Click Prediction for P2P Loan Ads Based on Support Vector Machine

Xiaoyao Zhang

School of Information Technology and Management, University of International Business and Economics, Beijing 100020, China
Corresponding author’s e-mail: zhangxiaoyao_uibe@qq.com

Abstract. This paper mainly contains the following three aspects of research. Firstly, this paper starts from the definition of advertisement click prediction, analyzes the distribution and characteristics of the dataset and preprocesses the dataset. On this basis, based on the understanding of advertising and the characteristics in practical application, ten different types of features were extracted. Secondly, the paper first uses SVM, naive Bayesian model, decision tree model and neural network model to predict the click-through rate and analyze their shortcomings. The experimental results show that the SVM model method used in this paper can obtain better prediction results than other methods under the same characteristics.

1. Introduction

The development of the Internet has fundamentally changed the mode of Internet advertising: the traditional form of advertising in which fixed images or texts are embedded in web pages is gradually transformed into a dynamic targeting mechanism based on web content and user characteristics. Computational advertising is a sub-discipline that arises in this demand environment. It is an advertising delivery mechanism that calculates the best matching advertisements based on given users and web content and delivers targeted targeting.

Accurate prediction of ad click-through rate is the basis of ad serving, because whether the ad will be clicked involves the interests of the platform, advertisers and users. Without accurate predictions, advertisers will not receive the corresponding ad conversion rate, the platform will not be able to get the corresponding advertising costs and traffic, and users will not be able to get personalized recommendations for their own.

The previous advertising model was one-way, that is, what companies want consumers to know, not from the consumer's point of view, and what consumers want, partly because of the information asymmetry, companies can't know themselves. What consumers need, data sparseness is a problem that was faced before. But in this era of big data, we have a huge amount of data, how to extract useful user characteristics information from these data, construct user images, predict the user's advertising click rate, and then accurately carry out advertising, to achieve "three wins". These are the issues to be solved in this article.

The ratio of click-to-advertising and non-click-to-advertising in the dataset of this paper is about 1:34.5, which is a serious imbalanced sample. Therefore, this paper will use random sampling undersampling to reduce the proportional difference between positive and negative samples, so that the sample will be balanced and improved. Predict the sensitivity of the customer click-through rate model.
2. SVM model
Support vector machine (SVM) is a statistical learning method. Its core content is proposed from 1992 to 1995 [1-4], and it is still in the stage of continuous development. For a given learning sample with limited data training samples. Task, SVM compromises both accuracy and machine capacity to get the best generalization ability. Given training set \{ (x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l) \}, where \( x_i \in \mathbb{R}^n \), \( y_i \in \{-1, +1\} \). If the input space is linearly separable, the goal is to find a generalized optimal classification hyperplane in the linear separable real space \( w^\top x + b = 0 \). Learning problem minimizing the objective function is

\[
R(w, a) = \frac{1}{2} ||w||^2 + C \sum_{i=1}^l a_i \text{ s.t. } y_i [w \cdot x + b] \geq 1 - a_i
\]

(1)

Where: the penalty factor \( C > 0 \), \( a_i \) is the slack variable. Using the Lagrange multiplier method, the formula (1) can be changed to its dual form.

\[
\max W(\alpha) = \sum_{i=1}^l T_i - \frac{1}{2} \sum_{i,j=1}^l T_i T_j y_i y_j (x_i \cdot x_j)
\]

s.t. \( \sum_{i=1}^l T_i y_i = 0, T_i \in [0, C], i = 1, 2, \ldots, l \),

\[
w = \sum_{i=1}^l T_i y_i x_i
\]

(2)

In the solution to this optimization problem, some \( T_i \) are not 0, and the corresponding training samples are support vectors. For the vector \( x \) of unknown generic class, the following linear decision function can be used for class determination.

\[
 f(x) = \text{sgn}(w \cdot x + b) = \text{sgn}\left(\sum_{i=1}^l T_i y_i (x_i \cdot x) + b\right)
\]

(3)

For nonlinear classification, the nonlinear mapping function \( h \) is first used to map the training data from the original space to a high-dimensional space, and then an optimal classification hyperplane is constructed in the high-dimensional feature space. Although the dimension of the high-dimensional feature space may be very high, but only consider the dot product operation in the high-dimensional feature space in the nonlinear space, so it is not necessary to know the mapping function explicitly, just substitute the kernel function \( K(x_i, x_j) = (h(x_i)h(x_j)) \) Equation (3), the decision function of the nonlinear support vector machine can be obtained.

\[
f(x) = \text{sgn}\left(\sum_{i=1}^l T_i y_i K(x_i, x) + b\right).
\]

(4)

3. Experiment

3.1 data set
We chose Finup as a data source. Finup can collect some third-party data from customers, analyze the customer's credit status, score the customer's credit, and finally give the customer the most appropriate amount of assessment. Its subsidiary money station is mainly engaged in cash loans, with installment loans and rapid small loans. In order to expand the market, Finup needs to place advertisements on specific websites, and also needs to deliver advertisements to specific people.

3.2 Data Processing
For the purposes of the study, we obtained user data from the Finup platform. Based on the exploration and analysis of the original data, the data that is not related to the analysis target or the model needs to be processed is found, and the data is processed. The data processing methods involved are: data cleaning and data transformation. In the original data, there are 28,069 P2P loan advertisements with 812 clicks, accounting for 2.9% of the statistics.

In the next step, in order to solve the collinearity problem between independent variables, the stepwise regression-forward method in logistic regression is first used to incorporate the variables that have significant influence on “whether to click” into the regression equation one by one, avoiding the inter-variable collinearity problem. The logistic regression model incorporates 61 attributes into the regression equation, and 10 of them are more important for interpreting the target variable and are
significantly related to the dependent variable. The variables are shown in Table 1:

| variable code | name                      |
|---------------|---------------------------|
| X1            | male                      |
| X2            | female                    |
| X3            | Weather Query_Active High |
| X4            | life practical            |
| X5            | BannerType                |
| X6            | online loan               |
| X7            | Weather Query_Active      |
| X8            | online loan P2P_active high |
| X9            | takeout_active low        |
| X10           | online shopping           |

Correlation analysis was carried out between the 10 most important variables and the target variable label. It was found that online loans, online loans P2P_active high, gender male, weather query_active were positively correlated with label, and the other six variables were negative correlation. To simplify the model later, modeling uses only the 10 attributes derived from this section.

Based on the previous preprocessing, 10 of the 11 variables are 0/1 variables, and 1 is a discrete variable. Descriptive statistics have little meaning, so they are omitted here.

### 3.3 Modeling and Results

#### 3.3.1 Preliminary modeling

Preliminary modeling refers to the use of the 11 variables selected in 3.2 and the results of data preprocessing to partition the data, 70% of the training set, 30% of the test set, and then use four modeling methods for predictive analysis. The evaluation criteria include accuracy and sensitivity. The accuracy is defined as the ratio of the total number of 0 and 1 correctly recognized by the model to the total data set. The sensitivity is the ratio of the number of 1 correctly recognized by the model to the number of the original data set. Table 2 compares the four models for the 30% test set results (excluding the default values for model predictions).

| model         | Sensitivity | Accuracy  |
|---------------|-------------|-----------|
| Decision tree | 0%          | 97.24%    |
| Neural network| 0%          | 97.23%    |
| Bayesian network| 0%       | 97.23%    |
| SVM           | 0%          | 97.19%    |

The reason why the accuracy of the four models is above 97%, but the sensitivity is 0, the number of labels in the sample set is only 2/71 of the data set, and the ratio of positive and negative samples is 1:34.5. It is an extremely unbalanced sample. To improve the sensitivity of the model, it is necessary to undersample the sample set to reduce the proportion of negative samples.

#### 3.3.2 Random sampling - modeling after undersampling

This article defines the meaning of the sampling degree \( a\% \) as follows:

\[
a\% = \frac{N - N_s}{N - P}
\] (5)
Where: N represents the number of negative samples in the original data set, which is 27118; P represents the number of positive samples in the original data set 785; Ns represents the number of negative samples after sampling. Modeling analysis of a=10, 30, 50, 70, 80, 85, 90, 100 respectively (partition processing is the same as preliminary modeling). The result shows that there is an optimal ratio between positive and negative samples between 1/4 and 1, so that the predicted sensitivity is the highest. If the optimal ratio is exceeded, the sensitivity will decrease.

After preliminary modeling and comparison of under-sampling modeling, we conclude that when under-sampling makes the positive-negative sample reach the optimal ratio range, the sensitivity of the model will be greatly improved. Of course, this process will sacrifice the accuracy of partial accuracy. It can be obtained from Figure1 and Figure2 that the order of the model prediction effects is: SVM model>Bayesian network>Neural network>Decision tree.

3.3.3 SVM Modeling

For the training data set, we have a total of 2971 user samples. For the test data set, there is a pool with a sample of 1283 users. We built five test data sets by randomly extracting 50% of the data pool. Each test data set contains 641 samples. From test data 1 to 5, the number of ad click users is 120, 138, 117, 132 and 114, respectively.

Then we first run the SVM model on the training set using the RBF kernel function. Figure 3 shows the model accuracy and sensitivity for all five test data sets.

Sensitivity represents the proportion of users who are correctly identified by the user who clicked on the ad. We are more concerned with the scale of sensitivity. It can be seen that the forest sensitivity of all test sets is above 0.7. In this way, the model can correctly obtain the user's willingness to click on the advertisement.
5. Conclusions

Evaluated by the third part of the model, the SVM model works best, advertising position, gender male, gender female, weather query _ active high, online loan, online loan P2P_ active high, takeaway _ active low, weather query _ active, online shopping 10 characteristics of life and practicality can help distinguish the sample space, and the importance of these features is reduced in turn. According to the results of SVM model, this paper believes that users with click-through tendencies are most likely to have the following characteristics: gender is male, the frequency of application for weather query is generally used, and the frequency of application of P2P online loan is high, often using take-out Applications, low-frequency use of life-use applications, not too much love online shopping; ads with an ad position of 2 are more likely to be clicked.

According to the relevance of the attributes, the specific analysis is as follows: This type of advertisement is a product advertisement for the P2P loan category. The first type of users, whose gender is male, will travel to different cities. These users have entrepreneurial or excessive consumption behaviors. Their existing assets cannot satisfy entrepreneurship and consumption, and they cannot meet their loan requirements in commercial banks. Therefore, they need make an online loan. Therefore, in the model, the advertisement click is positively correlated with the gender male, the weather query active, the online loan, and the online loan p2 active high. The second type of users, such groups are women, paying attention to the family, paying attention to the practicality of life, and because of the cost-effective consumption, online shopping, attention to the details of life, and frequent weather inquiries. Users with these feature tags have a stable life, and most of them consume less than existing assets, and do not need to invest or start a business. Therefore, there is no need to pay attention to the information of P2P loans, and there are fewer clicks on such advertisements. Therefore, in the model, advertising clicks are negatively correlated with gender females, practical life, low take-out activities, online shopping, and high weather queries.

For the accurate delivery of P2P loan advertisements, it is important to focus on the first category of users (characteristic labels: gender male, weather query active, online loan, online loan P2P active high positive correlation), such users have high ad click tendency can increase the click rate of ads and increase the company's revenue.

Future research can be performed by principal component analysis or factor analysis in the data preprocessing section to classify attributes. At the bottom of the data, the ratio of click-to-advertising users to non-click-to-advertising users in the dataset is about 1:34.5, which is a serious imbalanced sample, which is an important factor leading to the deterioration of the classifier performance. In this paper, random sampling undersampling is used to reduce the proportional difference between positive and negative samples. Although the sample can be balanced, it will lead to new problems.
References

[1] Boser B, Guyon I, Vapnik V. (1992) A training algorithm for optimal margin classifiers, Fifth Annual Workshop on Computational Learning Theory. Pittsburgh: ACM Press, 27:32-36.

[2] Cortes C, Vapnik V. (1995) Support vector networks. Machine Learning, 20: 273-297.

[3] Schölkopf B, Burges C, Vapnik V. Extracting support data for a given task (1995). Proc. First International Conference on Knowledge Discovery Data Mining, 262-267.

[4] Vapnik V. The Nature of Statistical Learning Theory. (1995) NY: Springer Verlag, 132-133.

[5] XU Y, CHEN Y. (2015) An improved dynamic framed slotted ALOHA Anti-collision algorithm based on estimation method for RFID Systems, Proceedings of the IEEE International Conference on RFID, 1-8.

[6] CANTADOR I, BELLOGIN A, IGNACIO F T, et al. (2011) Semantic contextualisation of social tag-based profiles and Item recommendations, E-Commerce and Web Technologies, 85: 101-113.