populations' nutrition status, facing many challenges, including obesity and chronic diseases.

Indeed, the fresh-cut sector - that includes RTE salads - plays a critical role in the daily consumption of FV, as recommended by the World Health Organization (World Health Organization. (2008)). The “fresh-cut” sector offers convenience, in terms of time saving for washing and preparation, and freshness.. In line with the definition given by the International Fresh-Cut Produce Association (IFPA), fresh-cut FV are minimally processed products that are washed, cut, mixed, and packed. Since their first appearance in Europe in the early 1980s, they have become more and more popular in consumers' market baskets because of time convenience, quality and safety attributes that are generally valued positively by consumers (Artes et al. (2019)). RTE products are defined by the European Commission as "food intended by the producer or the manufacturer for direct human consumption without the need for cooking or other processing to eliminate or reduce to acceptable level micro-organisms of concerns" (Regulation 2073/2005/EC). In China, RTE products regulation is included in the Food Safety Law of 2015 (HFG - Law and Intellectual Property, 2016). Given the risk of raw food contamination with dangerous bacteria, RTE food safety is currently the main concern for authorities. Therefore, encouraging proper handling and storage to avoid foodborne diseases is a priority for public and private food institutions.

Models

Logit

To explore the role of specific factors affecting the consumption of ready-to-eat salads in China, we firstly run a logit model. Logit models are primarily used in discrete choice analysis and are helpful to assess the odds ratio of an outcome exposure or food

choice. In consumer behavior and marketing research, logit models have been widely used to understand purchasing behavior (see for example Nevo (2011); Roberts & Lilien (1993)).

Our independent variable can be treated as a realization of a random variable that follows a binomial distribution $Y_i \sim B(n_i, \pi_i)$. Thus, we define generalized linear model with binomial response and link logit as:

$$\text{logit}(\pi_i) = \mathbf{x}_i' \boldsymbol{\beta} \quad (1)$$

where the logit of the underlying probability π_i is a linear function of the predictors where \mathbf{x}_i is a vector of covariates and $\boldsymbol{\beta}$ is a vector of regression coefficients. Thereby, β_j represents the change in the logit of the probability associated with a unit change in the j -th predictor holding all other predictors constant. Exponentiating Equation 1 we find that the odds for the i -th unit are given by

$$\frac{\pi_i}{1 - \pi_i} = \exp \{ \mathbf{x}_i' \boldsymbol{\beta} \} \quad (2)$$

The related log-likelihood function is defined as:

$$\log L(\boldsymbol{\beta}) = \sum \{ y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i) \}, \quad (3)$$

where π_i depends on the covariates \mathbf{x}_i and a vector of p parameters $\boldsymbol{\beta}$ through the logit transformation of Equation 1 .

In our specification, the dependent variable indicated if the respondent regularly (at least three times per week) eats fresh-cut salads. The explanatory variables are age, income, seniority, household size, knowledge, fitness, source of information regarding RTE products, purchase location, and whether consumption occurs for snacking or for lunch. The logit model was conducted on 70% of the original sample as required by predictive statistics.

Machine Learning

In this section, we deal with the wide framework of unsupervised learning which embraces various models such as Support Vector Machine, Neural Networks and Tree-based Models, used for both regression and classification problems.

According to the statistical learning theory (SLT) the problem of supervised learning is formulated as follows. Given a set of training data $D = \{(\mathbf{x}_1, y_1) \dots (\mathbf{x}_l, y_l)\}$ in $\mathbb{R}^n \times \mathbb{R}$ sampled according to unknown probability distribution $P(\mathbf{x}, y)$, and a loss function $L(y, f(\mathbf{x}))$ that measures the error done when, for a given \mathbf{x} , $f(\mathbf{x})$ is "predicted" instead of the actual value y .

It is worth pointing out that there is no information on the underlying joint probability functions. Therefore, it is necessary to perform a "distribution-free" approach, where the only information available is a training dataset.

The problem consists in finding a function f that minimizes the expectation of the error on new data, that is, find a function f that minimizes the expected error:

$$\int L(y, f(\mathbf{x}))P(\mathbf{x}, y)d\mathbf{x}dy$$

Since $P(\mathbf{x}, y)$ is unknown, we need to use some induction principle in order to infer from the l available training examples a function that minimizes the expected error. The principle used is Empirical Risk Minimization (ERM) over a set of possible functions, called hypothesis space. Formally this can be written as minimizing the empirical error:

$$\frac{1}{l} \sum_{i=1}^l L(y_i, f(\mathbf{x}_i))$$

Machine Learning tools are the so-called "nonparametric" models. "Nonparametric" does not mean that the ML models do not have parameters at all. On the contrary, their "learning" is the crucial issue here, indeed, unlike in classic statistical inference, the parameters are not predefined and their number depends on the training data used.

DNN

The term NN refers to a mathematical model inspired by the human brain. It allows computational models composed of multiple processing layers to learn representations of data characterized by multiple levels of abstraction, solving even the most complex problems. Its architecture includes neurons, synaptic connections that link the neurons, and learning algorithms (backpropagation). NN works as a weighted regression in which each unit gets “weighted” information through synaptic links from the other connected ones, returning an output by using an activation function that transforms the weighted sum of input signals. Albeit DL embodies a broad variety of networks differing from each other by the architectural structure, in this section the basic scheme of NN has been presented, the so-called feedforward structure. NN training involves an unconstrained optimization problem where the aim is to minimize a function in high dimensional space, the so-called loss function, that measures the difference between the predicted values and observed ones. The back-propagation is the most used algorithm for the training of NNs. The algorithm compares the predicted values against the desired ones (objective) and modifies the synaptic weights by back-propagating the gradient of the loss function. Schematically, the procedure alternates forward and backward propagation steps:

- in the forward step, the prediction is computed fixing the synaptic weights,
- in the backward step, the weights are adjusted in order to reduce the error of the network.

The NN iteratively performs forward and backward propagation and modifies the weights to find the combination that minimizes the loss function \mathcal{L} .

SVM

Support Vector Machines originated as an alternative to Neural Networks in the wide context of statistical learning, offering a solution to this trade-off by arbitrarily fixing

model accuracy and minimising model complexity; in this way it is able to manage overfitting.

In a classification problem, this translates into an optimal solution identified by a linear hyperplane which correctly classifies observed data and lies as far as possible from them.

Random Forest

Since Regression Trees usually produces low bias and high variance estimations, they became a good candidate for ensemble methods. Indeed, Random forests basically consist in building an ensemble of decision trees grown from a randomized variant of the tree, this method is useful to get the error reduction pulling down the prediction variance, preserving the bias Starting from a single learning set, the basic idea is to introduce a random perturbation into the learning procedure in order to introduce a differentiation among the trees and combine the predictions of all these trees using aggregation techniques. Breiman (1996) proposed a first aggregation method the so-called bagging in which the different trees, are built by using random bootstrap copies of the original data. Its natural evolution, the random forests, was developed by the same author in 2001 Breiman (2001). In the random forests the bagging approach has been extended and combined with randomization of the input variables that are used when considering candidate variables to split internal nodes t . In particular, instead of looking for the best split s^* among all variables, the algorithm chooses a random subset of K variables for each node and then determines the best split using these variables.

Performance assessment Cross-validation

In a k -fold cross-validation (CV), the original dataset is randomly partitioned into k subsets of approximately equal sizes. At each of the k CV iterations, one of the folds is chosen as the test set, while the $k - 1$ others are used for training. The considered performance measure is computed based on the test set. After the k iterations, the

performances are finally averaged over the iterations. In our study, we perform 10 repetitions of stratified 5-fold CV, as commonly recommended (Bischl et al. (2012)). In the stratified version of the CV, the folds are chosen such that the class frequencies are approximately the same in all folds. The stratified version is chosen mainly to avoid problems with strongly imbalanced datasets occurring when all observations of a rare class are included in the same fold. By "10 repetitions", we mean that the whole CV procedure is repeated for 10 random partitions into k folds with the aim to provide more stable estimates.

Accuracy Prediction

For binary target variables, we evaluate the level of accuracy and its 95% confidence interval (CI), true positive rate (Sensitivity), true negative rate (Specificity), and Cohen's Kappa (Cohen (1960)). We define $y = 1$ for a regular consumption of RTE and 0 otherwise. Then a 2×2 confusion matrix has elements $a_{\text{row}, \text{column}}$ with predicted conditions $\hat{y} = \{1, 0\}$ on rows and true conditions $y = \{1, 0\}$ on columns. The statistics are defined by,

$$\text{Accuracy} = \frac{a_{11} + a_{22}}{a_{11} + a_{12} + a_{21} + a_{22}} \quad (4)$$

$$\text{Sensitivity or True Positive Rate (TPR)} = \frac{a_{11}}{a_{11} + a_{12}} \quad (5)$$

$$\text{Specificity or True Negative Rate (TNR)} = \frac{a_{22}}{a_{21} + a_{22}}$$

We calculate accuracy that refers to the portion of customers correctly classified with respect to RTE regular consumption. We calculate the 95% confidence interval using accuracy's standard deviation generated through iterations. Sensitivity refers to the

proportion of regular RTE consumers correctly identified as such. Poor sensitivity implies a large number of inclusion errors, i.e. identifying RTE consumers when in fact they are not. Specificity refers to the proportion of customers correctly predicted to be occasional RTE consumers. Poor specificity implies a large number of exclusion errors. Cohen's Kappa statistic measures the agreement for categorical variable relative to what would be expected by chance. That is,

$$\text{Kappa} = 1 - \frac{1 - \text{Accuracy}}{1 - EP} \quad (6)$$

$$EP = \frac{(a_{11}+a_{12})(a_{11}+a_{21})+(a_{21}+a_{22})(a_{12}+a_{22})}{(a_{11}+a_{12}+a_{21}+a_{22})^2}$$

where, accuracy of prediction is the observed probability, and EP is the probability of random agreement or expected probability. Cohen's Kappa becomes zero if there is no agreement among the predicted and observed response other than what would be expected by chance. The reported output also provides the McNemar test and the numerical and graphical representation of the ROC curve and the relative area under the curve (AUC) values ².

Empirical Analysis

Data

The cross-sectional study examined the consumption of ready-to-eat salads among a sample of Chinese consumers. The consumption of prepackaged fruits and vegetables has increased in China in the last years, especially among young urban workers. Fresh salads are a versatile, healthy, and convenient meal that can easily substitute or accompany a lunch, a snack, or a dinner. A total of 524 questionnaires were collected by trained interviewers between January and March 2019, in the area of Suzhou and Shanghai. The study employed a convenient sample, and participants took part in the

²The area under the (ROC) curve, summarizes the classifier performance. The larger area under the curve the better the classifier.

survey voluntarily. The questionnaire was part of a larger project that collected information on ready-to-eat fruits and vegetables. It was designed following the literature on the consumption of fruits and vegetables and included several sections. Questions included in the analysis are available in the Appendix (7.1). The project's coordinators built the first version in English, and the questionnaire was later translated into Chinese by a bilingual speaker. Researchers conducted a questionnaire validation through a pilot test on a smaller sample of potential respondents to make sure that Chinese speakers correctly understood all the questions. The core analysis included behavioral questions on consumption of ready-to-eat-salads, socio-demographic variables, knowledge of fresh-cut products, subjective norms, and other questions on respondents' habits. All variables included in the model were categorical. Data were analyzed by the authors and are now currently stored on a private cloud owned by the authors.

Variable transformation Most variables did not require any modification, with the exception of two indicators that were built after data collection. The first indicator is "knowledge," a variable measuring respondent's level of awareness towards packaged ready-to-eat fruits and vegetables compared with their regular alternative. Respondents had to indicate whether six statements were true or false. They were: "RTE products must be washed before being consumed," "RTE products are more treated," "RTE products are healthier," "RTE products are industrial products," and "RTE products are more controlled and selected." If the respondent answered correctly, the variable value was 1, and 0 otherwise. We then built the variable "knowledge" so that each respondent could obtain a score ranging from 0 to 6. Knowledge was categorized as low if the score was below 2, medium if between 2 and 4, and high if above or equal to 5. The second variable that needed recoding was "subjective norm." Respondents were asked to indicate how important was the opinion towards fresh-cut products of their relevant people. Specifically, we included questions about family, friends, colleagues, and social media influencers. Following the statistical procedure indicated in Acock (2008), we built the

indicator "subjective norm" using the mean score method. Principal component analysis and alpha Cronbach were determined before calculating the scores. The subjective norm's eigenvalue was unique with a value of 2.18, and the Alpha Cronbach's between the four items was 0.71. The eigenvalue of knowledge was 1.95, and the correlation between items 0.60.

Table 1 reports the descriptive statistics of the whole sample. There is a slight prevalence of female respondents, and, overall, 57% of participants have less than 30 years old. The presence of young people in the sample is also reflected by the distribution of job seniority level, with only 15% of respondents occupying managerial roles. Income distribution is even among the four ranges of the variable. One-fourth of respondents regularly consume RTE salads. As specified in the question, regular consumption indicated that respondents purchase an RTE salad about three times per week. Consumption is mainly for snacking (60%) or for lunch (60%). Purchasing locations are the supermarket (75%) and convenience stores (25%). About half of respondents regularly practice some physical activity during their free time (48%). About 10% of respondents exhibit a low level of subjective norm, meaning they give little importance to others' opinion in their behavioral choices. About 52% and 37% have instead medium and high subjective norm's levels.

| Variable | Frequency | Variable | Frequency |
|---------------------------------|-----------|-------------------------------------|-----------|
| Gender | | Regular consumption of RTE | |
| <i>Female</i> | 52% | salads | |
| <i>Male</i> | 48% | <i>Yes</i> | 25% |
| Age | | <i>No</i> | 75% |
| <i>21-25</i> | 23% | Fitness | |
| <i>26-30</i> | 34% | <i>Yes</i> | 48% |
| <i>31-35</i> | 12% | <i>No</i> | 52% |
| <i>36-40</i> | 17% | Learning-advertising | |
| <i>>40</i> | 14% | <i>Yes</i> | 76% |
| Income | | <i>No</i> | 24% |
| <i>Below 15,000rmb</i> | 28% | Learning-social media | |
| <i>Between 15 and 20,000rmb</i> | 25% | <i>Yes</i> | 31% |
| <i>Between 20 and 25,000rmb</i> | 21% | <i>No</i> | 69% |
| <i>More than 25,000rmb</i> | 27% | Purchasing-supermarket | |
| Job seniority level | | <i>Yes</i> | 75% |
| <i>Entry</i> | 40% | <i>No</i> | 25% |
| <i>Middle</i> | 45% | Purchasing-convenience store | |
| <i>Managerial</i> | 15% | <i>Yes</i> | 49% |
| Household size | | <i>No</i> | 51% |
| <i>One or two people</i> | 18% | Consumption for snacking | |
| <i>Three people</i> | 50% | <i>Yes</i> | 60% |
| <i>More than three people</i> | 32% | <i>No</i> | 40% |
| Knowledge of RTE | | Consumption for lunch | |
| <i>Low</i> | 25% | <i>Yes</i> | 60% |
| <i>Medium</i> | 39% | <i>No</i> | 40% |
| <i>High</i> | 36% | | |
| Subjective norm | | | |
| <i>Low</i> | 10.79% | | 60% |
| <i>Medium</i> | 51.76% | | 40% |
| <i>High</i> | 37.44% | | |

Table 1: Descriptive Statistics

Table 2 reports the logit regression results performed on the 70% of the overall sample. The table reports the coefficients, standard errors, and significance level. Concerning socio-demographic results, we observed that male respondents are less likely to regularly consume RTE salads compared to female respondents ($p < 0.000$). Respondents between 26 and 30 years old were more likely to consume RTE salads than younger respondents ($p < 0.05$). A lower income was also significantly associated with the consumption of fresh salads. A similar result was observed for respondents in a

| | | β_j | s.e. |
|------------------------------------|-------------------|-----------|--------|
| (Intercept) | | −3.63*** | (0.99) |
| Gender (F) | | −1.04** | (0.39) |
| Age (21-25) | 26-30 | 1.32* | (0.58) |
| | 31-35 | 0.44 | (0.64) |
| | 36-40 | 0.35 | (0.58) |
| | >40 | −0.76 | (0.72) |
| Income (Inc. <15k) | 15k< Inc. ≤ 20k | −1.61** | (0.53) |
| | 20k< Inc. ≤ 25k | −1.12 | (0.58) |
| | Inc. >25k | −0.38 | (0.49) |
| Job Seniority (Entry) | Middle | 1.61*** | (0.45) |
| | Managerial | 0.91 | (0.64) |
| Household size (1-2 people) | 3 people | 0.20 | (0.57) |
| | >3 people | 1.17* | (0.59) |
| Knowledge (Low) | Medium | 0.10 | (0.44) |
| | High | −0.31 | (0.50) |
| Fitness (No) | | 1.39*** | (0.38) |
| Learning | Advertising (0-1) | −1.18** | (0.42) |
| | Social (0-1) | −1.05* | (0.42) |
| Purchasing | Supermarket (0-1) | 1.04 | (0.56) |
| | Conv. store (0-1) | 0.51 | (0.42) |
| Consumption | Lunch (0-1) | 0.61 | (0.38) |
| | Snacking (0-1) | −0.84* | (0.41) |
| Subjective norm (Low) | Medium | 0.42 | (0.70) |
| | High | 1.52* | (0.76) |
| AIC | | 284.67 | |
| BIC | | 372.58 | |
| Log Likelihood | | −118.33 | |
| Deviance | | 236.67 | |

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Table 2: Logit model estimation

middle-level job compared to those at the entry-level. Regarding the other variables, we observed that fitness and subjective norm were positively associated with RTE salad consumption. Social and advertising did not seem to be the main channels through which consumers find information about fresh-cut products. Frequent consumers were more likely to consume RTE salads for lunch and buy them at the supermarket. These results, altogether, are helpful to outline the profile of the typical Chinese consumer of RTE salads. Young, prevalently females, at middle-career level, healthy oriented, pragmatic, and inclined to listen to his/her group reference opinions for food purchasing.

Prediction

In this section, we describe the training phase procedure. Dataset has been randomly split into a training and independent-test set, containing respectively 70% and 30% of the total data. Machine learning model training involves a procedure of hyperparameter fine-tuning until the model is optimized. Specifically, we use 10-fold cross-validation, which means the chosen 70% of the sample is further randomly partitioned into ten subsamples with equal size. Thus, nine subsamples are used as training data and the remaining one is used for validation. The cross-validation process is repeated ten times (folds) such that each of the subsamples is used only once for validation. The hyperparameters are updated based on the average of ten results. The learning phase completes when hyperparameters are optimal, producing minimum prediction errors, then, we test the predictive performance of the final model using the unobserved 30% test sample. Random Forest and Support Vector Machine, require few specifications during the training phases, compared to Deep Neural Networks. In this regard, to select the optimal number of hidden layers, neurons, and parameters used in the DNN, it is common practice to perform a fine-tuning phase. The aim of this phase is to choose an appropriate structure of the model (number of hidden layers, neurons) according to the training error minimization. This choice depends on the type of data that remains a heuristic problem in the field of neural networks. The process consists of feeding

the model the training set and subsequently assessing its accuracy. About the results of the tuning, we have noticed that the architectures with 5 hidden layers using 100 neurons work better than the others on our data. Furthermore, the Rectified Linear Unit (ReLU) activation function, outperformed the other functions tested, (Glorot and Bengio (2011)), ensuring faster learning in networks with many layers.

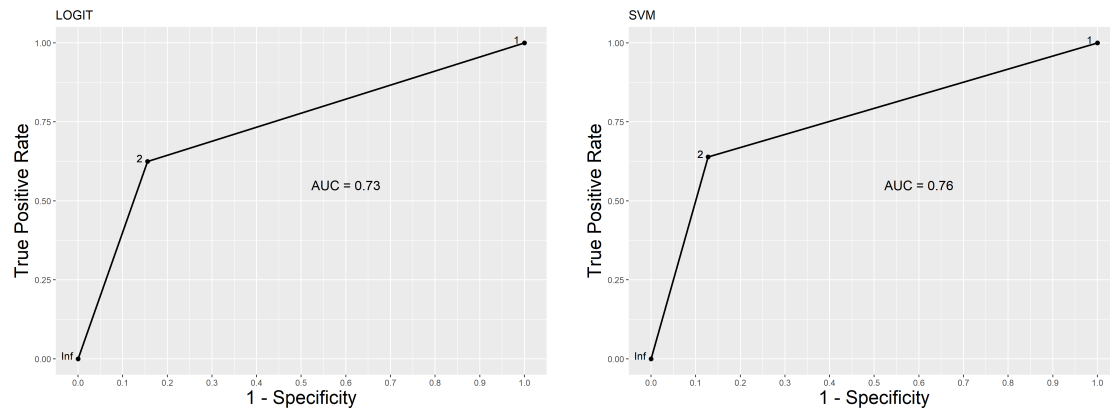


Figure 2: AUC: Logit and SVM

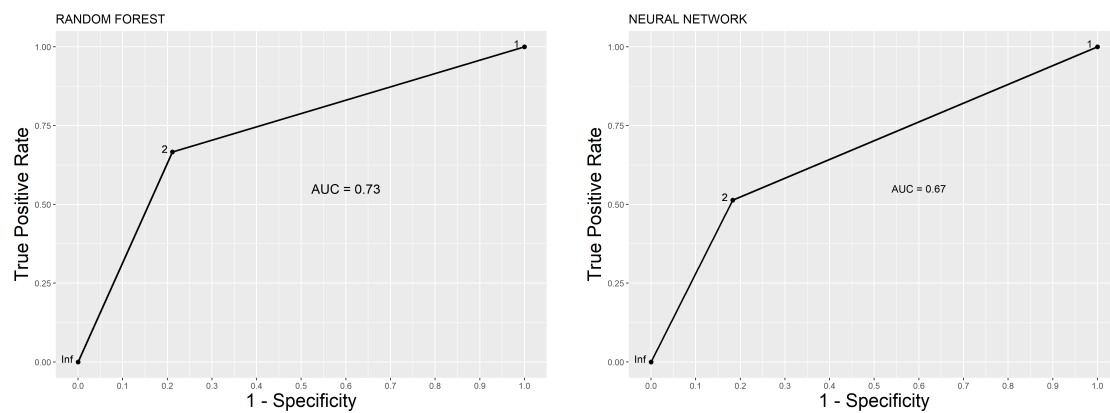


Figure 3: AUC: Random Forest and Deep Neural Network

| Model | Accuracy | AUC | McNemar's Test P-Value | Sens.-Spec. | Kappa |
|----------------|---------------------|------|---------------------------|-------------|-------|
| Logit | 0.78 (0.70-0.85) | 0.73 | 0.84452 | 0.864-0.588 | 0.460 |
| Random Forest | 0.77 (0.68-0.84) | 0.73 | 0.0045 | 0.931-0.352 | 0.332 |
| SVM | 0.80 (0.72-0.87) | 0.76 | 0.83826 | 0.852-677 | 0.519 |
| Neural Network | 0.73 (0.64-0.80) | 0.67 | 0.98 | 0.809-0.528 | 0.33 |

Table 3: Model Performance: Random Forest and Deep Neural Network

Table 3 presents the statistics on the prediction performance in test data. Columns 2–6 show Accuracy, AUC, McNemar test, sensitivity-specificity, and Kappa estimators for each estimated model. We find that overall prediction accuracy ranges between 72% and 80% for our selected ML and Logit models, indicating the robustness of the sample. Furthermore, in each model true positive rates (sensitivity) exceed true negative rates (specificity), which means models perform better at detecting the target consumption behavior. We need to point out that Random Forest shows the McNemar test with a p-value <0.05 , which means that the observed and predicted distribution differs from each other. This finds evidence also in a low Kappa and the highest sensitivity, expressed by Random Forest compared to other models, it seems specialized to detect positive examples.

Discussion

Consumer profiling

The consumer profile that emerges from our analysis is in line with the lifestyle patterns followed by urban young Chinese consumers. RTE salads' versatility contributes to maintaining a healthy lifestyle without giving up on taste and personalization. Our results are consistent with studies conducted in other countries, but there are also

significant differences regarding fresh-cut foods' consumption habits. Looking at the socio-demographic characteristics, similarly to our results, the literature on Western consumers found that the consumption is prevalent among females, middle-income earners, and individuals below forty years old (Massaglia et al. (2019); Sgroi et al. (2018); Van Loo et al. (2010)). Our study also confirms the role of convenience as an essential purchasing factor (Stiletto et al. (2020); Vidal et al. (2013)). The indicator "knowledge" was not significant in our model specification. Still, in its formulation, it aimed at capturing the role of objective information regarding RTE products, including food safety, quality, and freshness. Regarding socio-demographic factors, our study is similar to the findings of a Korean study (Bae et al. (2010)). Preferences for RTE products are significantly higher among young women who care about health. The similarities between the two realities might be associated with the so-called "Hallyu," the Korean wave, meaning South Korea's influence on other countries regarding lifestyles and consumer trends. Another similarity is found with Indian consumers, who obtain information about RTE products through advertising (Hirekenchanagoudar (2008)). Another aspect worth mentioning is the role of subjective norms that is the degree of influence that other people exert on our choices. In our study, respondents giving importance to others' opinions are more likely to consume regularly RTE salads. This finding probably captures the awareness of how important it is to eat fresh vegetables to be healthy. In a collectivist society like China, following social rules is culturally vital for consensus and belonging. The importance of subjective norms in the Chinese context has been acknowledged in several behavioral studies. At the psychological level, it is, for example, adopted to reduce the perceived risk of adopting certain behaviors (Zhong et al. (2012)), to increase self-preventive behaviors (?), or to encourage sustainable choices, like waste recycling (Zhang et al. (2015)).

As an additional contribution, this study provides new insights into the fresh-cut customers' behavior. Thereby using ML models, almost neglected in the analysis of consumer behavior, we offer a comprehensive perspective about predictive tasks, providing

peculiar information for both policymakers and practitioners. At first glance, it may seem surprising that our "non-ML" method would be comparable to our ML methods. The success of the logistic regression depends on how closely the conditional odds ratio can be approximated by a linear function. Indeed, our predictive investigation brings evidence that RF and DNN seem to not bring improvement to our predictions. The latter seems to suffer from the lack of high data dimensionality, requiring more examples to achieve deep abstraction levels (see Lecun et al. (2015)). Our predictor sets have a relatively small number of covariates, and several of the predictors are binary, limiting the scope for model complexity and making it more likely that a simple linear model will yield an adequate approximation. With a larger pool of predictors, we might find that RF or DNN outperform our non-ML methods, which methods are likely to work best will depend on the pool of available predictor sets. On the contrary, the marginal gain in prediction accuracy obtained when using the SVM is larger than other models, confirming the SVM ability as a powerful classifier.

Undoubtedly, the accuracy level would depend on the study context, to the best of our knowledge, literature does not provide a gold standard in predicting RTE consumption behavior, even more in China market. Nevertheless, in the ML field, the 80% of accuracy seems to be reasonable. We can therefore suggest that even in the field of consumer behavior a good combination of statistical modeling in lockstep with machine learning tools might be the best practice³. In this way, it is possible to provide a reasonable solution to the trade-off between interpretability and better predictive performance.

Conclusions

This study's results are a preliminary contribution to an expanding food sector in China and have important policy implications. First, companies and institutions need to build communication campaigns to promote healthy eating habits among the pop-

³Taking Kaggle competitions as a benchmark, a widespread platform among ML practitioners.

ulation. Previous research in the literature showed how healthy marketing campaign towards RTE high-fiber cereals determined a significant growth in the purchasing level that lasted in the long term (Levy & Stokes (1987)) Secondly, other reasons call for food regulator intervention within this market. A concern is, for example, the excessive use of plastic in the packaging (Tarancón et al. (2021)). Consumers worldwide are becoming more sensitive to all the forms of pollution and prefer buying environmentally friendly products (Hartmann & Apaolaza-Ibáñez (2012)). Another criticism of fresh-cut fruits and vegetables is the high amount of waste food that, currently, the production process generates. However, there is evidence that the adoption of novel technologies can significantly reduce its amount (Plazzotta et al. (2017)). Finally, as mentioned above, RTE products are subject to contamination and spoilage if not properly handled (Evans & Redmond (2016)). Therefore, recommending food safety practices for RTE products, such as refrigeration or eating instructions, is another important aspect that requires intervention from policymakers. It also highlighted the cold chain system, required in the transport of fresh cut products in a territory as large as that of China, which is certainly one of the factors that can facilitate the development of local companies. This can undoubtedly create additional entry barriers for international companies targeting to enter the promising Chinese fresh cut products market (ACE Group (China)).

Appendix

Questions included in the analysis

1. Do you regularly consume (about 3 times per week) the RTE salads?
2. What is your gender?
3. Please indicate your age range
4. What is your highest level of completed education?

5. What is your role within the company or organization you work for?
6. How many people live in your household?
7. How do you spend your free time? Fitness
8. Regarding RTE products, which statements do you think are true?
9. RTE products must be washed before being consumed Can be consumed directly
Are treated more than the fresh alternative Are more healthy than the cool ones
Are industrial products very far from the quality of the fresh produce Are more
controlled and selected than the cool
10. How did you learn about these products? Advertising Internet/social media/search
engines
11. Where do you usually buy RTE fresh-cut or packaged fruits and vegetables? Su-
permarket Convenience Store
12. When do you consume RTE products? (multiple options allowed) Snacking
Lunch
13. What is your family's opinion about cut and packaged products?
14. What is your friends ' opinion about cut and packaged products?
15. What is the opinion of your working/study colleagues on cut and packaged prod-
ucts?

Support Vector Machine

Support Vector Machines originated as an alternative to Neural Networks in the wide context of statistical learning, offering a solution to this trade-off by arbitrarily fixing model accuracy and minimising model complexity.

In a classification problem, this translates in an optimal solution identified by a linear hyperplane which correctly classifies observed data and lies as far as possible from them.

Consider the problem of binary classification or dichotomization. Considering the case of a two-dimensional input space, i.e., $\mathbf{x} \in \mathbb{R}^2$. Training data containing l observation of two explanatory variable X_1, X_2 and the realisation of a target class Y , are given as:

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_l, y_l), \mathbf{x} \in \mathbb{R}^n, y \in \{+1, -1\}$$

Data are linearly separable and we aim at find the optimal separating function (among many different hyperplanes) without knowing the underlying probability distribution $P(\mathbf{x}, y)$.

By using given training examples, during the learning stage, our machine finds parameters $\mathbf{w} = [w_1 w_2 \dots w_n]^T$ and b of (a discriminant) or decision function $d(\mathbf{x}, \mathbf{w}, b)$ given as

$$d(\mathbf{x}, \mathbf{w}, b) = \mathbf{w}^T \mathbf{x} + b = \sum_{i=1}^n w_i x_i + b$$

where $\mathbf{x}, \mathbf{w} \in \mathbb{R}^n$, and the scalar b is called a bias. After the successful training stage, by using the weights obtained, the learning machine, given unseen pattern \mathbf{x}_p , produces output according to the following decision rule:

- if $d(\mathbf{x}_p, \mathbf{w}, b) > 0$, the pattern \mathbf{x}_p belongs to $\{C_0\}$ (i.e. , $o = y_1 = +1$) and
- if, $d(\mathbf{x}_p, \mathbf{w}, b) < 0$ the pattern \mathbf{x}_p belongs to $\{C_1\}$ (i.e., $o = y_2 = -1$).

Where $\{C_0, C_1\}$ are the two possible output classes and b is a bias term.

Under these assumptions we define the setting outlined in Figure 4, where the two dashed lines describe the so-called *margin* to be maximised, which equals to

$$M = \min_{\mathbf{x}: y=1} \frac{\mathbf{w}\mathbf{x}}{\|\mathbf{w}\|} - \max_{\mathbf{x}: y=-1} \frac{\mathbf{w}\mathbf{x}}{\|\mathbf{w}\|}, \quad (7)$$

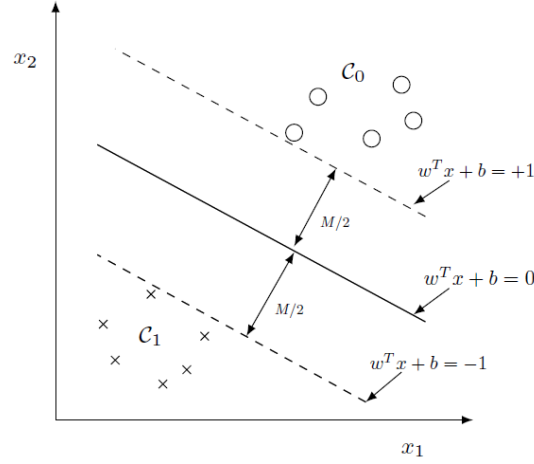


Figure 4: SVM decision boundary and margin.

Since from the decision rule and (7) follows that

$$M = 2 \frac{|\mathbf{w}\mathbf{x}|}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|} \quad (8)$$

that is a geometrical translation of margins. Therefore we can write the SVM optimisation problem as

$$\begin{aligned} \min \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} \\ \text{sub} \quad & y_n [\mathbf{w}^T \mathbf{x}_n + b] \geq 1 \quad \forall n = 1, \dots, N \end{aligned} \quad (9)$$

This is a classic quadratic optimization problem with inequality constraints. Such an optimization problem is solved by the saddle point of the Lagrange functional (Lagrangian):

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{n=1}^N \alpha_n \{y_n [\mathbf{w}^T \mathbf{x}_n + b] - 1\}$$

where the α_i are Lagrange multipliers. The search for an optimal saddle point $(\mathbf{w}_0, b_0, \alpha_0)$ is necessary because Lagrangian L must be minimized with respect to \mathbf{w} and b , and has to be maximized with respect to nonnegative α_i (i.e., $\alpha_i \geq 0$ should be found).

Deep Neural Network

NN training involves an unconstrained optimization problem where the aim is to minimize a function in high dimensional space the so-called loss function, that measures the difference between the predicted values and observed ones. The back-propagation is the most used algorithm for the training of NNs. The algorithm compares the predicted values against the desired ones (objective) and modifies the synaptic weights by back-propagating the gradient of the loss function. Schematically, the procedure alternates forward and backward propagation steps:

- in the forward step, the prediction is computed fixing the synaptic weights,
- in the backward step, the weights are adjusted in order to reduce the error of the network.

The NN iteratively performs forward and backward propagation and modifies the weights to find the combination that minimizes the loss function \mathcal{L} .

It is possible to formally describe the forward pass, to aim at obtaining the target output. Specifically, we have the activation function f^k , weights matrix $W^{(k)}$, hidden layers $H^{(k)}$

where \mathbf{X} e $\mathbf{W}^{(k)}$ are respectively, input and weights vectors and \mathbf{b} the bias column.

Thus for the k -th hidden layer $\mathbf{H}^{(k)}$:

$$\mathbf{H}^{(k)} = f^k(\mathbf{W}^{(k)} \mathbf{H}^{(k-1)} + \mathbf{b}^{(k)}) \quad (10)$$

and thus, we can rewrite as function of the input \mathbf{X} :

$$\mathbf{H}^{(k)} = f^{(k)}\left(\mathbf{W}^{(k)} \underbrace{f^{(k-1)}\left(\dots f^{(1)}\left(\mathbf{W}^{(1)} \mathbf{X} + \mathbf{b}^{(1)}\right) \dots\right)}_{=\mathbf{H}^{(k-1)}} + \mathbf{b}^{(k)}\right) \quad (11)$$

After the estimation procedure, which implies to estimate the weights, we obtain the final output, \hat{y} , resulting from the application of the DNN parameters (weights matrix $\hat{\mathbf{W}}$ and bias $\hat{\mathbf{b}}$) obtained by the optimization procedure described in the following.

NN, like other machine learning techniques, requires the splitting of the dataset into a training and a testing set. The training set stands for supervised learning, while the testing set is used to validate the model. After the training phase, the network has learned the input–output functional relationship and it should be able to predict future values using only the input. Practically, NN parameters are obtained by the minimization of the overall loss function \mathcal{L} . In the classification problem the most widely used loss function is the binary cross-entropy:

$$\mathcal{L}(y, \hat{y}) = - \sum_{i=1}^n y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)$$

To minimize the loss function, we use the Gradient Descent optimization algorithm, which iteratively moves in the direction of steepest descent as defined by the negative of the gradient.

This algorithm proceeds by minimizing \mathcal{L} at each step t , therefore differentiating the loss function with respect to the weights (\mathbf{W}). For the generic weights $w_{n,n}^{(k)}$ and the k -th layer, the algorithm proceeds using the chain derivation rule described in the following equation:

$$\frac{\partial \mathcal{L}(y, \hat{y})}{\partial w_{n,n}^{(k)}} = \frac{\partial \mathcal{L}(y, \hat{y})}{\partial H_n^{(k)}} \frac{\partial H_n^{(k)}}{\partial z_n^{(k)}} \frac{\partial z_n^{(k)}}{\partial w_{n,n}^{(k)}} \quad (12)$$

where $z_n^{(k)} = w_n^{(k)} H_n^{(k-1)} + b_n^{(k)}$. To update the weights ($\tilde{\mathbf{W}}$), the gradient of the loss function, $\nabla \mathcal{L}_t(y, \hat{y})$, is multiplied by a scalar, η , often called learning rate, according to the following scheme:

$$\tilde{\mathbf{W}} = \mathbf{W} - \eta \nabla \mathcal{L}_t(y, \hat{y}) \quad (13)$$

In a figurative way, the idea behind the gradient descent is similar to “climbing down a hill” until a global or local minimum is reached. At each update, the search moves in the opposite direction of the gradient and the learning rate η determines the amplitude of this movement, controlling the adjustment in the weights, thus determining how fast or slow we will move towards the optimal weights. A very large learning rate leads to

a sub-optimal solution. A very small learning rate involves too many iterations to find the optimal solution. Then, the learning rate can be considered the most important hyper-parameter for tuning NN.

The search for the optimal parameters is then carried out through an optimization process where the NN initial weights are selected in an arbitrary (random) way so they are not optimal parameters. The iterations of the algorithm lead to the optimization of the weights and minimization of the error. The choices concerning the type of architecture (e.g., the number of hidden layers, units for each layer) and the hyperparameter (e.g., learning rate, activation functions, and loss function), remains a heuristic problem for NN users: the choice often depends on the type of data and it is not an easy step.

Random Forest

The random forest (RF) algorithm creates a collection of decision trees from a casually variant of the tree. Once one specific learning set is defined, the RF presents a random perturbation to the learning procedure and in this way a differentiation among the trees is produced. Successively the predictions of all these trees is derived through the implemmentation of aggregation techniques. The first aggregation procedure was described by Breiman (1996); the authors proposed the well know bagging based on random bootstrap copies of the original data to assemble different trees. Later in 2001 the same authors Breiman (2001)r proposed the random forest as an extention of the procedure of the bagging such that it combines the bootstrap with randomization of the input variables to separate internal nodes t . This means that the algorithm does not identify the best split $s_t = s^*$ among all variables, but firstly creates a random subset of K variables for each node and among them determines the best split.

The RF estimator, for both regression and classification problem, of the target variable \hat{y}_{R_j} is a function of the regression or classification tree estimator, $f^{\text{tree}}(\mathbf{X}) = \sum_{j \in J} \hat{y}_{R_j} \mathbf{1}_{\{\mathbf{X} \in R_j\}}$, where $\mathbf{X} = (X_1, X_2, \dots, X_p)$ is the vector of the predictors, $\mathbf{1}_{\{-\}}$ represents the indicator function and $(R_j)_{j \in J}$ are the regions of the predictors space

obtained by minimizing the binary cross entropy Loss Function Therefore, denoting the number of bootstrap samples by B and the decision tree estimator developed on the sample $b \in B$ by $\hat{f}^{\text{tree}}(\mathbf{X} | b)$, the RF estimator is defined as follows:

$$f^{RF}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B f^{\text{tree}}(\mathbf{X} | b)$$

References

- Acock, A. C. (2008). A gentle introduction to Stata. Stata press.
- Artés, F., Gómez, P., Aguayo, E., Escalona, V., Artés-Hernández, F. (2009). Sustainable sanitation techniques for keeping quality and safety of fresh-cut plant commodities. *Postharvest Biology and Technology*, 51(3), 287–296.
- Bae, H.-J., Chae, M.-J., Ryu, K. (2010). Consumer behaviors towards ready-to-eat foods based on food-related lifestyles in Korea. *Nutrition Research and Practice*, 4(4), 332.
- Baselice, A., Colantuoni, F., Lass, D. A., Nardone, G., Stasi, A. (2017). Trends in EU consumers' attitude towards fresh-cut fruit and vegetables. *Food Quality and Preference*, 59, 87-96.
- Bischl, B., Mersmann, O., Trautmann, H., Weihs, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary computation*, 20(2), 249-275.
- Brennan, M. A., Derbyshire, E., Tiwari, B. K., Brennan, C. S. (2013). Ready to eat snack products: The role of extrusion technology in developing consumer acceptable and nutritious snacks. *International Journal of Food Science & Technology*, 48(5), 893–902.
- Breiman, L. (1996). Bagging predictors. *Machine learning* 24 (2), 123-140

- Breiman, L. (2001). Random forests. *Machine learning* 45 (1), 5-32
- Buckley, M., Cowan, C., McCarthy, M. (2007). The convenience food market in Great Britain: Convenience food lifestyle (CFL) segments. *Appetite*, 49(3), 600-617.
- Cassady, D., Jetter, K. M., Culp, J. (2007). Is price a barrier to eating more fruits and vegetables for low-income families?. *Journal of the American Dietetic Association*, 107(11), 1909-1915.
- Cohen J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*. 20(1):37-46.
- The State Council of China. China Nutrition Improvement Action Plan (1996-2000). The State Council of China, Beijing; 1997.
- Chonpracha, P., Ardoin, R., Gao, Y., Waimaleongora-Ek, P., Tuuri, G., Prinyawiwatkul, W. (2020). Effects of intrinsic and extrinsic visual cues on consumer emotion and purchase intent: A case of ready-to-eat salad. *Foods*, 9(4), 396.
- Daxue Consulting. (2021). Survey on Health Perceptions in China. <https://daxueconsulting.com/health-perceptions-in-china/>
- Dinnella, C., Torri, L., Caporale, G., Monteleone, E. (2014). An exploratory study of sensory attributes and consumer traits underlying liking for and perceptions of freshness for ready to eat mixed salad leaves in Italy. *Food Research International*, 59, 108–116.
- Evans, E. W., Redmond, E. C. (2016). Older adult consumer knowledge, attitudes, and self-reported storage practices of ready-to-eat food products and risks associated with listeriosis. *Journal of Food Protection*, 79(2), 263–272.
- Faostat, F. A. O. (2017). Available online: <http://www.fao.org/faostat/en/#data.QC> (accessed on January 2020).

- Frewer, L. J., Risvik, E., Schifferstein, H. (Eds.). (2013). Food, people and society: a European perspective of consumers' food choices. Springer Science Business Media.
- Gibson, A. A., Partridge, S. R. (2019). Nutritional qualities of commercial meal kit subscription services in Australia. *Nutrients*, 11(11), 2679.
- Gibson, A., Edgar, J. D., Neville, C. E., Gilchrist, S. E., McKinley, M. C., Patterson, C. C., Young, I. S., Woodside, J. V. (2012). Effect of fruit and vegetable consumption on immune function in older people: A randomized controlled trial. *The American Journal of Clinical Nutrition*, 96(6), 1429–1436.
- Gordon, R., McDermott, L., Stead, M., Angus, K. (2006). The effectiveness of social marketing interventions for health improvement: what's the evidence?. *Public health*, 120(12), 1133-1139.
- Glorot, X. and Bengio, A. B. . Y. (2011). Deep sparse rectifier neural networks. In Gordon, G., Dunson, D., and Dudík, M., editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA. JMLR Workshop and Conference Proceedings
- Gu, Y., He, Y., Ali, S. H., Harper, K., Dong, H., Gittelsohn, J. (2021). Fruit and Vegetable Intake and All-Cause Mortality in a Chinese Population: The China Health and Nutrition Survey. *International Journal of Environmental Research and Public Health*, 18(1), 342.
- Hansstein, F., Keqiang, W., Hongmei, L. (2017). Perceptions of food quality: Evidence from a survey in Shanghai. *International Journal of Consumer Studies*, 41(6), 754–760.
- Hartmann, P., Apaolaza-Ibañez, V. (2012). Consumer attitude and purchase intention toward green energy brands: The roles of psychological benefits and environmental concern. *Journal of Business Research*, 65(9), 1254–1263.

- Hastie, T., Tibshirani, R. and Friedman, J. The Elements of Statistical Learning. Springer Series in Statistics Springer New York Inc., New York, NY, USA, (2009)
- The State Council of China. Healthy China Action Plan (2019–2030). The State Council of China, Beijing; 2019.
- HFG - Law and Intellectual Property. (2016). Food Safety Law of the People's Republic of China (2015). www.hfgip.com
- Hicks, D. T., Pivarnik, L. F., McDermott, R., Richard, N., Hoover, D. G., Kniel, K. E. (2009). Consumer awareness and willingness to pay for high pressure processing of ready to eat food. *Journal of Food Science Education*, 8(2), 32–38.
- Hirekenchanagoudar, R. (2008). Consumer behaviour towards ready-to-eat food products.
- James, J. B., Ngarmsak, T., Rolle, R. S. (2010). Processing of fresh-cut tropical fruits and vegetables: A technical guide. RAP Publication (FAO) eng no. 2010/16.
- Kim, J. (2007). Fresh-cut market potential and challenges in Far-East Asia. 33–38.
- Kwon, Y.-S., Ju, S. (2018). Sensory evaluation of commercial ready-to-eat rice between trained panelist and consumer. *British Food Journal*.
- Lecun, Y. and Bengio, Yoshua Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.
- Levy, A. S., Stokes, R. C. (1987). Effects of a health promotion advertising campaign on sales of ready-to-eat cereals. *Public Health Reports*, 102(4), 398.
- Liem, D., Bolhuis, D., Hu, X., Keast, R. (2016). Influence of labeling on Australian and Chinese consumers' liking of milk with short (pasteurized) and long (UHT) shelf life. *Journal of Dairy Science*, 99(3), 1747–1754.

- Lilavanichakul, A., Chaveesuk, R., Kessuvan, A. (2018). Classifying Consumer Purchasing Decision for Imported Ready-to-Eat Foods in China Using Comparative Models. *Journal of Asia-Pacific Business*, 19(4), 286–298.
- Marshall, D., Anderson, A. S., Lean, M., Foster, A. (1994). Healthy eating: fruit and vegetables in Scotland. *British Food Journal*.
- Massaglia, S., Merlino, V. M., Borra, D., Bargetto, A., Sottile, F., Peano, C. (2019). Consumer Attitudes and Preference Exploration towards Fresh-Cut Salads Using Best–Worst Scaling and Latent Class Analysis. *Foods*, 8(11), 568.
- The Ministry of Health of China. Management Measures for Nutrition Improvement. The Ministry of Health of China, Beijing;2010.
- Mazzocchi, M., Traill, W. B., Shogren, J. F. (2009). *Fat economics: nutrition, health, and economic policy*. Oxford University Press.
- The State Council of China. National Nutrition Plan (2017–2030). The State Council of China; 2017.
- Nevo, A. (2011). Empirical models of consumer behavior. *Annu. Rev. Econ.*, 3(1), 51–75.
- Pilone, V., Stasi, A., Baselice, A. (2017). Quality preferences and pricing of fresh-cut salads in Italy: New evidence from market data. *British Food Journal*.
- Plazzotta, S., Manzocco, L., Nicoli, M. C. (2017). Fruit and vegetable waste management and the challenge of fresh-cut salad. *Trends in Food Science Technology*, 63, 51–59.
- Pollard, J., Kirk, S. L., Cade, J. E. (2002). Factors affecting food choice in relation to fruit and vegetable intake: a review. *Nutrition research reviews*, 15(2), 373–387.

- Poti, J. M., Mendez, M. A., Ng, S. W., Popkin, B. M. (2016). Highly processed and ready-to-eat packaged food and beverage purchases differ by race/ethnicity among US households. *The Journal of Nutrition*, 146(9), 1722–1730.
- Rico, D., Martin Diana, A. B., Barat, J. M., Barry-Ryan, C. (2007). Extending and measuring the quality of fresh-cut fruit and vegetables: a review. *Trends in Food Science Technology*, 18(7), 373-386.
- Roberts, J. H., Lilien, G. L. (1993). Explanatory and predictive models of consumer behavior. *Handbooks in Operations Research and Management Science*, 5, 27–82.
- Santeramo, F. G., Carlucci, D., De Devitiis, B., Seccia, A., Stasi, A., Viscecchia, R., Nardone, G. (2018). Emerging trends in European food, diets and food industry. *Food Research International*, 104, 39–47.
- K. Seiders & R., Petty. (2004). Obesity and the Role of Food Marketing: A Policy Analysis of Issues and Remedies. *Journal of Public Policy & Marketing*
- Si, H., Shen, L., Liu, W., Wu, G. (2021). Uncovering people's mask-saving intentions and behaviors in the post-COVID-19 period: Evidence from China. *Sustainable Cities and Society*, 65, 102626.
- Singla, G., Chaturvedi, K., Sandhu, P. P. (2020). Status and recent trends in fresh-cut fruits and vegetables. In *Fresh-cut fruits and vegetables* (pp. 17–49). Elsevier.
- Sgroi, F., Piraino, F., Donia, E. (2018). Determinants of ready-to-eat products purchase intentions: An empirical study among the Italian consumers. *HortScience*, 53(5), 656–660.
- Soliva Fortuny, R. C., Biosca, M., Grigelmo Miguel, N., Martín Belloso, O. (2002). Browning, polyphenol oxidase activity and headspace gas composition during storage of minimally processed pears using modified atmosphere packaging. *Journal of the Science of Food and Agriculture*, 82(13), 1490-1496.

- Stiletto, A., Giampietri, E., Trestini, S. (2020). Heterogeneity in consumer preferences for ready-to-eat pomegranate: An empirical study in Italy. *British Food Journal*.
- Storm, H., Baylis, K., Heckelee, T. (2020). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*, 47(3), 849-892.
- Tarancón, P., Fernández-Serrano, P., Besada, C. (2021). Consumer perception of situational appropriateness for fresh, dehydrated and fresh-cut fruits. *Food Research International*, 140, 110000.
- Thienhirun, S., Chung, S. (2018). Consumer attitudes and preferences toward cross-cultural ready-to-eat (RTE) food. *Journal of Food Products Marketing*, 24(1), 56–79.
- Thike, T. Z., Saw, Y. M., Lin, H., Chit, K., Tun, A. B., Htet, H., Cho, S. M., Khine, A. T., Saw, T. N., Kariya, T. (2020). Association between body mass index and ready-to-eat food consumption among sedentary staff in Nay Pyi Taw union territory, Myanmar. *BMC Public Health*, 20(1), 206.
- Van Loo, E. J., Ricke, S. C., Milillo, S. R., Seideman, S., Crandall, P. G. (2010). Consumer food safety perceptions of ready-to-eat deli foods in Northwest Arkansas. *Food Protection Trends*, 30(11), 635-643.
- Vidal, L., Ares, G., Giménez, A. (2013). Projective techniques to uncover consumer perception: Application of three methodologies to ready-to-eat salads. *Food Quality and Preference*, 28(1), 1–7.
- Watada, J., Roy, A., Kadikar, R., Pham, H., Xu, B. (2019). Emerging trends, techniques and open issues of containerization: a review. *IEEE Access*, 7, 152443-152472.
- World Health Organization. (2008). Microbiological hazards in fresh leafy vegetables and herbs: meeting report (Vol. 14). World Health Organization.
- Xiao, Y., Su, C., Ouyang, Y., Zhang, B. (2015). Trends of vegetables and fruits consumption among Chinese adults aged 18 to 44 years old from 1991 to 2011.

Zhonghua Liu Xing Bing Xue Za Zhi= Zhonghua Liuxingbingxue Zazhi, 36(3), 232–236.

Zhang, B., Zhong, Z., Min, M., Ding, K., Xie, Q., Ruan, R. (2015). Catalytic fast co-pyrolysis of biomass and food waste to produce aromatics: Analytical Py–GC/MS study. *Bioresource technology*, 189, 30-35.

Zhong, B., Huang, Y., Liu, Q. (2021). Mental health toll from the coronavirus: Social media usage reveals Wuhan residents' depression and secondary trauma in the COVID-19 outbreak. *Computers in human behavior*, 114, 106524.

ACE Group (China) Co., Ltd. (<https://www.acejpn.com/about/group.html>)