



International Journal for Innovative Engineering and Management Research

A Peer Reviewed Open Access International Journal

www.ijiemr.org

COPY RIGHT



ELSEVIER
SSRN

2020 IJIEMR. Personal use of this material is permitted. Permission from IJIEMR must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. No Reprint should be done to this paper, all copy right is authenticated to Paper Authors

IJIEMR Transactions, online available on 2nd Jan 2021. Link

[:http://www.ijiemr.org/downloads.php?vol=Volume-09&issue=ISSUE-12](http://www.ijiemr.org/downloads.php?vol=Volume-09&issue=ISSUE-12)

DOI: 10.48047/IJIEMR/V09/I12/163

Title: **FEATURE ENGINEERING FOR SEARCH ADVERTISING RECOGNITION**

Volume 09, Issue 12, Pages: 954-958

Paper Authors

PARIKI MOUNIKA, POGULA PROSWI, P. SAHITHYA, PODUGU MANASA, PODUGU MANASA



USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per **UGC Guidelines** We Are Providing A Electronic Bar Code

FEATURE ENGINEERING FOR SEARCH ADVERTISING RECOGNITION

PARIKI MOUNIKA¹, POGULA PROSWI², P. SAHITHYA³, PODUGU MANASA⁴, PODUGU MANASA⁵

^{1,2,3,4} B TECH Students, Department of CSE, Princeton Institute of Engineering & Technology For Women, Hyderabad, Telangana, India.

⁵ Assistant Professor, Department of CSE, Princeton Institute of Engineering & Technology For Women, Hyderabad, Telangana, India.

Abstract: Feature engineering is one of the key technologies in the research of search advertising recognition. Most of the existing search advertising methods are selected according to the prior knowledge, which is too subjective to be popularized. Taking the advertisement of Ali search advertising as the research object, a feature processing method based on the pre-analysis of store and user data is put forward, and then the conversion rate is predicted with XGBoost (eXtreme Gradient Boosting). Experiments show that compared with other priori Feature Engineering, the proposed method can significantly improve the prediction results.

Keywords: feature engineering; pre-analysis; XGBoost; search advertisement recognition

I. Introduction

Search advertising is a common way of Internet marketing. Businesses purchase specific keywords according to the characteristics of the goods. When users enter these keywords, the corresponding advertising goods will be displayed in the pages that the user sees. The conversion rate of search advertisements is used as an index to measure the effect of advertising transformation, that is, the probability of advertising products being bought by users after clicking. With the rapid development of Internet, search advertising has become more and more popular in Internet advertising, and has become one of the most important business models in the Internet industry. In Feature Engineering, traditional feature processing methods are linear combination of original features, one-hot coding and so on. It is difficult to improve the recognition rate in traditional ways. This paper, taking Ali search advertising as the research object, proposes a feature

processing method based on store and user data pre-analysis, which aims to pre-analyze the features, that is, the first prediction processing of the features of users and stores, and as a new feature. The results of this experiment take the size of Logarithmic Loss (Logless) as the evaluation standard. In general, we must correctly handle the features and reduce the Logless value as much as possible, which is the next problem we need to solve.

II. Theory of Feature Engineering

A. Feature engineering concept Feature engineering is the most important concept in machine learning field. It can generally be considered as the work of designing feature sets for machine learning applications. For feature-based machine learning methods, the selection of feature sets determines the extreme value of the algorithm to be iterative to the optimal condition. It is also the quality of the design of the feature system that determines the

performance of the whole model in essence. Therefore, how to extract effective and highly correlated features from the existing historical data is the most important issue that we should consider next. The quality of feature design will directly affect the prediction effect of the prediction model, and also have a great impact on the spatio-temporal complexity and convergence speed of the model [1]. Data is the carrier of information, but the original data contains a lot of noise, and the expression of information is not concise enough. Therefore, the purpose of feature engineering is to use a series of engineering activities to express this information in a more efficient way of coding, and its purpose is to get better training data. Using the information represented by the feature, the information loss is less, and the law contained in the original data is still preserved. In addition, new coding methods also need to minimize the impact of uncertainties in raw data [4].

III. Common method of Feature Engineering

In general, feature selection refers to selecting a feature set that obtains the best performance of the corresponding model and algorithm. The commonly used methods in engineering are as follows: Calculate the correlation between each feature and response variable: The commonly used methods in engineering include Pearson coefficient and mutual information coefficient, Pearson coefficient can only measure linear correlation, and mutual information coefficient can measure various correlations well. The calculation is relatively complicated. Fortunately, many toolkits include this tool (such as sklearn's MINE). After getting correlation, you can sort the selection features

[2-3]. The model of a single feature is constructed, and the feature is selected by the accuracy of the model, and then the final model is trained when the target features are selected. The feature is selected by the L1 regular term: the L1 canonical method has the characteristics of sparse solution, so naturally has the characteristics of feature selection, but it should be noted that the features not selected by L1 do not represent unimportant because of two features with high correlation. Only one may be retained. If it is important to determine which features are important, cross-checking with the L2 regular method is again required. After feature selection, features are selected again: if the user id and user characteristics are combined to obtain a larger feature set and then select a feature, this practice is more common in recommendation system and advertisement systems. This is also known as mega rating or even billion rating. The main source of the feature is that the user data is sparse and the combined features can take into account both the global model and the personalized model. Through depth learning to feature selection: at present, this means is becoming a means with the popularity of deep learning, especially in the field of computer vision, due to the ability of deep learning to have automatic learning features, which is also the reason for deep learning called unsupervised feature learning. After selecting the characteristics of a neural layer from the deep learning model, it can be used to train the final target model [5].

Experiment and Analysis

In order to ensure the rigorousness and accuracy of the experimental results, the experimental data we use comes from the Alibaba International Advertising Algorithm

Competition. The data used this time includes five categories (basic data, advertising product information, user information, context information, and store information). The basic data table provides the most basic information about search advertising, as well as the "whether to trade" tag. Four types of data, such as advertising product information, user information, contextual information, and store information, provide supplementary information that may help in the conversion rate estimate. The experimental code is implemented in python and optimized with XGBoost. Each piece of data has its own instance_id, a total of 478,138 pieces of data. The characteristics of specific data are as follows:

Table 1: Basic Data

<i>Fields</i>
instance_id
is_trade
item_id
user_id
context_id
shop_id

Table 2: Advertising Product Information

<i>Fields</i>
item_id
item_category_list
item_property_list
item_brand_id
item_city_id
item_price_level
item_sales_level
item_collected_level
item_pv_level

Table 3: User Information

<i>Fields</i>
user_id
user_gender_id
user_age_level
user_occupation_id
user_star_level

Table 4: Context Information

<i>Fields</i>
context_id
context_timestamp
context_page_id
predict_category_property

Table 5: Store Information

<i>Fields</i>
shop_id
shop_review_num_level
shop_review_positive_rate

The real data collected from the experimental data processing often result in unpractical data due to the low degree of correlation. In order to make data analysis or data prediction work scientific and reliable, these "dirty data" often cannot be used directly. Before this, we need to extract the original data from these data. 1) User information processing This part of the user's natural attribute features and the user's registration information can be extracted directly from the original data, but the user's registration information is not perfect, some



users' sex, age data are unknown, and some users are home users. So we mainly solve the conversion rate according to the data given, that is, in the case of all given data and labels, we can find the conversion rate of the purchase behavior of different sex and the conversion rate of purchase behavior at different age grades, so as to produce new characteristics for us to use.

2) Advertising goods information processing. As for the attribute list of advertising goods, because of its classification, we need to calculate the number of attributes and the weights of attributes. For the list of items, the first column of the list of items calculated is exactly the same, so we only have the latter two and divide them into two features, and we are prepared for the later code. We count the number of advertising products, the brand number of advertising products, and the city code of advertising products, respectively, and calculate the conversion rate.

3) Context information processing Whether users buy goods or not has a great relationship with time, for example, double eleven Shopping Festival, Spring Festival, Christmas, Valentine's day and so on will use the consumer to produce the purchase behavior. Therefore, the display time of the commodity is also an important attribute. In this regard, we convert the time to the normal format based on the original data given, and add the weekly attributes to add new features. For the prediction category attributes of commodities, because of their missing values, we can count the number, maximum and sum of category attributes.

4) One_hot coding processing In many machine learning tasks, features are not always continuous values, but may be classified values. It will be much more efficient if the characters are represented by numbers. But when converted to digital

representation, data cannot be directly used in classifiers. Because classifier often acquiesces data is continuous and orderly. In order to solve these problems, a feasible solution is to adopt One-Hot Encoding. Also called an effective encoding, the method is to use the N bit state register to encode the N state, each state is made by his independent register bit, and at any time, only one of them is valid.

5) Initial prediction processing in order to solve the drawbacks of traditional feature processing, new innovation points are added. For this purpose, we perform prediction processing on the relevant data of shop and user and generate two new features, namely `predicted_shop_score` and `predicted_user_score`. User data is very important, and conversion is closely related to the user's wishes, so it is very important to deal with user-related features. The user's gender, age, occupation, and star rating are all relevant characteristics of the user. These characteristics directly affect the probability of the user's willingness to purchase an advertisement commodity. So for these characteristics of the user, we perform the following processing: Combine the relevant features of each user, and then use XGBoost to calculate the conversion rate of the combined data and then regenerate the one-dimensional features.

This paper discusses the prediction method of advertising conversion rate based on store characteristics and user needs, and makes a reasonable prediction for search advertising conversion rate. This method can be used to predict the conversion rate of commercial network platforms (Jingdong, Tmall, Taobao, etc.) due to the universality of the two features of store features and user needs.

IV. Conclusion

In this paper, the prediction of advertising conversion rate is studied by experiments. For training data, the selection of features has a crucial impact on the prediction performance of the model. In the prediction of advertising click rate, there are many features that can be used, including basic data, advertising commodity information, user information, context information and store information. To achieve a good accuracy rate, the model should be fully excavated as a feature, and the better the combination of these features, the better the performance of the model. The feature learning method proposed in this paper estimates the ad click rate, only considering that the full advertising data is not considered to show inadequate advertising. In the next work, how to estimate the click rate of sparse advertising from the point of view of characteristic learning is a problem worthy of study, and it is also an urgent problem to be solved at present. At the same time, we should also pay attention to the research of different models integration.

V. References

- [1] Zeng D S, Huang F L, Pan C D. Feature Engineering for Product Review Spam Identification[J]. Journal of Fujian Normal University, 2017.
- [2] Cormack G V. Feature engineering for mobile (SMS) spam filtering[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2007:871-872
- [3] Li J, Liang X, Ding W, et al. Feature engineering and tree modeling for author-paper identification challenge[C]// Kdd Cup 2013 Workshop. 2013:1-8.
- [4] Stankevich M, Isakov V, Devyatkin D, et al. Feature Engineering for Depression Detection in Social Media[C]// International Conference on Pattern Recognition Applications and Methods. 2018:426-431.
- [5] Chen J H, Li X Y, Zhao Z Q, et al. A CTR prediction method based on feature engineering and online learning[C]//International Symposium on Communications and Information Technologies. IEEE, 2018.
- [6] Chen T, Tong H, Benesty M, et al. xgboost: Extreme Gradient Boosting[J]. 2015.
- [7] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016:785-794.