Ad Click-through Rate Prediction

Project Abstract

DS 5220

Kaushik Nishtala
nishtala.k@northeastern.edu

Shubhi Saxena
saxena.shu@northeastern.edu

Satyanarayana Vadlamani
vadlamani.s@northeastern.edu

Dataset: https://www.kaggle.com/c/kddcup2012-track2/data

Search advertising is a multi-billion dollar internet industry that has served as one of the most lucrative stories in the domain of machine learning. It has relied extensively on the ability of learned models to predict ad click–through rates (CTR) accurately while promoting authenticity and low latency. The Kaggle dataset derived from session logs of the Tencent proprietary search engine, soso.com has been the point of attraction for our project as this dataset is indicative of Legitimacy, consistency and Relevance to our problem statement. Each of the 155,750,158 observations in the data identifies an instance of a search session which describes an impressed/displayed ad (AdID) by a user (UserID) who issued a particular search query (QueryID) with a certain number of ad attributes such as the depth (number of ads shown), position, title, landing page, description, keywords and the advertiser of the ad (AdvID). The fields, Impression (number of search sessions in which the ad was shown to the user) and click (number of times the ad has been clicked among the above impressions), are critical for our problem to obtain the target variable CTR by performing their division (#clicks / #impressions). Moreover, this dataset is accompanied by five additional data files, each line of which maps an ID to a list of tokens corresponding to the query, keyword, ad title, ad description and user information (age, gender), respectively.

Many proprietary search engines owned by Google, Microsoft, Yahoo etc., have effectively tackled the economic model underlying the prediction of ad CTR, which works in accordance with cost-per-click (CPC) advertising system where several ads, bidded by advertisers, are selectively picked and ranked by the product of the CPC and CTR (revenue). So the business objective for these companies centralizes on the balance between maximizing profit (and thus the CPC) and user satisfaction (and thus the CTR). The mathematical objective of our project solely revolves around finding pCTR, which is the probability that a certain ad is clicked while being conditioned on the occurrence of the ad (AdID), user (UserID) and relevant context (P(Click|AdID,UserID,Context)). Thus, accurately predicting the probability of click (pCTR) of ads is critical for maximizing the revenue and improving user satisfaction.

The given data can provide valuable insight into understanding the distribution of various measurements and their coherent relationships in order to reveal any hidden factors that might affect ad CTR by allowing us to use hypothesis testing (p-values) and multivariate analysis to answer interesting questions such as the effect of frequency of words in search query, ad title and description on the ad CTR, the kind of users lucrative ads are targeted at, the way number of clicks vary with number of impressions, the way ad attributes (such as depth, position, etc.) affect CTR, studying the role of advertisers and investigating the distinguishing factors between high performing and low performing ads based on ad frequency and ad CTR. This understanding can enlighten us to effectively model the problem as both a regression (CTR = #clicks / #impressions) and a classification (pCTR = P(Click|AdID,UserID,Context)) problem. The goodness of the predictions can be evaluated by the area under the ROC curve (AUC) or Log loss corresponding to an emphasis on the ranking of ads or the accuracy of the pCTR prediction. Although Logistic Regression seems to have a pivotal role in the early analysis in this domain, appropriate models like Ridge regression, Support vector regression, Naive Bayes, Kernel SVM, GBDT and XGBoost regression etc. may be studied and evaluated based on their model assumptions. Gradually, we could exploit the blending of individual models to boost the performance on the validation and test set.