Slides script

SLIDE 1

Hi. I am X here with my teammates Y and Z to present our project on Ad Click through rate Prediction. We'll start off by giving you an overview of the domain this project deals with.

SLIDE 2

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.
Click-through rate or CTR for short is an important metric for ranking and pricing ads in the internet marketing world. It is the ratio of the number of times a specific Ad is clicked to the total number of times that Ad is displayed or impressed to all users.

Many proprietary search engines owned by companies like Google, Microsoft have effectively tackled this problem by making use of the cost-per-click advertising system where several ads, bidded by advertisers, are selectively picked and ranked by the ad revenue.

So the business objective for these companies centralizes on the balance between maximizing profit and user satisfaction.

Therefore, the objective of our project solely revolves around finding pCTR, which is the probability that a certain ad is clicked based on the occurrence of the ad, user and relevant context. Thus, accurately predicting the probability of clicking the ads is critical for maximizing the revenue and improving user satisfaction.

SLIDE 3

Coming to the dataset, each of the 155M observations in the data identifies an instance of a search session which describes an ad displayed to a user who issued a particular search query with a certain number of attributes such as the depth, position, title etc.

Due to the heavy computational resources needed to work on the entire data, we randomly sampled 1M train and test instances for an easier local implementation.

The features such as Impression - which is the number of search sessions in which the ad

was shown to the user - and clicks, are critical for our problem to obtain the target variable CTR.

Moreover, this dataset is accompanied by five additional data files, each of which maps an ID to a list of tokens corresponding to the query, keyword, ad title, ad description and user information respectively.

SLIDE 4

We now talk about our findings from Exploratory data analysis. EDA has been an essential part of the project in deriving valuable insights of various features and their coherent relationship on the ad CTR.

As you can notice from the slide, the 3 features Depth, Position and Age seem to have a correlation with the Average CTR.

Specifically, Users above the age of the 30 years are more likely to click an Ad as compared the younger counterparts.

Additionally, Ads appearing on the top of the search session labelled position 1 have a higher chance of being clicked than the other positions.

Similarly, we found conclusive evidence from our plots which helped us engineer new variables in order to extract information that is otherwise hidden by the given structure of the data.

SLIDE 5

Next, we move on to data wrangling phase where we started off by restructuring the data which formed the crux for modeling the data as a binary classification and regression problem.

The data composed largely of categorical variables which proved useful in our model. We began by expanding them into binary features resulting in a huge sparse matrix and for each category, we computed the average click-through rate as an additional one-dimensional feature.

In order to deal with the imbalance of the categories, Additive smoothing technique had to be used over these variables. Additionally, the dataset contained descriptive data which was utilized by using cosine similarity between TF-IDF vector of tokens.

Predicting CTRs for new ads that were not found in the training set proved challenging to address. So to overcome the problem, we made use of the existing Ad CTRs and their keyword similarities.

SLIDE 6

The CTR prediction problem could easily be modeled as a regression task by training the model to use the continuous valued CTR which is clicks-by-impressions as target variable.

On the other hand, to model this as a classification problem, we had to unroll and expand each instance of the data corresponding to the number of impressions. The number of clicks determine the tuples which will carry a 1 in the response field and the number of non-click tuples given by (impression - clicks) were assigned to 0.

Therefore, for every impression, the user may or may not choose to click an ad, which can be represented as 1 or 0.

The flexibility in modeling this hypothesis as both a regression as well as a classification problem allowed us to experiment with a diverse set of models such as Ridge and Logistic Regression, SVM, Naïve Bayes, XGBoost, Random Forests and so on.

SLIDE 7

The goodness of the predictions is evaluated by the area under the ROC curve, which is equivalent to the probability that a random pair of a positive sample, the (clicked ad) and a negative sample, the (unclicked ad) is ranked correctly using the predicted click-through rate.

We found AUC to be favorable for evaluating the performance of the hypothesis as our objective is concerned only with the order/ranking of the highest performing ads rather than the real values of their probabilities.

Finally, ROC is insensitive to changes in class distribution or skewness which makes it a better metric to evaluate with.

SLIDE 8

A wide spectrum of methods were used to reduce the number of features in each model. We managed to identify the most important features using pre-processing methods such as manually selecting features based on the domain knowledge from our Literature survey.

We also incorporated univariate selection where in we considered features that exceeded a certain threshold of correlation between the feature and the target variable.

We then made use of automated step-wise methods like Forward selection and Backward elimination in trimming out irrelevant features and finally adopted the least interpretable but effective method of PCA to analyze the variables with highest variance.

The various approaches we used helped us summarize the effects of various features on AUC as given in the table.

SLIDE 9

Initially we considered logistic regression where its internal clicking probability is taken as the ranking criteria. However, it suffered from multicollinearity and class imbalance problem which significantly hurt the interpretability of the weights and resulted in an AUC of 0.5.

To overcome this, we introduced polynomial or interaction features and calibrated the model to account for data imbalance which led to an increase in AUC value to 0.6.

The SVM which also suffered from the class imbalance issue tends to generalize majority class which is the non-clicks and discards rare class which is the clicks as can be observed from the confusion matrix resulting in an AUC value of 0.5 as shown in the plot.

It is also very expensive to train for multiple iterations to analyze the effect of smoothing and is therefore not considered.

SLIDE 10

Another model we eyeball at particularly is Decision trees which made use of several interaction features to account for the polynomial nature of the decision boundary

inherently.

The tree structure is ideal for capturing interactions between features in the data which led to a reasonable AUC of 0.62 as shown in the ROC plot. However, We observed a lack of smoothness in performance as slight changes in the input feature had a big impact on the predicted outcome which is not desirable.

Random Forests on the other hand performs reasonably well with an AUC of 0.68 but is also quite unstable as a few changes in the training dataset could create a completely different set of trees. Finally, we observed that decision trees are very interpretable as long as there are only a handful of features, unlike our data where we have over 80 different variables.

SLIDE 11

We anticipated that Naive Bayes could be an interpretable model on the modular level because of the independence assumption. It is found to be very clear for each feature how much it contributes towards a Click or no click by interpreting their conditional probability.

It performs similar to Logistic regression with a test AUC of 0.599 as summarized in the table.

XGBoost is a good starting point if the classes are not skewed too much, because it internally ensures that the bags it trains on are not imbalanced.
Due to in-built L1 and L2 regularization, efficiently handling missing values, and effective tree pruning, XGBoost Classifier outperformed other models with an AUC of 0.74 with a depth of 2 and 800 estimators as highlighted in the table.

SLIDE 12

It was an interesting challenge to appropriately evaluate regression using AUC by an algorithm that uses uncalibrated CTR scores to calculate it. Initially, the samples are sorted in the decreasing order of CTR and for each successive sample, the trapezoidal area is computed by cutting it into vertical slices at every change of CTR.

This resulted in an approximate AUC value for the hypothesis with clicks and non-clicks in the place of True Positive rate and False Positive rate.

Ridge regression performed well in addressing multicollinearity in the data where the estimation of the hyper parameter, alpha is done using GridSearch on the validation set as shown in the plot.

We observed that our regression model has low RMSE on the test data but the R^2 value is negative number with an AUC of 0.5 which meant that the predicted curve was performing worse than the mean hypothesis.

Looking at the residual and quantile-quantile plots led us to the conclusion that that model failed to capture non-linear relationships from the data and further processing must be performed to accommodate polynomial features.

On the contrary, XGBoost regressor performed decently well with an AUC 0.68 alongside providing information about the most important features.

SLIDE 13

In conclusion, we managed to get a good grip over the data and obtained a sufficiently small subset of it for local implementation.

We have obtained a best AUC score of 0.74 on test set using XGBoost Classifier and found that features corresponding to users, ad title and description semantics, search query and advertiser information are the most prominent in effecting the ad CTR.

We modeled users' behavior on click-through rate by using various linear and non-linear models and for each individual model, we used several combinations of features to capture patterns from different perspectives.

We believe the success of our methods is based on capturing various information in the data and utilizing those information effectively.
