Accounting and Corporate Management (2020) 2: 67-70
Clausius Scientific Press, Canada
DOI: 10.23977/acccm.2020.020106
ISSN 2523-5788

Discriminant analysis text to predict customer loss in the real estate industry

Yu Zhao[1,a], Tong Bai[1,b], and Yingnan Jia[3,c]

[1]*Institute of Date Science, City University of Macao, Avenida Padre Tomás Pereira, Macao, China*
[2]*Institute of Business Adminstration, Krirk University, Ram Inthra Rd, Bangkok, Tailand*
[3]*School of Economic and Management, SHIHEZI University, Shibeisi Rd, Xinjiang, China*
[a] *D19091105079@cityu.mo,* [b] *91666@88.com,* [c] *1445006464@qq.com*

Keywords: Linear Discriminant Analysis, Customer Loss, Real Estate Industry

Abstract: Customer churn in a broad sense means that customer service is terminated because the behavior of the customer or real estate operator is contrary to the service agreement. In fact, in real life, the main reason for the loss of customers is because customers are not satisfied with the real estate operator's service attitude and manner, or other real estate operators give more favorable prices. Customer churn forecasting is the process of using historical data recorded by customers to identify potential churning customers. He is a very important concern for the various service industries. Especially in the highly competitive financial, telecommunications, real estate, and other industries.

Customer loss is one of the most important measures of whether the real estate industry can develop healthily. According to the figures, in recent years, the customer turnover rate in all walks of life is in a relatively high state, the average monthly customer turnover rate of about 2.2%. Therefore, maintaining customers and reducing customer churn is an immediate problem.

1. Research Purposes

A large number of customer churn is now the general status quo in the service industry. Customer churn increases not only the cost of sales but also the opportunity cost. At the same time, it will significantly reduce the number of new customers introduced by old customers, and it will also reduce the number of new customers attracted. According to statistics, the cost of developing a new customer is about 5-6 times the cost of maintaining an old customer. Therefore, we should reasonably maintain old customers and reduce customer churn.

The purpose of this paper is to determine the customer base for the implementation of the strategy and make reasonable predictions about the potential loss of customers. Through the establishment of a judgment analysis model, analysis of customer churn, and give reasonable maintenance advice.

2. Principle of Linear Discriminant Analysis

2.1 The idea of linear discriminant analysis

Linear Discriminant Analysis, or LDA for short, is a classic dimensional reduction method. LDA is supervised learning, a classic dimensional reduction technique, which means that each sample of its dataset has a category output. Its main characteristics are the projection of the class within the smallest variance, the largest variance between classes. As the following illustration shows, we project the dataset, and the desired result is that the projection points of each category of data are as close as possible, and the projection points of different categories of data are as far away as possible.

Simply put, let's put it this way: Let's say there are now two types of data, both two-dimensional, and our goal is to project this data onto a one-dimensional line so that the projection points of each category of data are as close as possible, that is, the projection points of data of the same color are as close as possible, and the projection points of different types of data are as far away as possible, that is, the distance between the red and blue data centers.

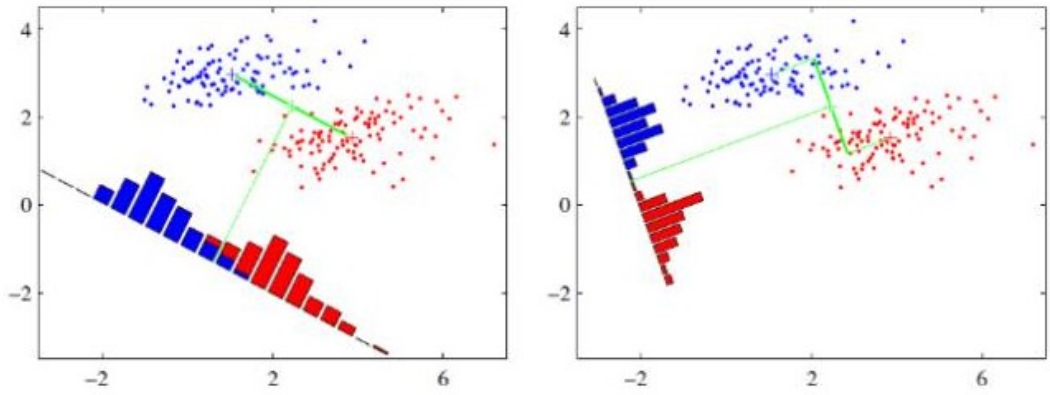

Figure 1: Linear Discriminant Analysis Projection Point Graph.

According to the image above, the figure on the right is much better than the projection effect on the left, first of all, the data of the same color on the right is more aggregated, and second, the distance between the data of different colors is significantly larger.

2.2 Principle and process of linear discriminant analysis

Given the data set: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$

Collection of examples of type I: X_i

The mean vector of the i-th example: $\boldsymbol{\mu}_i$

Covariance matrix of the i-th example: $\boldsymbol{\Sigma}_i$

Projection of the centers of two types of samples on a straight line: $\boldsymbol{w}^{\mathrm{T}}\mu_0$ $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1$

Covariance of two types of samples: $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w}$ and $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}$

The goal of LDA: to maximize the generalized Rayleigh quotient $J = \dfrac{\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}}$

(W multiplied zoom does not affect the J value, only the direction is considered)

Let maximize the generalized Rayleigh quotient equivalent form as $\min_{\boldsymbol{w}} \; -\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}$ $\text{s.t. } \boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w} = 1$

Using the Lagrange multiplier method, there are $\mathbf{S}_b\boldsymbol{w} = \lambda\mathbf{S}_w\boldsymbol{w}$

The direction is always might $\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ as $\mathbf{S}_b\boldsymbol{w} = \lambda(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$ well make so $\boldsymbol{w} = \mathbf{S}_w^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$

Singular value $\mathbf{S}_w = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$ decomposition

Then $\mathbf{S}_w^{-1} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}$

Suppose there are N classes

Global divergence matrix: $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w = \sum_{i=1}^{m}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$

Within-class divergenc matrix: $\mathbf{S}_w = \sum_{i=1}^{N}\mathbf{S}_{w_i} \quad \mathbf{S}_{w_i} = \sum_{\boldsymbol{x}\in X_i}(\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^T$

Inter-class divergence matrix: $\mathbf{S}_b = \mathbf{S}_t - \mathbf{S}_w = \sum_{i=1}^{N} m_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$

3. Experimental results

3.1 Feature preprocessing and selection

Convert the Boolean value in the data to 0 and 1. Draw the age histogram, we can see that the distribution is roughly normal, and the missing values are imputed after age segmentation. Perform feature extraction on the field to achieve the purpose of dimensionality reduction. Finally, after processing and integration, it is found that there are 44 standardized features.

3.2 Classification model selection and training

Data set division: use K-fold cross-validation, train_test_split autonomously split the data set

Model selection: using decision tree, boosting tree (GBDT/XGBoost), SVM (libsvm) neural network (multi-layer perceptron algorithm, LDA algorithm to train the model separately, the final result :

Table 1: LDA algorithm separately trained model results.

	LDA	DT	XGBoost	Libsvm	BP
accuracy	0.91	0.86	0.91	0.89	0.9
precision	0.88	0.86	0.88	0.84	0.88
recall	0.92	0.86	0.91	0.89	0.9
F1 score	0.89	0.86	0.89	0.85	0.87

Perform analysis based on the above icons. It can be seen that: LDA algorithm and XGBoost algorithm have the highest accuracy; LDA algorithm, XGBoost algorithm and BP neural network algorithm have the highest accuracy; LDA algorithm has the highest recall rate; LDA algorithm and XGBoost algorithm have the highest F1 index. However, the LDA algorithm is better for all indicators.

4. Conclusion and Suggestion

4.1 Reasons for customer churn

(1) The attitude and ability of real estate services are limited, which is detrimental to the interests of customers.

(2) The lack of innovation in real estate services makes customers choose other companies.
(3) Lack of communication with customers and not establishing close relationships.
(4) The rapid flow of people has taken away some customers.
(5) Natural loss.

4.2 Policy to prevent customer loss

(1) Adopt flexible service methods.
(2) To establish a corresponding motivational physique. Divide customers into: core customers, key customers, general customers, and unimportant customers.
(3) Communicate more with customers and establish a relationship of mutual trust.
(4) Improve customer satisfaction and cultivate customer loyalty.
(5) Strive to achieve honest marketing.

4.3 Advantages of the LDA model

(1) LDA is an unsupervised learning model, in the process of dimensionality reduction, the prior knowledge experience of the category can be used.
(2) The classification of LDA is based on the mean rather than the variance.
(3) LDA can be used for both classification and dimensionality reduction.
(4) LDA uses the method of matrix eigen decomposition in the process of dimensionality reduction
(5) The calculation process of LDA conforms to Gaussian distribution

References

[1] WEICP, TANGCI.Turning telecommunications call details tochurn prediction: a data mining approach[J] . Expert Systems with Applications, 2002, 23(2) : 103- 112.
[2] LOUIS A C. Data mining and causal modeling of customer[J] . Telecomm unication Systems, 2002, 21(2) : 381- 394.
[3] ROSSET S, EINAT N. Integrating customer value considerations intopredictive modeling[C] / /Proc of the 3rd IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society,2003: 283- 290.
[4] CARDELLN S, GOLOVNYA M, STEINBERG D. Churn modeling for mobile telecommunications[EB /OL] . (2003-06) . http: / / www.salford-systems. com/ doc / churnwinF08. pdf.
[5] QI Jia-yin, ZHANG Yang-ming, ZHANG Ying-ying, et al. TreeLogitmodel for customer churn prediction[C] / /Proc of Asia-Pacific Conference on Services Computing. Washington DC: IEEE Computer Society, 2006: 70- 75.
[6] LUO Bin, SHAO Pei-ji, LIU Juan. Customer churn prediction basedon the decision tree in personal handyphone system service[C] / /Procof International Conference on Service Systems and Service Management. 2007: 1- 5.
[7] KIM H S, YOON C H. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market[J] . Telecommunications Policy, 2004, 28(9) : 751- 765.