CLICK-THROUGH RATE (CTR) PREDICTION

By - Prithviraj Patil

UOA- MSDS C5 - BATCH 3607

AGENDA

Introduction

Data Set - reading and understanding

Data Visualizing and Preparation

Model Building

Insights

Conclusion

INTRODUCTION

In online advertising, CTR is a very important metric for evaluating ad performance. CTR prediction systems are thus essential and widely used for sponsored search and real-time bidding. As part of this assignment, we are required to predict whether a user will click on an ad or not

THE FLOW OF MACHINE LEARNING

DATA PREPARATION

Check for missing or null values, outliers

DATA VISUALIZATION - EDA

Performs Exploratory data Analysis for Features

DATA TRANSFORMATION

Pre Process Data . Do test train split, Encoding, Scaling, Split, Feature Selection

MODEL SELECTION

Family of Algorithm implementation .Predictive ML Models

INSIGHTS AND CONCLUSION

Benchmark the Result.

DATASET

- The dataset contains rich information about the advertisement with 99999 rows & 27 attributes of which 8 are anonymous features maybe related to user specific or advertisers' profiles which have no readable insights for privacy reasons
- Classify these attributes into four sets and one target variable 'click' , Site specific and app specific attributes , Device specific attributes , Ad specific attributes , Anonymous attributes
- Time based columns are already broken down into hour, day, dayofweek, month
- Missing values – Not Found
- Null Values – Non
- Outliers – Capped to the 0.98 quantile percentage
- We know that click=0 means the ad was not clicked and click=1 means the ad was clicked. From the mean value observer "click" column, we can see that the number of ads clicked was 16.98% — value of "click" can only be 0 or 1, so the mean value is also the click ratio.
- Data is imbalanced - The overall click through rate is approx. 17%, and approx. 83% is not clicked.
- We have imported the necessary libraries for Data modelling and pre-requisites

DATA PREPARATION AND VISUALIZATION

- **HeatMap –** allows us to see the correlations between attributes. Recommend that the quality features should be independent of each other. If two features could mean nearly the same, it is not worth including in the feature set as it will reduce the accuracy of the model. Columns dropped basis to correlation observations stated in the workbook
'y','month','C1','C14','C15','C16','C20','C21','device_ip','device_id'
- **Feature Engineering -** The process to filter and choose quality features for the machine learning algorithms to learn from and predict the target value accurately . We generate the CTR as the number of clicks to the number of the impressions and plot them as a graph for Hours , Day of Week , Banner Position, Device Type.
- **Observations**: Highest number of clicks is at hour 13 and 14 (1pm and 2pm), and the lowest number of clicks is at hour 0 (mid-night). The highest CTR happened in the hour of 1, 7 and 15. While Tuesdays and Wednesdays have the highest number of impressions and clicks, their CTR are among the lowest. Saturdays and Sundays enjoy the high CTR. Apparently, people have more time to click over the weekend. Although banner position 0 has the highest number of impressions and clicks, banner position 5 enjoys the highest CTR. Increasing the number of ads placed on banner position 5 seems to be a good idea. Device type 1 gets the most impressions and clicks, and the other device types only get the minimum impressions and clicks. The highest CTR comes from device type 0.

DATA TRANSFORMATION

- **ENCODING** –

  One Hot encoding - All categorical variables have lot of unique values in it, one hot encoding is not a scalable approach for considering each as it increases high dimensionality. We will create dummies for features of interest only with some numerical variables too. Incase Categorical attributes we will go with Target encoding (TargetEncoder)

- **SCALING**

  Scaling StandardScaler() approach is a better idea for this. There are anonymous attributes, they are hashed to unique values and hence it becomes seamless when we need to reduce the dimensional space of the dataset in hand.

- **DATA SPLIT –**

  We have split X and y into training and testing sets 70:30 ratio into X_Train, X_Test, y_train, y_test

- **CLASS BALANCE:**

  SMOTE is an oversampling technique that generates synthetic samples from the minority class.

MODEL BUILDING

- We wrangle the dataset and use the data preparation techniques. Post these steps, we build the decision tree, the logistic regression, and the random forest machine learning models discussed using the libraries mentioned below, to name a few like NumPy, pandas , Matplotlib , Sklearn , Seaborn , metrics , etc

- **Algorithm** – Based on studied recommendation of ecommerce products to the users and the relevant literatures that worked on predicting the click through rate of any advertisement dataset, used the following machine learning algorithms to predict the probability that a user will click the ad.
- Dummy Classifier - helps in baseline the model performance w.r.t dominant class.
- Logistic regression - for interpretability and finding linear relationship
- Decision Tree Model - for interpretability and for non-linear relation
- Ensembles - Bagging , Random Forest and Boosting Algorithm's - for accuracy and improving the model
- We have predominantly used the Scikit learn libraries to implement these model

- **Bench Marking Techniques -** In this study we use the following evaluation techniques Viz. The ROC AUC Curve , The Confusion matrix, Precision, Recall, F1-score, Cross validation mean

INSIGHTS

Logistic Regression

- The feature engineering techniques and sampling strategies were some of the key parameter tuning steps that helped improve the overall performance of the research. This model seems to have
- One important thing to note when using this algorithm remove the correlated features to avoid the over fitting or under fitting of the trained model . This has been used for finding linear relationships and explainability. This model seems to have low accuracy , precision and recall as compared to the Other models.

INSIGHTS

Decision Trees

- ROC AUC indicates that the decision tree model is able to classify strictly between the positive and negative data points, click vs no click in this case. As indicated in the confusion matrix, the model has a very good precision but is inefficient with the recall bench marking strategy.
- Overfitting is observed in decision trees
- The Grid Search on decision Trees seemed to improve the balance on Test train validation . This has been used for finding non-linear relationships and explainability

INSIGHTS

Ensemble Techniques

- Random Forest algorithm selects different set of features and if we look closely at the implementation of the Random Forest, it could intuitively figure out which of the features that are used as the parameters for the training are strong candidates for the prediction of the target variables. This way, the ensembles models is able to learn the importance of any feature relatively.

- The results of the Bagging & Boosting technique are better balanced as they help in reducing the Bias and Variance respectively. The ensemble model when trained using imbalanced and balanced data gave the overall best result. We could also tune the model user the hyper parameters and max features to be used.

- **Here the models were built with focus on High predictability rather than interpretation. When using each of the above discussed classifier for the prediction,**

CONCLUSION

- The rapid growth of the social media as the online advertising platform has made it compelling for the researches to focus on improving the misclassification rate in the prediction of the click through rate of an advertisement. This will help the advertiser choose the right set of attributes to target the potential audience for their business. The accuracy can also help the advertising platforms to decide on the cost of the advertisement. The results of each step and benchmarks were set
- We see that all the classifiers predict the non-clicks better than the clicks due to the imbalance in the data. Additionally, it is also observed that data points are the mispredicted by one model is predicted rightly by other models.
- If given access to all the anonymized variables, noise in the data can be perfectly removed. We could tune the parameters wisely which could help us solve the more complex problem of choosing the right advertisement at the right time for a user. Interpretability thus is not prime importance .
- **Hence since as the categories and data available is anonymized we prefer the use of Ensembles Techniques (Bagging, Random Forest, Boosting) with Feature Engineering to get the best possible feature selection and take a wise decision on bias and variance Trade-off to avoid overfitting and underfitting.**

THANK YOU

Prithviraj Patil
prithvip.g89@gmail.com