NORTHEASTERN UNIVERSITY
DS 5220 Milestone-1 Report

Ad Click-Through Rate Prediction

Kaushik Nishtala nishtala.k@northeastern.edu
Shubhi Saxena saxena.shu@northeastern.edu
Satyanarayana Vadlamani vadlamani.s@northeastern.edu

1 INTRODUCTION

While there has been some significant progress made in our project since the proposal, a cursory review of the literature has revealed that some of the original-proposed milestones might take longer than anticipated. The reason could be partly attributed to the degree of pre-processing/tidying that is essentially required to help derive new information to feed to the model that is otherwise hidden due to the unconventional structure of the data.

Reiterating our mathematical objective, the focus of the 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|x,w)).

2 PROGRESS SO FAR

We have been successful, for the most part, in deriving valuable insights by understanding the distribution of various measurements and their coherent relationships that determine their cascading effect on Ad CTR. We began by performing extensive univariate analysis to study the individual distribution of the count of words in ad title, description and query which allowed us to effectively investigate how the word count in each of the title, description and query affects the ad CTR . We found that as the number of words in a search query increases, there is a gradual decline in the Click-Through Rate of ads (Fig 2b).

The way that the data is structured, it is inefficient to inadvertently fit the models on the raw features without considering the idea of deriving hidden information to feed to the model. For instance, from our analysis so far, we found that more impressions of an ad lead to more clicks, users with above 30 years of age have higher average CTR as compared to younger users, female users with above 30 years of age are more likely to click an ad as opposed to male users, higher performing ads have lower position (Fig 1), increase in depth of the search session leads to a decrease in average CTR, advertisers regardless of whether they have High CTR or Low CTR ads use almost the same median word count in the title and description, High CTR Advertisers have higher mean number of impressions, frequent advertisers and ads have higher average and median CTR (Fig 2a), positive correlation between IDs and impressions suggests that the IDs may contain time information, among others. Following an extensive analysis, we look to construct features exploiting the aforementioned relationships.

Owing to an abundance of categorical features in the data, we learned from our literature survey that there exists a need to adequately engineer suitable variables to capture novel information that is otherwise inhibited by the structure of the data before we fit the models. Therefore, we expanded categorical variables into binary features and for each categorical feature, we computed the average click-through rate as an additional one-dimensional feature. This essentially represents the estimated click-through rate given its category.

3 SURPRISES

Having said that, we failed to notice any statistical significance in the new features through our hypothesis testing. After scoping out the underlying problem which is due to a huge class imbalance issue, we applied simple additive smoothing method on the response, pCTR to overcome this problem. Since feature engineering and modeling go hand in hand, we extensively studied diverse models and their assumptions that optimize our problem statement.

Initially, we considered logistic regression as it works well on large datasets/high dimensions and models our hypothesis by the following a probability distribution P(Click|x,w) using a maximum likelihood problem, where the clicking probability is taken as the ranking criteria. Although Logistic regression was helpful in promoting interpretability using feature importance or hypothesis testing, it suffered from multicollinearity problem which significantly hurt the interpretability of the weights. To overcome this, we introduced polynomial or interaction features and calibrated the model to account for data imbalance.

These surprises have not adversely affected our workflow. If anything, they allowed us more flexibility and room to brainstorm model assumptions and disregard the use of models such as KNNs and Kernel LR as it is apparent that they perform very poorly on our dataset and are very expensive to train as their runtime complexities prove terrible for this problem.

4 REVELATIONS

Our theoretical contemplation of the supervised model assumptions so far yielded fresh insights in dealing with models such as Naïve Bayes which exploits a strong independence assumption to efficiently handle a large dataset. We can adopt a multinomial model for each feature in Naïve Bayes and directly use values of the estimated conditional probability as the ranking criteria, although we expect it to perform relatively poorly as it inherently assumes all the features, the probability of click is conditioned on are independent-so it may need additional calibration to be done using some weighted schemes.

Another model we eyeball at particularly is Gradient boosted Decision trees which promote low latency with their shallow trees and hence prove useful with their low run-time complexity. We realized that with GBDTs, we would have the flexibility with any loss function - so we can opt to use it with performance metrics like pseudo residual logloss or AUC that we're optimizing for. Moreover, we could make use of several interaction features to account for the polynomial nature of the decision boundary inherently with GBDTs. So with these theoretical revelations in mind, it will be quite interesting to identify how they perform in practice.

5 NEXT STEPS

In a similar trend, we plan to apply several different approaches to capture different concepts and learn a diverse set of models. I feel that enhancing this diversity in models could boost the performance when the models are appropriately aggregated. We then want to use validation set to select and compare models and then utilize the predicted results to aggregate all our generated models in the test set. So to sum it all up, our next week's work is going circle back to feature engineering and then tap into generating aforementioned individual models and assess their performance, and the later part of the project which involves blending with the validation set, and ensemble learning with the test set will be done in the week after that with a potential implementation of a Shiny Application to demonstrate the results.

6 VISUALIZATIONS

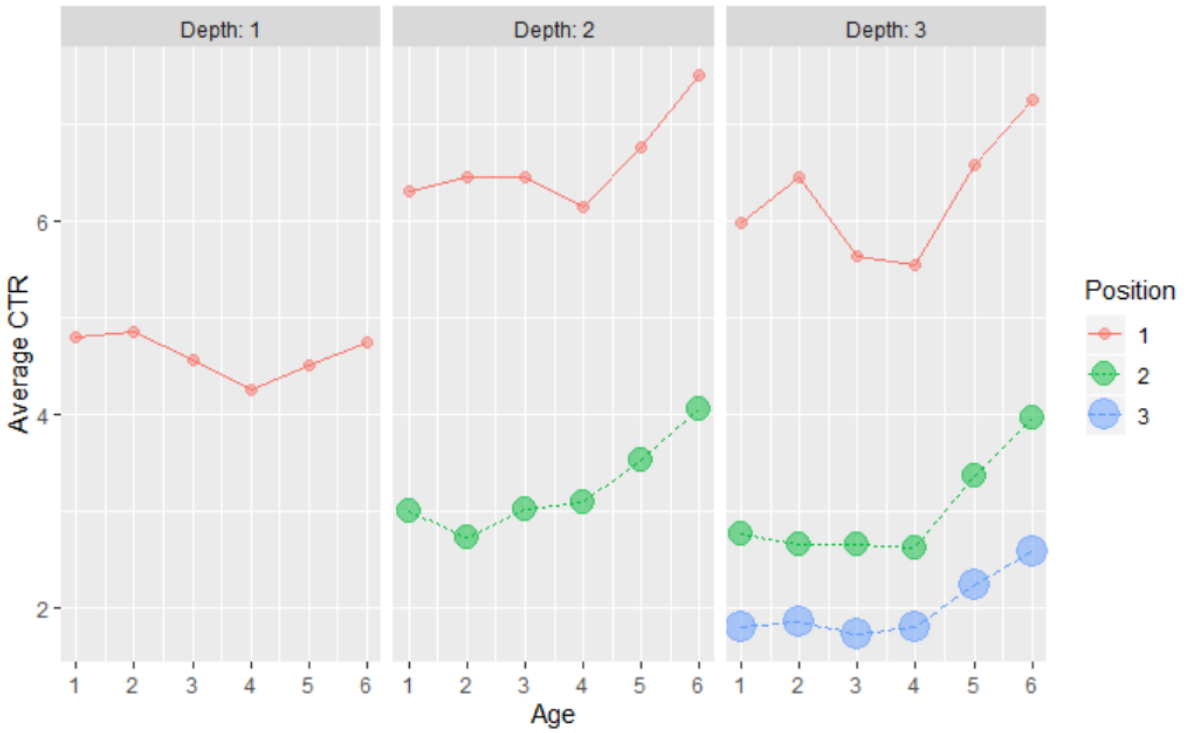
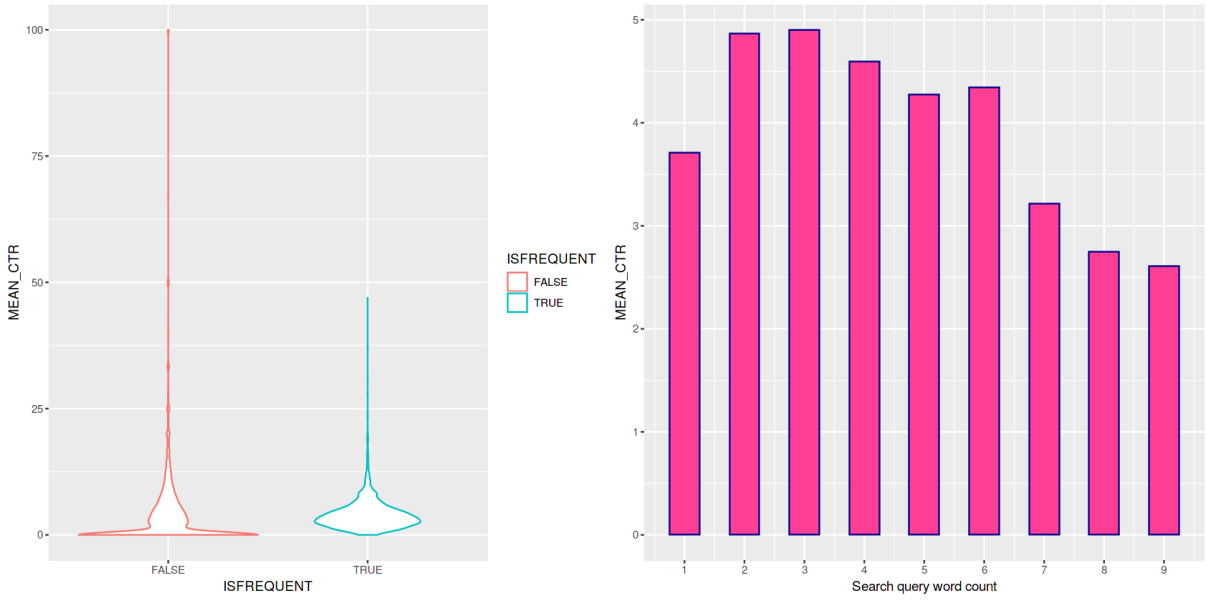


Fig 1. How Age, Position and Depth affect Ad CTR



(a) Frequent advertisers have higher average CTR (b) lengthy search queries tend to have lower CTR

Fig 2.