A Study on a Predictive Model of Customer Defection in a Hotel Reservation Website

Shaoyong Han*

Business School, University of Shanghai for Science and Technology, Shanghai200093, China

Abstract. This paper examines a hotel reservation website's customer defection. Applying statistic and data mining technology including logistic regression and random forests, we examine customer database to identify the attributes that affect customer attrition and develop a model of customer defection in the hotel reservation website. The empirical evaluation results showed the model has 78.9% accuracy, which suggest that the proposed churn prediction technique exhibits satisfactory predictive effectiveness.

1 Introduction

As markets mature and competitive pressure intensifies, companies can no longer ignore the importance of customer retention as their existing customer bases have become their precious assets. [1 – 3]

This is particularly true in the tourism industry; many companies start the hotel reservation business, website of booking, Airbnb, etc. This article aims to help hotel reservation website address the churn issue by providing them with insights into the effects of customer demographic and behaviour characteristics on churn rate through the use of the well-known statistical procedure called Logistic regression and Random forests.

2 Literature review

Many enterprises use data mining technology to analyze customer churn. Data mining refers to the process of extracting previously unknown, effective and controllable patterns or knowledge from a large database in order to establish a business decision support system. [4] Based on the types of knowledge, which can be found in the database, data mining technology can be divided into several categories, including classification, clustering, dependency analysis, data view and background mining.

To solve the problem of churn, some scholars abroad have established prediction models with Artificial Neural Network (ANN), genetic algorithm (GA), Decision Tree (DT), support vector machine (SVM) and so on, and a certain application effect has been obtained. [5-9]

Logistic regression analysis is a classification method in data mining technology. As a computer-based identification technique based on statistical theory, it has many advantages. The output of a logistic regression is more informative than other classification algorithms. Like any regression approach, it expresses the relationship between an outcome variable (label) and each of its predictors (features). The logistic regression not only gives a measure of how relevant a predictor is (coefficient size) but also its direction of association (positive or negative).

Random forest is an ensemble method in which a classifier is constructed by combining several different Independent base classifiers. The independence is theoretically enforced by training each base classifier on a training set sampled with replacement from the original training set. This technique is known as bagging, or bootstrap aggregation. In Random Forest, further randomness is introduced by identifying the best split feature from a random subset of available features. The ensemble classifier then aggregates the individual predictions to combine into a final prediction, based on a majority voting on the individual predictions. It can be shown that an ensemble of independent classifiers, each with an error rate e, when combined significantly reduces the error rate.

3 Research methodology and data selection

We used a hotel reservation website as the research object, and obtained 12 months' data from its central database. The initial research goal was established: what kind of customers should be regarded as the churn customer? it was determined that customers whose total quantity of orders in the last three months decreased by 80% compared to that in the first three months were considered to be the churn customers. In this article, users specifically refer to group customers.

We decided to use supervised machine learning to predict the list of customer churn. About model selection,

*Corresponding author: hanshaoyong@163.com we used two methods for modelling-random forests and logistic regression. Based on the model of Confusion Matrix, Precision Rate, Recall Rate, and visualized ROC map, the two modelling methods are compared. Finally, the model was explained.

Figure 1 presents a sample of decision tree model. Now there are a customer data, which contains 14 samples, then the decision tree will branch and classify the samples by the most important feature values. Among all the feature values, form of settlement is the most important variable, so it is taken as the first branch. For the group customer's payment method advance and present settlement, order quantity in the first three months and the number of membership cards are the most important variables of these two branches respectively. The decision tree creates a rule that clearly shows the relationship between each branch and node.

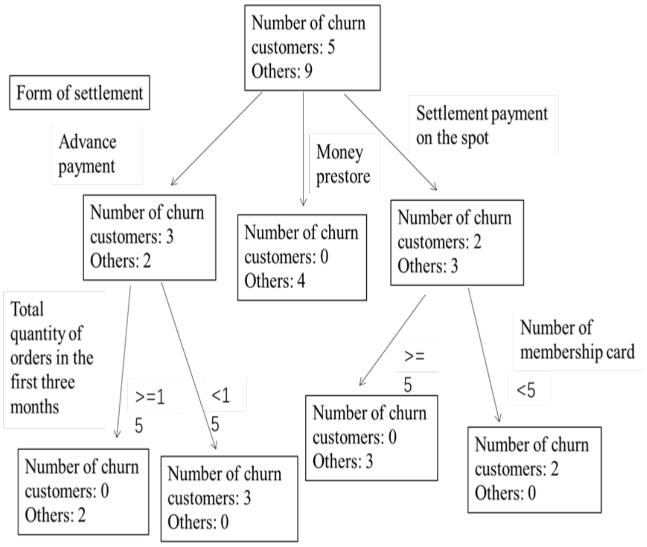

Number of churn customers: 5 Others: 9
Form of settlement
Advance payment
Money prestore
Settlement payment on the spot
Number of churn customers: 3 Others: 2
Number of churn customers: 0 Others: 4
Number of churn customers: 2 Others: 3
Total quantity of orders in the first three months
>=1 5
<1 5
>= 5
Number of membership card
<5
Number of churn customers: 0 Others: 2
Number of churn customers: 3 Others: 0
Number of churn customers: 0 Others: 3
Number of churn customers: 2 Others: 0

Figure 1. A sample of decision tree model.

Figure 2 presents a method of building a model of customer churn.

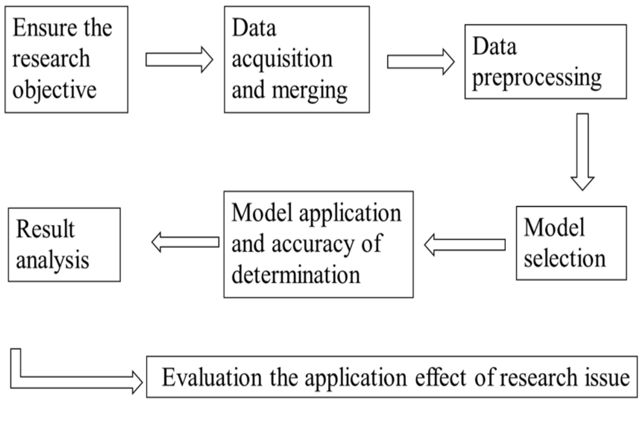

Ensure the research objective
Data acquisition and merging
Data preprocessing
Model selection
Model application and accuracy of determination
Result analysis
Evaluation the application effect of research issue

Figure 2. A method of building a model of customer churn.

Data description. The data for modelling can be extracted from various operating systems. There are several basic elements of modelling: the sample size is as large as possible; each variable chosen fits the actual situation of the business as closely as possible; the goal of modelling is clear, and this will directly affect the final model selection.

According to the current international literature research, we believe that examining whether the volume of transactions between customers and organizations is reduced helps predict customer churn; [10] Changes in customer behaviour can affect the probability of customer churn, and changes in address can indicate potential the loss of customers; [11] The discounts obtained by customers and the channel used by customers are also associated with customer churn; [12] Customer complaints have a positive impact on customer churn; [13] The length of service time and reduction of account balance are also two variables that can help predict customer churn.

Data acquisition and merging. We determined the variables that affect the customer churn model, Table 1 presents the argument Lists. We extracted 14 feasible variables from the database as the standard of our modelling. Considering specific business conditions, data collected from January 2015 to January 2016 are available. The data includes all data for each company ID per month, as well as customers who use self-service platform to place orders. The limit range of data is the first 3 months (excluding the current month) and the order quantity is more than five.

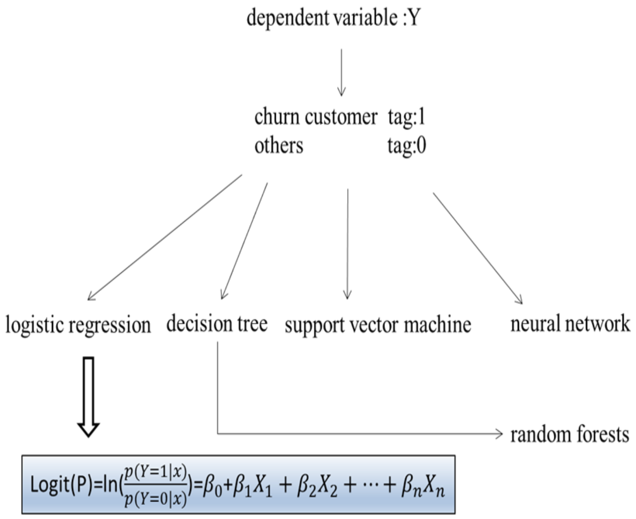
Table 1. Argument Lists.

Classification	Serial number	Index
Customer background information	1	Form of settlement
	2	Registration duration until the current month(time unit: month）
	3	Industry
	4	Number of calls until the current month
	5	City
	6	Number of membership card in the current month
	7	Registration method
Consumption data	8	Average number of rooms reserved in recent 3 months
	9	Average star of rooms reserved in recent 3 months
	10	Average price of rooms reserved in recent 3 months
	11	Order number in website in recent 3 months/total order number in recent 3 months
	12	Order number in mobile app in recent 3 months/total order number in recent 3 months
Service feeling	13	Number of complaints in recent 3 months
	14	Number of complaints in recent 3 months/order number in recent 3 months

Data preprocessing. First, through information entropy and anova analysis of variance, the information entropy after visualization can clearly reflect the importance of each variable for Response Variable. The most important variables are the company's settlement method and the order amount in the first three months. The degree of importance of the variable will also be reflected later in the modeling. In statistics, when the information value of the variable is less than 0.1, it means that this variable is negligible for the entire data and can therefore be removed.

Secondly, through the Pearson correlation, we can get the correlation between each variable and other variables. Therefore, by calculating the Pearson correlation, We can also filter out some variables that are relatively from 14 variables and the variables whose collinearity value VIF>2.

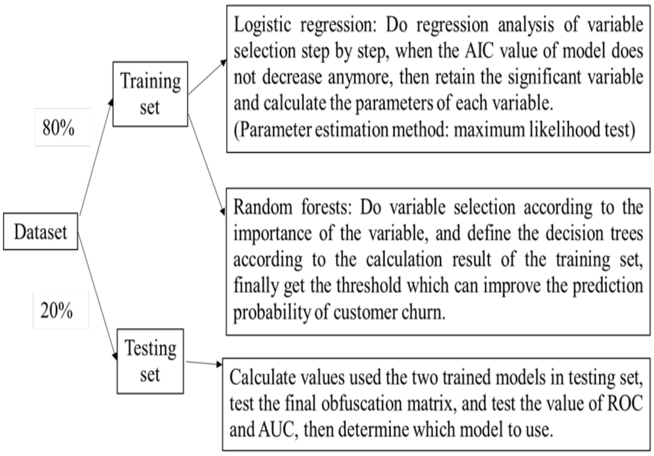
Model selection. The models we used are logistic regression and random forests. Extract 80% of the data as a training set and 20% of the data as a testing set, and use the 10-fold cross-validation verification method to ensure the feasibility of the model. The two models are cross-checked at the end to verify the accuracy of the data predicted using the two models. Figure 3 presents the model selection strategy.



$$\text{Logit(P)}=\ln\left(\frac{p(Y=1|x)}{p(Y=0|x)}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n$$

Figure 3. Model selection strategy.

Model establishment and case analysis. Figure 4 presents the model establishment process.



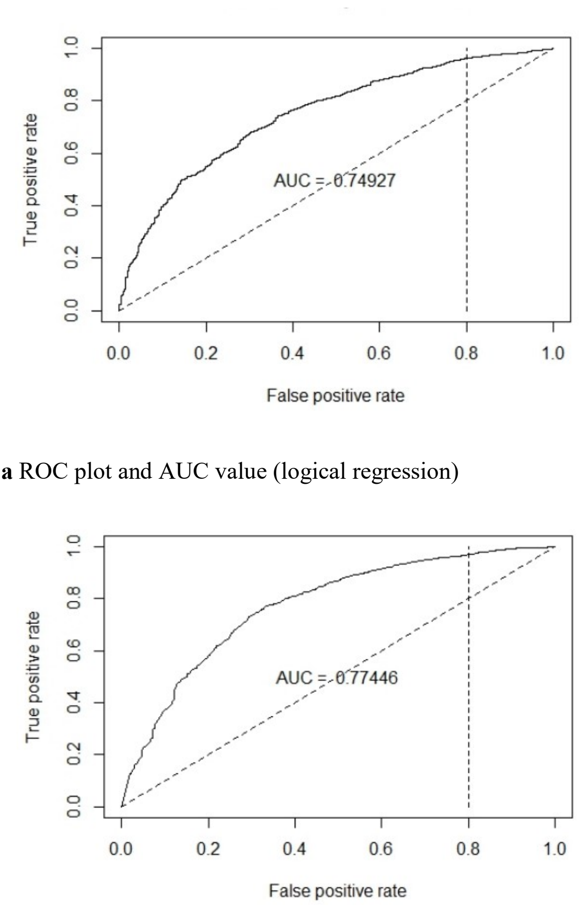
Logistic regression: Do regression analysis of variable selection step by step, when the AIC value of model does not decrease anymore, then retain the significant variable and calculate the parameters of each variable. (Parameter estimation method: maximum likelihood test)

Random forests: Do variable selection according to the importance of the variable, and define the decision trees according to the calculation result of the training set, finally get the threshold which can improve the prediction probability of customer churn.

Calculate values used the two trained models in testing set, test the final obfuscation matrix, and test the value of ROC and AUC, then determine which model to use.

Figure 4. The process of Model building.

Figure 5 presents the simulation results (Using Logistic Regression Model and Random Forest Model).



5a ROC plot and AUC value (logical regression)

5b ROC plot and AUC value (random forests)

Figure 5. The ROC plot and AUC values of logical regression and random forests.

The AUC value calculated by the logistic regression model is smaller than the value calculated by random forests, therefore, random forests model is selected. Table 2 presents the experimental data calculated by random forests.

Table 2. Random forests experimental data.

Random forests training set	Number of churn customers	Number of retention customers
Number of churn customers	384	223
Number of retention customers	655	2496

4 Results

The model was evaluated and the conclusions are as follows:

1. The accuracy of the model is 78.9%.

2. Without changing the threshold of the model, the accuracy of predicting the customers churn is 37%. By changing the threshold value, the accuracy of predicting the customers churn can be improved.

3. Calculate the score of each company through the model, the higher the score is, the smaller the probability

of customer churn is. Similarly, the lower the score, the greater the churn probability of customers.

5 Conclusion

Faced with high customer acquisition costs and a fast maturing market, hotel reservation websites are urged to shift their marketing focus from customer acquisition to retention activities.

Through the use of random forests, this article presents a model that marketers can use to identify customer segments that are sensitive to churning behavior.

6 Limitations and future research

Although this study contributes to the knowledge on customer retention in tourism industry about hotel reservation, some limitations and opportunities for further research deserve to be mentioned. First of all, this research only analyses data provided by a hotel reservation website in china, so there is a geographical bias. Second, this research focuses on the group customers. Future studies can explore individual customers and develop explanatory variables that are applicable to such customer segment.

In summary, the findings of this article may help researchers and practitioners in the tourism industry about hotel reservation adopt a holistic perspective and make informed decisions about the opportunities and challenges that churn management would present.

Acknowledgements

This work is supported by University of Shanghai for Science and Technology (Grant No. 18HJ-DSDG-ZD-004).

References

1. Athanassopoulos A.D.. Customer satisfaction cues to support market segmentation and explain switching behaviour. Journal of Business Research, 2000, vol.47, pp.191-207.
2. Jones M., Mothersbaugh D. and Betty S.. Switching barriers and repurchase intentions in services. Journal of Retailing, 2000, 76 (2): 257-272.
3. Thomas, J.S. A methodology for linking customer acquisition to customer retention. Journal of Marketing Research, 2001, 38 (2): 262-268.
4. Berry M J A, Linoff G. Data mining techniques: for marketing, sales, and customer support [M]. New York: Wiley, 1997.
5. Mozer M.C., Wolniewicz R., Grimes D.B., et al. Churn Reduction in the Wireless Industry [J]. Advances in Neural Information Processing Systems, 2000, (12): 935-941.
6. Lemmens A., Croux C.. Bagging and Boosting Classification Trees to Predict Churn [J]. Journal of Marketing Research, 2006, 43 (2): 276-286.
7. Chiang D., Wang Y., Lee S., Lin C.. Goal-Oriented Sequential Pattern for Network Banking Churn Analysis [J]. Expert Systems with Applications, 2003, 25 (3): 293-302.
8. Eiben A. E., Koudijs A. E., Slisser F. Genetic Modeling of Customer Retention [C]. Lecture Notes in Computer Science, 178-186, 1998.
9. Zhao Y., Li B., Li X., Liu W., Ren S. J. Customer Churn Prediction Using Improved One-Class Support Vector Machine [C]. Lecture Notes in Computer Science, 2005, (3584): 300-306.
10. Zeithaml V, Berry L, Parasuraman A. The behavioral consequences of service quality [J]. Journal of Marketing, 1996, 60 (4): 31-46.
11. Hamilton R.. How croft B. A practical approach to maximizing customer retention in the credit card industry [J]. Journal of Marketing Management, 1995, 11: 151-163.
12. Ainslie A, Pitt L. Customer retention analyses [J]. Journal of Direct Marketing, 1992, 6 (3): 31-43.
13. For nell C, Wernerfelt B. Defensive marketing strategy by customer complaint management: a theoretical analysis [J]. Journal of Marketing Research, 1987, 24 (11): 37-46.