Serv Bus
https://doi.org/10.1007/s11628-018-0371-z

EMPIRICAL ARTICLE

A predictive investigation of first-time customer retention in online reservation services

Yen-Chun Chou[1] · **Howard Hao-Chun Chuang**[1]

Received: 25 August 2017 / Accepted: 29 March 2018

Abstract This paper reports a predictive investigation of first-time customer retention in an emerging service business—online reservation services. We work with an online platform that enables customers to make reservations for various types of restaurants. With numerous first-time users on the platform, the focal company is eager to effectively identify recurring customers. However, the business problem is challenging due to that each first-time customer has one and only one booking record hinders the use of well-established marketing models that demand multiple booking records for a customer. By analyzing more than 100,000 observations, we extract booking-related features that are useful in predicting first-time customer retention. Our feature extraction is potentially applicable to other service sectors (e.g., hotel, travel) with similar booking information fields (e.g., reservation timing, party size). We further conduct a comparative study in which surprisingly, the seemingly simplistic generalized additive model (GAM) for our test cases consistently outperforms computationally intensive ensemble learning methods, even the cutting-edge *XGBoost*. Our analysis indicates that there is no silver bullet for applied predictive modeling and GAM should definitely be included in the arsenal of business researchers. We conclude by discussing the implications of our study for online service providers and business data analytics.

Keywords E-services · First-time customer retention · Prediction · Analytics · Statistical learning

✉ Howard Hao-Chun Chuang
chuang@nccu.edu.tw

Yen-Chun Chou
yenchun@nccu.edu.tw

[1] College of Commerce, National Chengchi University, Taipei 11605, Taiwan

Published online: 02 April 2018

1 Introduction

Owing to the Internet, various online service platforms for travel, hotel, and rental industries have emerged. With the continued advent of new online activities, our study focuses on an emerging service business: *online reservation services*. The focal company in our paper develops an online platform for customers to make reservations for various types of restaurants at desired time slots. Such services utilize the wide access of the Internet to bring online customers across regions to offline restaurants (online-to-offline) across countries in East Asia. So far, the company has over one million users, and its reservation services cover several thousands of restaurants and five-star hotels. Through its online-to-offline services, the platform attracts many new customers and accumulates numerous customers' booking records. Each booking record is composed of various variables including (1) customer's basic characteristics such as age and gender, (2) customer's booking information such as the party size and timing of a dining reservation, and (3) restaurant attributes such as locations and types.

Having obtained a huge amount of booking records, one of the most challenging questions the focal company seeks to answer is how to predict whether a *first-time customer* would reuse the online reservation service or not. Being able to predict the return of those first-timers would help service business grow. However, unlike a repeated-customer who has multiple historical booking records such that well-established marketing models (e.g., Pareto-NBD) can be applied to predicting probabilities to return, each first-time customer has *one* and *only one* available booking record. Given highly limited customer features, we aim to empirically address two questions: "What are the critical features that can be extracted from booking records to predict whether a first-time customer re-uses online reservation services or not? Which methods would be efficient and effective in predicting first-time customer retention in our case?"

To answer the research questions, we obtain more than 100,000 unique first-time customers' booking records from the platform to develop predictive models of customer retention. Our study includes two types of statistical learning methods: *regression-based* models and *ensemble-based* models. Regression-based models are commonly seen in prior studies because their linear and additive structure is good for *inference*. In contrast, ensemble learning models have non-linear and non-additive functional forms that allow for model *flexibility*. In this study, we adopt both types of methods that represent different levels of model *interpretability* and *flexibility*, and compare their prediction performance. Specifically, for the regression-based prediction, we start with the often-used logistic regression for the binary response of customer retention (return or not), and further relax the assumed linearity between response and predictor variables using the generalized additive model (GAM) (Hastie and Tibshirani 2016). For the ensemble-based prediction, we test tree-based bagging, random forest, and several boosting algorithms to see whether extra flexibility indeed improves prediction performance.

To our surprise, the prediction performance of regression-based GAM for our test cases is consistently better than other flexible and computationally intensive

ensemble-based techniques. Instead of increasing model complexity, we find that identifying meaningful variables from the business domain of reservation services is more useful than adopting sophisticated machine learning algorithms. Our analysis indicates that variables about member booking behaviors such as timing and final status on the dining date would substantially enhance prediction performance. The result is consistent across all of the tested model specifications.

Our predictive investigation into emerging online services contributes to the research and practice as well. First, extant e-commerce/e-service literature tends to put a predominant focus on explanatory modeling but under-investigates predictive modeling that has rich applications in different industry sectors (Shmueli and Koppius 2011). However, variables that are valuable for explanation may not be useful for prediction. In our study, while almost all of the listed variables are statistically significant, only some of the variables are influential for predicting first-time customer retention. For instance, widely used age and gender (Swart and Roodt 2015) have low predictive power despite their statistical significance. Such differences reiterate the necessity of distinguishing between explanatory and predictive modeling.

Second, our results inform service providers the relevance of features about booking behaviors in predicting first-time customer retention. Despite our research context in the booking behaviors in the dinning industry, our feature extraction and model construction are potentially applicable to other service contexts where customers book services with many information fields (e.g., reservation timing, party size) identical to our context. Moreover, booking-related variables (e.g., status of first reservation) can be easily collected and less prone to missing/entry errors often seen in variables like age and gender. Last, advances in computing power ease the training of complex models, and hence some practitioners may intrinsically prefer sophisticated prediction algorithms. However, in our case, even the cutting-edge *XGBoost* (Chen and Guestrin 2016) method only results in prediction performance comparable to the seemingly simplistic GAM. Our analysis suggests that predicting first-time customer retention is a fairly challenging problem that cannot be easily tackled by ensemble learning.

The rest of this article is organized as follows. In Sect. 2, we review related literature, and in Sect. 3, we describe in detail the problem setting and data. In Sect. 4, we present prediction results of generalized additive modeling under different model specifications. In Sect. 5, we compare and contrast the regression-based approach and other five ensemble learning techniques. We conclude the paper with a discussion of implications for managers and researchers.

2 Literature review

The use of statistical modeling for causal *explanation* is distinct from that for *prediction*. However, most social science/management studies tend to put significantly more emphasis on explanatory models aimed at minimizing bias to test theories/hypotheses, but underemphasize predictive models aimed at minimizing the total of bias and variance for predicting new observations (Shmueli 2010).

Searching the 1072 papers published in the top-rated journals—*Information Systems Research* and *MIS Quarterly*—between 1990 and 2006, Shmueli and Koppius (2011) found only 52 empirical papers with predictive claims, of which only seven carried out proper predictive investigation. In fact, many existing studies fail to recognize that *predicting* customer behaviors is an entirely different story from *explaining* customer behaviors. For example, Netflix is known for its predicting algorithm for customer preferences. As Netflix rolled out to 130 new countries in 2015, one might expect *demographic data* as an important input to its algorithm. However, its VP of product Todd Yellin said, "Geography, age, and gender? We put that in the garbage heap" (Morris 2016).

Even though in recent years more and more offline/online service firms call for accurately *predicting customer behaviors* in platforms such as Kaggle, most empirical analyses (among prior e-commerce/e-service studies on customer behaviors) still focus on *explanatory* power of their underlying causal/relational models. Those studies concentrate on factors that can *explain* customers' online behaviors such as perceptions about unique features of mobile data services (e.g., Hong and Tam 2006; Lee et al. 2009), online auction decisions (e.g., Vakrat and Seidmann 2000; Bapna et al. 2009), online browsing and perceptions about websites (Hong et al. 2004; Schmutz et al. 2010), and so forth. The foregoing studies put a strong emphasis on in-sample explanation for theory testing and building. In spite of the practical value of predictive models in business services (Olson 2007), relatively few studies have applied *predictive modeling* to online customer behaviors. Among the studies wishing to generate out-of-sample predictions, predicting a *customer's purchase behavior* is of primary interest.

On one hand, a customer's purchase behavior can be depicted as whether a customer would purchase or not, i.e., *conversion*. The use of statistical modeling and machine learning to predict conversion has received increasing attention from service researchers in the last few years. Van den Poel and Buckinx (2005) show that detailed clickstream information are influential to predict whether a customer is converted from a viewer to a buyer. Kim et al. (2013) develop a support vector machine approach (SVM) to predict customers' purchase response to catalog mailing. Migueis et al. (2017) address the same question of predicting customers' response to direct marketing in banking services using random forests to in banking. On the other hand, a related customer purchase behavior is whether a customer buys or uses again, i.e., *retention*. Morrisonn et al. (2001) posit that most factors affecting whether an online customer would buy or not are quite different from those affecting whether an online customer would *return or not* (i.e., retention). Customer retention in different service industries (e.g., telecom, insurance) is a fruitful area for predictive modeling (Olson 2007). In fact, empirical evidence shows that just a small percentage of improvement in retention rates could lead to non-trivial profit increase (Van den Poel and Lariviere 2004). This offers a compelling example that motivates us to empirically analyze customer retention in service business. That said, our problem setting—predicting first-time customer retention—is unique and hinders the direct use of probabilistic marketing models for repeated customers with multiple historical records. Hence, our analysis protocol (e.g., feature engineering, model fitting) is different from previous studies on predicting customer retention.

In addition, most prior studies on predictions of customer retention are related to offline service industries. For example, Xie et al. (2009) apply improved balanced random forests to enhance performance of retention prediction for banks in China. Similarly, Coussement and Van den Poel (2008) predict customer retention of newspaper subscription services. As e-commerce becomes prosperous, more and more consumers shop online. Different from the offline service context, the online stores open 24/7 in order to provide convenient and reachable services to customers. In addition, through the online storefront, firms could collect customer browsing and purchasing information, and provide customized offerings and promotions accordingly. Shankar et al. (2003) argue that customer loyalty to an online service provider is higher than customer loyalty to an offline service provider. Thus, customer retention is of great importance in the context of online service business. In response to the call for more applied predictive modeling research (Shmueli 2010) and the practical need of an online service provider, we build prediction models for customer retention using a large dataset. Our paper differentiates from prior studies on customer retention (e.g., Hosseini and Bideh 2014; Swart and Roodt 2015) by focusing on first-time customers with highly limited information, creating challenges for model development and leaving research gaps to be filled.

3 Data and problem setting

We obtain a dataset that contains over 100,000 first-time customers' bookings records from an online reservation service provider in Taiwan. Each record stands for a reservation made by a unique first-time member. Being a leading intermediary in the online reservation business, the company charges no fees from its members and earns its profits from charging restaurants and selling vouchers. In the process of expanding its service business, the online platform keeps attracting many new users (i.e., first-time customers) from different regions. The firm is eager to identify first-time customers who are likely to reuse its service, as its revenue comes from every successfully executed dining reservation. The company must make predictions based on limited features regarding each customer with one and only one historical booking record. Figure 1 summarizes variables in our problem setting. The focal

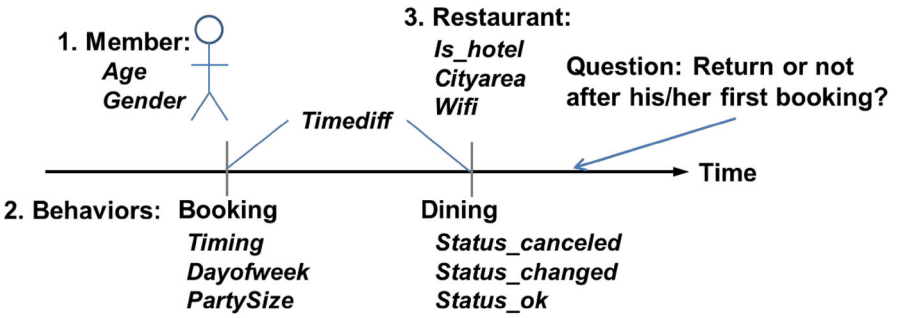

Fig. 1 Data and problem setting

question is whether a member will return to use the service again, after booking for the first time via the platform (i.e., website or mobile APP). Effective prediction models are then needed to address the focal question.

As shown in Fig. 1, the data available for our predictive investigation contains three sets of information related to member demographics, booking behaviors, and restaurant attributes. *Age* and *gender* (1 Male; 0 Female) are the only two variables on member demographics. Such a small number is primarily due to that the company deliberately simplifies registration processes to attract people and requires minimum information from members. Restaurant attributes are also fairly limited with only *Is_hotel* (1 Affiliated with a hotel; 0 Not), *Cityarea* (county and districts of restaurants), and *Wifi* (1 Wifi is available; 0 Not). That said, several features about booking behaviors can be extracted from the timing and outcome of each reservation. Those features can be potentially useful for characterizing service usage patterns and predicting customer retention.

From the booking date (the time reservation being made), we calculate *Timing* (the booking date is during which week of which year), *DayofWeek* (day of the week of the booking date), and *PartySize* (number of people expected to appear on the dining date). *Timing* allows us to capture some unobserved exogenous random shocks during a particular week in that year. For instance, in a certain week some restaurants may offer incentives to allure booking due to holidays, promotions, new-opening, etc. *DayofWeek* also partially captures unobserved exogenous events. *PartySize* partially captures the purpose of dining. For instance, a large *PartySize* is likely to be associated more formal/rate gathering.

Given the booking date and dining date, we further calculate *Timediff*—the number of days between booking date and dining date—for each record. This variable to some extent reflects behavior heterogeneity, i.e., make reservation early before the dining date or not. Finally, not all reservations would be eventually executed by those first-time customers. On the dining date, one of the following four outcomes—no show, okay, cancelled, or changed—would be observed by restaurants and later reported to the service provider. While the focal firm has no way to know underlying causes of realized outcomes, those outcomes may be related to members' intention to make reservation via the platform again after the dining date. For example, a cancelation could be due to something unexpected/customers' badwill, which may be associated with chances of booking again. Hence, we incorporate three binary variables—*Status_canceled*, *Status_changed*, and *Status_ok*—into our modeling framework.

In total, there are over 100,000 booking records (each made by a unique first-time member) in our data. Following standard practices in data mining/predictive modeling, we randomly split the data into training and test sets. The training set has 62,083 records ($\sim$ 60% of all observations), and the test set has 41,546 records ($\sim$ 40% of observations). Table 1 shows summary statistics of variables in training and test sets. The response variable to be predicted—*Return*—is a binary attribute where 1 denotes that the member returns to use the e-service to make a reservation again in 90 days after the current booking in our data. Note that the threshold of 90 days is determined by the focal company and we have no access to exact number of days until the next booking. Our prediction target—customer retention

Table 1 summary statistics of training and test sets

Variables	*n*	Mean	SD	Min	Max
Response variable					
Return (training)	62,083	0.20	0.40	0.00	1.00
Return (test)	41,456	0.20	0.40	0.00	1.00
Member demographics					
Age (training)	39,839	32.33	9.45	0.00	114.00
Age1	62,083	0.15	0.36	0.00	1.00
Age2	62,083	0.29	0.45	0.00	1.00
Age3	62,083	0.15	0.35	0.00	1.00
Age (test)	26,566	32.37	9.46	0.00	114.00
Age1	6144	0.15	0.35	0.00	1.00
Age2	11,874	0.29	0.15	0.00	1.00
Age3	6046	0.15	0.35	0.00	1.00
Gender (training)	62,083	0.45	0.50	0.00	1.00
Gender (test)	41,456	0.45	0.50	0.00	1.00
Booking behaviors					
Timing (training)	62,083	85.15	34.47	1	147
Timing (test)	41,456	85.21	34.49	1	147
Dayofweek (training)	62,083	3.95	1.90	1.00	7.00
Dayofweek (test)	41,456	3.98	1.89	1.00	7.00
Timediff (training)	62,083	9.75	13.57	0.00	142.00
Timediff (test)	41,456	9.68	13.39	0.00	147.00
Partysize (training)	62,083	4.05	2.93	1.00	45.00
Partysize (test)	41,456	4.06	3.00	1.00	39.00
Status (training)					
Status_ok	62,083	0.73	0.45	0.00	1.00
Status_canceled	62,083	0.19	0.40	0.00	1.00
Status_changed	62,083	0.01	0.12	0.00	1.00
Status (test)					
Status_ok	41,456	0.72	0.45	0.00	1.00
Status_canceled	41,456	0.20	0.40	0.00	1.00
Status_changed	41,456	0.02	0.12	0.00	1.00
Restaurant attributes					
Is_hotel (training)	62,083	0.36	0.48	0.00	1.00
Is_hotel (test)	41,456	0.36	0.48	0.00	1.00
Cityarea (training)					
Area1	62,083	0.12	0.33	0.00	1.00
Area2	62,083	0.19	0.39	0.00	1.00
Area3	62,083	0.03	0.17	0.00	1.00
Area4	62,083	0.06	0.23	0.00	1.00
Cityarea (test)					
Area1	41,456	0.12	0.33	0.00	1.00
Area2	41,456	0.19	0.39	0.00	1.00
Area3	41,456	0.03	0.17	0.00	1.00

Table 1 continued

Variables	*n*	Mean	SD	Min	Max
Area4	41,456	0.06	0.23	0.00	1.00
Wifi (training)	62,083	0.61	0.49	0.00	1.00
Wifi (test)	41,456	0.61	0.49	0.00	1.00

(*Return*)—is binary and slightly imbalanced as the ratio of return-to-not return is about 1:4.

For the predictor variables introduced above, we categorize *Age* and *Cityarea* into groups. *Age* is noisy and has a non-trivial fraction with faulty/arbitrary entries (e.g., minimum of 0 and maximum of 142 in the training set). Based on the available observations, we create three binary variables *Age*1 (1 Age 16–25), *Age*2 (1 Age: 26–35), and *Age*3 (1 Age 36–45). The three variables would be zero for records with *missing Age*, $Age < 16$, or $Age > 45$. For *Cityarea*, we create four binary variables for districts/counties considered much more developed than others. Further, for *Timing* we set the earliest week-year combination in our data as 1, and *Timing* increases by 1 unit on a week-to-week basis. Also, *Day of Week* ranges from 1 to 7 where 1 stands for Sunday.

In short, as the full data are randomly split, and given the large number of observations, the training and test sets show neither statistical nor practical differences in all attributes listed in Table 1. In the next section, we will fit different kinds of predictive models using the training set and evaluate out-of-sample prediction performance using the test set.

4 Predictive modeling: generalized additive model

To begin with, we apply generalized additive models (GAM) developed by Hastie and Tibshirani (1990). Compared to linear and interpretable regression models, GAM is a much more powerful technique that allows predictor variables and response variables to follow linear as well as highly non-linear associations. Compared to flexible "black box" machine learning algorithms, GAM offers much higher interpretability. In our context, the response variable *Return* is dichotomous as such *Return* for the i_{th} member can be modeled as a Bernoulli random variable

$$\text{Re}turn_i \sim Bernoulli(p_i), \tag{1}$$

where $p_i \in (0, 1)$ is the expected value of $Return_i$. Given a vector of predictors (X_1, $X_2 \ldots$, X_k), the GAM model is

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 S_1(X_{1i}) + \beta_2 S_2(X_{2i}) + \cdots + \beta_k S_k(X_{ki}) + \varepsilon_i,$$

where $\log\left(\frac{p_i}{1-p_i}\right)$ is the logit transformation of p_i and ε_i is random noise. With its

additive form in the RHS, GAM applies semi-parametric smooth function $S()$ to continuous predictors. We adopt the smoothing spline (James et al. 2013), where $S()$ is derived from minimizing the objective function

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - S(x_i))^2 + \lambda \int \left(S''(t)\right)^2 dt,$$

where the first term is the mean squared error (MSE) of using curve $S(x)$ to predict y. A more wiggly function could lower MSE but has higher risk of overfitting. We prevent overfitting by penalizing the curvature $S''()$ with a non-negative tuning parameter λ in the second term. In general, a higher λ would lead to a smoother curve S. Essentially, the smoothing spline $S()$ has to strike a balance between reducing MSE and increasing curvature penalty.

Using the three sets of variables—member demographics, booking behaviors, and restaurant attributes—introduced in Sect. 3, we test four models with different predictor variables.

$$M1 : Age1 + Age2 + Age3 + Gender;$$
$$M2 : S(Timediff) + S(Timing) + S(PartySize) + DayofWeek;$$
$$M3 : M2 + Status\_canceled + Status\_changed + Status\_ok;$$
$$M4 : M1 + M3 + Is\_hotel + Wifi + Area1 + Area2 + Area3 + Area4.$$

$M1$ uses just member demographics and reduces to the ordinary logistic regression because it only has binary variables. $M2$ introduces smooth splines to three continuous variables related to booking behaviors. $M3$ extends $M2$ by introducing three indicators for possible outcomes on the dining date. Finally, M4 uses all information available by adding *Age*, *Gender*, and restaurant attributes into M3. We fit the four models using the *gam* function in *R* and its built-in backfitting algorithm (Hastie and Tibshirani 2016). We also estimate those models via restricted maximum likelihood using *mgcv* package in *R* and obtain nearly identical results. Hence, we report only the findings from the *gam* routine. Note that our GAM model has three continuous predictor variables (i.e., *Timediff*, *Timing*, *PartySize*) and includes more than a dozen of categorical variables. That said, the ordinary logistic GLM (generalized linear model) is already capable of processing such a mix of categorical and continuous variables (Ranganathan et al. 2017) or even exclusively categorical variables (SAS software 2018). The GAM, nevertheless, is a generic extension of the logistic GLM and has been developed to capture non-linear relationships between predictor and response variables. GAM applies non-parametric smoothing to continuous predictors and simply treats categorical predictors the same as GLM does. The validity of $M1$–$M4$ has also been checked through triangulating estimates from different estimation approaches (i.e., backfitting algorithm, restricted maximum likelihood).

We assess the prediction performance of different model specifications using the receiver operating characteristics (ROC) curve (Fawcett 2006) for binary classification problems. Given a predicted value and an observed outcome, there are four possible outcomes: true positive (TP), false positive (FP), true negative (TN), and

false negative (FN). Accordingly, we can compute true positive rate (TPR, TP/(TP + FN)) and false positive rate (FPR, FP/(FP + TN)) (Fawcett 2006). An ROC curve is constructed by plotting TPR on the *Y* axis and FPR on the *X* axis. The ROC curve illustrates tradeoffs between TPR and FPR, since each point on the curve represents a combination of TPR and FPR under a particular threshold value in [0, 1] to make a positive prediction (i.e., return = 1). Given the ROC curve, the key performance metric of interest is area under curve (AUC), which is bounded in [0, 1]. A perfect classifier would have AUC = 1, while random guesses should result in AUC = 0.5. Models with adequate predictive power are expected to achieve AUC significantly higher than 0.5. The AUC is a fairly standard metric for evaluating prediction performance of different classification models.

Figure 2 shows prediction performance (in terms of AUC) for *M*1–*M*4. Even though *Age* and *Gender* are traditionally considered useful predictors of consumer behaviors, *M*1 with the two member demographics perform poorly and its AUC 0.531 is only 7.7% higher than the one of a randomly guessing model (AUC = 0.5). Interestingly, user demographics like gender and age, which are commonly seen in customer behavior studies, are not effective predictors. The finding from *M*1 also echoes to the afore-mentioned anecdotes by Netflix (Morris 2016). *M*2 achieves AUC $\sim$ 0.6 by instead using three timing features and *PartySize*. *M*3 improves its AUC to 0.669 by including three *Status* variables on the dining date. However, *M*4 that uses all available predictor variables listed in Table 1 only increase the AUC of *M*3 by 0.004. The results indicate that booking behaviors (i.e., reservation timing, party size, dining appearance) are much more useful than member demographics and restaurant attributes in predicting customer retention to reuse online reservation services in our context. Note that most of the weak predictors in regarding demographics and restaurants are statistically significant. Apparently, the statistical significance is a product of our large sample size and it is not synonymous with predictive power. Overall, the prediction performance of the GAM is acceptable (-maximum accuracy $\sim$ 0.8) but not entirely satisfactory. Hence, in the next session,

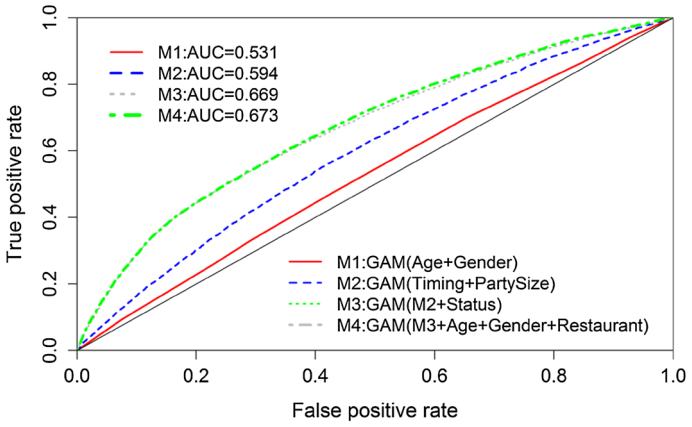

Fig. 2 GAM prediction performances

we apply more sophisticated 'black box' learning algorithms to see whether we can further improve prediction performance.

5 GAM versus ensemble learning

In this section, we try to outperform the best-performing GAM *M*4 (AUC = 0.673) and tackle the prediction problem using several *ensemble learning* techniques, which obtain strong predictive power by creating ensembles of training observations. After that outcomes from multiple ensembles are aggregated to generate predictions based on the spirit of using collective wisdom. Those ensemble techniques, when combined with out-of-sample validation procedures, not only prevent overfitting and achieve substantially lower variance of prediction error.

Specifically, we test five powerful ensemble learning methods. The first is *Bagging* that generates B (B is an integer $\gg 1$) bootstrapped samples to train B models, respectively. Each of the B models is typically a regression/classification tree, and the B models are then aggregated to generate predictions. The second—*Random Forest*—improves upon *Bagging* by de-correlating the B models/trees, in which only a subset of all available predictor variables is used to train a model for each sample. We implement *Bagging* using the *bagging* function in *R*, and implement *Random Forest* using the *randomForest* function in *R*. For both algorithms, we set $B = 500$ to ensure that sufficient ensembles are created for training.

The last three methods are grounded on *Boosting*, a class of sequential learning algorithms. That is, unlike *Bagging* and *Random Forest* where learning outcomes from each of the B models are independent, *Boosting* carries dependencies over all rounds of learning processes such that error can be reduced from round to round. For our classification problem (i.e., return or not), we test two classical algorithms—*Stochastic Gradient Boosting* and *AdaBoost* (Kuhn and Johnson 2013). We implement both algorithms using the *gbm* function in *R* with 1500 ensembles, respectively. Finally, we test the state-of-the-art *XGBoost* (extreme gradient boosting) tree boosting system developed by Chen and Guestrin (2016). The *XGBoost* outperforms other gradient boosting in *gbm* by searching over all the possible splits for a full tree according to percentiles of variable distribution. In many occasions and contests, the *XGBoost* with extraordinary computational efficiency has been shown to outperform many other statistical learning techniques, including support vector machines and neural network that have much higher computational costs.

Table 2 shows the prediction performance of GAM (*M*3 and *M*4) versus the five tree-based ensemble learning techniques discussed above. The first column reports AUC values using only variables on booking behaviors, and the second column reports AUC values using all available information in Table 1. Three important observations are made from this modeling exercise. First, all of the five computationally intensive learning techniques are surprisingly outperformed by the GAM. This finding sheds light on the practical utility of the seemingly simplistic yet flexible GAM for large-sample predictive modeling. Second, the rudimentary

Table 2 GAM versus ensemble learning

Model: *M*3 (booking behavior)	AUC	Model: *M*4 (all variables)	AUC
GAM	0.669	GAM	0.673
Bagging	0.614	Bagging	0.614
Random forest	0.669	Random forest	0.651
Stochastic gradient boosting	0.651	Stochastic gradient boosting	0.653
AdaBoost	0.651	AdaBoost	0.653
XGBoost	0.665	XGBoost	0.672

ensemble learning method, *Bagging*, leads to the worst performance among all, and its advanced variant, *Random Forest*, performs similar to *Stochastic Gradient Boosting* and *AdaBoost*. Interestingly, when all features are available for randomly drawn to impute a tree (i.e., *M*4), *Random Forest* performs worse (AUC = 0.651). Its performance improves when the features for random draws are limited (i.e., *M*3 and AUC = 0.669). This is a direct result of allowing non-relevant predictors to be in the set of candidate features for training. Third, in our case, only the most sophisticated *XGBoost* performs nearly identical to the GAM, corroborating the rapidly growing popularity *XGBoost* in practice. That said, after carefully fine-tuning the learning rate and iteration parameters of *XGBosst* for multiple times, we still could not obtain significant improvement in AUC or make its prediction performance superior to GAM. Finally, on top of ensemble learning, other two sophisticated algorithms—support vector machines and deep learning (with multilayer neural networks)—are also tested in initial phases of our study. However, the two classification algorithms converge slowly and perform poorly in our context with numerous categorical features. We will reflect on this somewhat unexpected result of GAM versus ensemble learning in the next section.

6 Discussion and conclusions

In this paper, we reported a predictive investigation of first-time customer retention in emerging service business—online reservation services. After analyzing more than 100,000 booking records from the focal company, we found that features about booking behaviors are more useful than user demographics and restaurant attributes in predicting first-time customer retention. Even though our empirical tests are based on datasets from online reservation services in the dining industry, the predictive modeling procedures—i.e., the GAM and input features—shown in our paper are potentially applicable to other service sectors. Specifically, nearly all our predictor variables pertaining to member demographics (age, gender), booking behaviors (timing, dayofweek, timediff, status, partysize), and geographical attributes (city area) are also relevant in hotel, travel, and rental industries, where customers' booking/reservation has numerous information fields identical to dinning

reservations. Therefore, while we cannot extrapolate our results according to a single study, we believe that our findings from the dining industry have certain generalizability and could be useful for other service industries that have to predict first-time customer retention.

Moreover, predicting the retention of first-time customers with only one past observation is by no means an easy problem for service business. In our case of online reservation services, we applied various contemporary learning methods, but found that there is not a silver bullet to the business problem. Instead of model complexity, variable selection is more critical to improving prediction performance. In this study, most useful predictors come from our creation of variables on booking behaviors such as timing labels and reservation status. Those are context-specific features instead of ordinary demographics (e.g., age, gender) widely used for database marketing in service industries (Swart and Roodt 2015). Our findings remind business researchers the importance of using domain know-how to identify relevant predictors. Without extracting meaningful features, big data and complex algorithms would not be sufficient for attaining satisfactory prediction outcomes.

Accordingly, we urge the online service provider to allocate more data collection efforts to members' booking behaviors in order to enhance performance of predicting first-time customer retention. Especially, data on booking behaviors are easy to obtain during transaction processes, since those data are required inputs to make online reservations. For instance, customers have to provide party size information and information about the booking time/time till dining date cannot go wrong. In contrast, demographic variables like age and gender are easily suffered from missing and mis-specified errors. Since the data quality of booking behaviors is assured, our findings suggest a good opportunity for the service provider take more initiatives on tracking booking behaviors, e.g., Is a reservation made under promotion? What is the purpose of dining? The extra data can be utilized to improve customer retention predictions. As the focal company relies on transaction fees from restaurants for every reservation made by customers, identifying those who would continuously use its services not only is critical to its revenues, but also could reduce ineffective marketing costs.

In addition to contextual findings described above, on the methodological front, we conducted a comparative study where surprisingly, the prediction performance of GAM outperforms ensemble learning methods, even the state-of-the-art *XGBoost*. In contrast to computationally intensive methods like random forest and boosting, the old-fashioned GAM has been under-utilized in data analytics. Nevertheless, while most methods are exclusively either in the camp of interpretability or flexibility, GAM strikes a nice balance between the interpretable, yet biased, linear model, and the extremely flexible, "black box" machine learning algorithms (Larsen 2015). Given the balanced nature of GAM and the findings of our study, we recommend researchers to include GAM in their arsenal when facing applied predictions problems.

Last, we would to echo the concept that predictive modeling is different from explanatory modeling (Shmueli 2010). Explanatory modeling is used to evaluate the explanatory power of underlying theory-driven models. In contrast to explaining phenomena from theory, predictive modeling is aimed at predicting new

observations at the empirical level. Shmueli and Koppius (2011) articulate the utility of predictive modeling in terms of improving existing theories, comparing competing theories, and evaluating practical relevance of theoretical models. For instance, in explanatory modeling, t test is commonly used to evaluate statistical significance of relationships among constructs. However, it is easy to achieve large t $(\hat{\beta}/(SE_\beta/\sqrt{n}))$ values and reject H_0: $\beta = 0$ when sample size (n) becomes big (Lin et al. 2013). Take our case ($n > 100{,}000$) for example, almost all predictors (including weak predictors) in our models are significant. Given that statistical significance can be an artifact of large samples, substantial significance of research models/variables should be validated by out-of-sample prediction. Predictive modeling can complement explanatory modeling by serving as a stringent test of practical usefulness.

A major limitation of our study is that the prediction performance is not truly outstanding, primarily due to that first-time customers (compared to repeated customers) have much less information available for predictive modeling. Given the limited features that could be extracted from booking records for model training, sophisticated statistical learning methods could not overcome this limitation in our case. Even though the reservation service provider possesses a large number of member observations, to make this seemingly "big" (in terms of sample size but not the number of features) data more useful, the online service provider has to expand its data collection effort. In a nutshell, the prediction performance reported in our study leaves room for improvement. We strongly encourage subsequent studies to investigate customer behaviors (e.g., conversion, retention) in this kind of emerging service business, and develop creative features or models in order to improve prediction accuracy that is indispensable to firm performance.

References

Bapna R, Chang SA, Goes P, Gupta A (2009) Overlapping online auctions: empirical characterization of bidder strategies and auction prices. MIS Q 33(4):763–783

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the twenty-secondth ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, CA, pp 785–794

Coussement K, Van den Poel D (2008) Churn prediction in subscription services: an application of support vector machines while comparing two parameter selection techniques. Expert Syst Appl 34(1):313–327

Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874

Hastie T, Tibshirani R (1990) Generalized additive models. Chapman and Hall/CRC, Boca Raton

Hastie T, Tibshirani R (2016) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York

Hong SJ, Tam KY (2006) Understanding the adoption of multipurpose information appliances: the case of mobile data services. Inform Syst Res 17(2):162–179

Hong W, Thong YLJ, Tam KY (2004) Designing product listing pages on e-commerce website: an examination of presentation mode and information format. Int J Hum-Comput 61(4):481–503

Hosseini SY, Bideh AZ (2014) A data mining approach for segmentation-based importance performance analysis (SOM-BPNN-IPA): a new framework for developing customer retention strategies. Serv Bus 8(2):295–312

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning with applications in R. Springer, New York
Kim G, Chae BK, Olson DL (2013) A support vector machine (SVM) approach to imbalanced datasets of customer response: comparison with other customer response models. Serv Bus 7(1):167–182
Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, New York
Larsen K (2015) GAM: the predictive modeling silver bullet. http://multithreaded.stitchfix.com/blog/2015/07/30/gam/. Accessed 9 Feb 2018
Lee S, Shin B, Lee HG (2009) Understanding post-adoption usage of mobile data services: the role of supplier-side variables. J Assoc Inf Syst 10(12):860–888
Lin M, Lucas HC, Shmueli G (2013) Too big to fail: large samples and the p-value problem. Inform Syst Res 24(4):906–917
Migueis VL, Camanho AS, Borges J (2017) Predicting direct marketing response in banking: comparison of class imbalance methods. Serv Bus 11(4):831–849
Morris DZ (2016) Netflix says geography, age, and gender are "garbage" for predicting taste. http://fortune.com/2016/03/27/netflix-predicts-taste/. Accessed 9 Feb 2018
Morrisonn AM, Jing S, O'Leary JT, Cai LA (2001) Predicting usage of the Internet for travel bookings: an exploratory study. Inform Technol Tour 4(1):15–30
Olson DL (2007) Data mining in business services. Serv Bus 1(3):181–193
Ranganathan P, Pramesh CS, Aggarwal R (2017) Common pitfalls in statistical analysis: logistics regression. Perspect Clin Res 8(3):148–151
SAS software (2018) The logistic procedure: example 53.2 logistic modeling with categorical predictors. https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_logistic_sect060.htm. Accessed 28 Mar 2018
Schmutz P, Roth SP, Seckler M, Opwis K (2010) Designing product listing pages—effects on sales and users' cognitive workload. Int J Hum-Comput Stud 68(7):423–431
Shankar V, Smith AK, Rangaswamy A (2003) Customer satisfaction and loyalty in online and offline environments. Int J Res Mark 20(2):153–175
Shmueli G (2010) To explain or to predict? Stat Sci 25(3):289–310
Shmueli G, Koppius OR (2011) Predictive analytics in information systems research. MIS Q 35(3):553–572
Swart MP, Roodt G (2015) Market segmentation variables as moderators in the prediction of business tourist retention. Serv Bus 9(3):491–513
Vakrat Y, Seidmann A (2000) Implications of the bidders' arrival process on the design of online auctions. In: Proceedings of the thirty-third annual Hawaii international conference on system sciences, Maui, Hawaii
Van den Poel D, Buckinx W (2005) Predicting online-purchasing behaviour. Euro J Oper Res 166(2):557–575
Van den Poel D, Lariviere B (2004) Customer attrition analysis for financial services using proportional hazard models. Euro J Oper Res 157(1):196–217
Xie Y, Li X, Ngai EWT, Ying W (2009) Customer churn prediction using improved balanced random forests. Expert Syst Appl 36(3):5445–5449