Click-Through Rate Prediction

Data Analytics Project Report

Ashish Malpani
Masters, Computer Engineering
Virginia Tech
Blacksburg, Virginia, USA
ashish76@vt.edu

Vidur Kakar
Masters, Computer Engineering
Virginia Tech
Blacksburg, Virginia, USA
vidurk@vt.edu

PROBLEM STATEMENT

"Predict whether a mobile advertisement will be clicked by the user or not.**"**

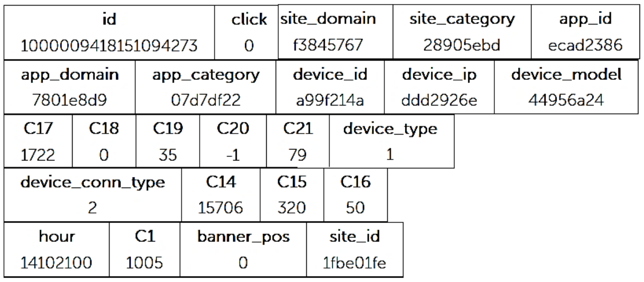
Advertising is very crucial for the businesses and plays a very significant role in the life cycle of a product or service. It discovers the customers based on the right category, makes them aware of the products and services for their benefit, increases sales potential, helps companies be aware of their competition. For businesses, especially small businesses, online advertising offers a number of benefits from being a very cost-effective solution to building a solid brand name both online and offline. Google AdWords, Google Adsense, Facebook ads, Yelp ads, Avazu and many more advertising platforms exist with coverage spanning the entire globe. In this era of digital advertising, data analytics is revolutionizing various segments of advertising ranging from optimizing pricing, detecting frauds and identifying relevant customers with similar characteristics [1].

In online advertising, Click Through Rate (CTR) or click-through rate is a metric which predicts the likelihood of an advertisement to be clicked. CTR is given as the ratio of clicks against impressions. As a result of this, click prediction systems are extremely essential and widely used for sponsored search and real-time bidding. Search engines like google selects an ad based on expected value $(Cost\ Per\ Click * CTR)$. Our work is centered around exploring the various attributes that affect CTR and how can we optimize the placements of these ads for maximum clicks. In the first section, we describe the dataset being used in detail and analyze various attributes of each row present in the dataset. In the second section, we perform pre-processing of the data available to refine it to a more useful state. In the next section, explores the dataset completely before we begin our formal data analysis. We then move on to model building and model evaluation where we work with various classification algorithms and examine the output probabilities. In the end, we present important and valuable insights for the advertising domain experts to help increase the efficiency of the digital advertising process.

DATA DESCRIPTION

For this project, we are using the dataset provided by Avazu – which is a digital advertisement agency on the open source Kaggle platform [2], which has 11 days worth of data to build and test prediction models (10 days of data for training and 1 day of data for testing). Training set is ordered chronologically with non-clicks and clicks subsampled based on different strategies. We have 40.4 million row entries with 24 attributes in the training set while test set contains 23 attributes with 4.6 million entries. Entire dataset is a little over 6 GB in size (training data) and 700 MB (testing data).

Table 1 shows the attributes, attribute type, attribute description, estimated unique values for each attribute, attribute's data type with the mean and median of the attribute. Attribute's type can be either nominal (categorical data without any quantitative significance), ordinal (whose order has significance), interval (order and numerical scales are important) or ratio (order, numerical scale along with exactness when compared to other values of similar nature). Attribute 'Id' has all unique string values and is of nominal type. Out of the 24 attributes, C1 and C14 to C21 are anonymized categorical data. Click attribute shows whether the advertisement was clicked (1) or not clicked (0). device_id and device_ip have a lot of unique values. 10 of the 24 attributes are of string data type while the remaining are of integer data types.



id	click	site_domain	site_category	app_id
1000009418151094273	0	f3845767	28905ebd	ecad2386

app_domain	app_category	device_id	device_ip	device_model
7801e8d9	07d7df22	a99f214a	ddd2926e	44956a24

C17	C18	C19	C20	C21	device_type
1722	0	35	-1	79	1

device_conn_type	C14	C15	C16
2	15706	320	50

hour	C1	banner_pos	site_id
14102100	1005	0	1fbe01fe

Figure 1: 1st row of the dataset.

Fig. 1 shows the first row of the dataset showing how the attribute values look like. A lot of the values are hashed, for example app_domain = 7801e8d9 or device_ip = ddd2926e. Hour field has the format YYMMHHDD, so in the above figure, the value 14102100 represents 00 hr. 21st October, 2014. On careful observation and analysis of the hierarchy of the unknown features in the dataset, C14 appears to be the advertisement id, C17 appears to be the advertisement group and C20 seems to be the sponsor id. Also, attributes C15 and C16 are the dimensions of the advertisement i.e. width and height of the advertisement.

We performed data transformation and feature extraction in the next section to make sure that the classifiers give the maximum ad-click prediction accuracy.

Table 1: Attributes information

Attribute	Type	Description	Unique values	Data type	Median	Mean
Id	Nominal	Ad Identifier	All unique	String	-	-
Click	Nominal	0 = non-click, 1=click	2	Integer	0	0.17
Hour	Interval	YYMMDDHH format	240	Integer	1.410e+7	1.410e+7
C1	Nominal	Anonymized data	7	Integer	1,005	1,004.968
banner_pos	Nominal	Position of the ad	7	Integer	0	0.288
site_id	Nominal	Id of the website	4,729	String	-	-
site_domain	Nominal	Domain of the website	7,755	String	-	-
site_category	Nominal	Category of the website	26	String	-	-
app_id	Nominal	Id of the application	8,529	String	-	-
app_domain	Nominal	Application domain	556	String	-	-
app_category	Nominal	Application category	36	String	-	-
device_id	Nominal	Id of the device	2,695,591	String	-	-
device_ip	Nominal	IP of the device	6,737,543	String	-	-
device_model	Nominal	Model of the device	8,254	String	-	-
device_type	Nominal	Type of the device	5	Integer	1	1.015
device_conn_type	Nominal	Connection type	4	Integer	0	0.331
C14-C21	Ordinal	Anonymized Data	-	Integer	-	-

DATA PREPROCESSING

Data pre-processing is one of the most important step in the process of data mining and analyses. It is important to ensure that the data being fed to the classifiers is not garbage and has a lot of relevance to the attribute that we are trying to predict. Data acquisition is a very loosely controlled process which results in a lot of inconsistent data combinations such as (Shape: Circle, Corners: 12) or out of range data such as (area = -10 cm^2) which puts an even higher stress on maintaining the quality of the data from processing it for any classification or evaluation [3].

First, based on intuition, we have classified the data attributes according to Fig. 2.

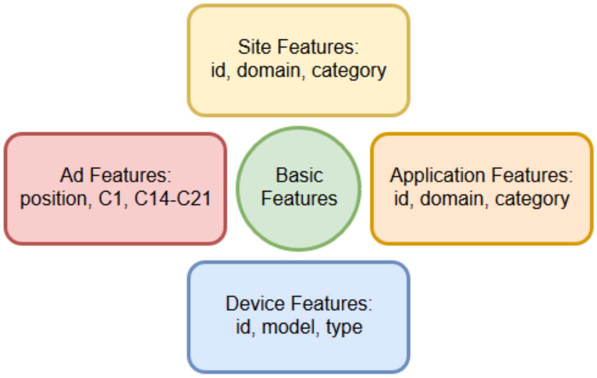


Figure 2: Intuitive categorization of the attributes.

We checked the dataset for inconsistent/incomplete or corrupt data values. The dataset on the Kaggle provided by Avazu appears to be clean and consistent, therefore, we did not have to substitute the remaining fields with custom data. Also, not all the attributes appear to be that critical for the ad-click prediction. Compared to attributes like site_id, site_domain, site_category, app_domain etc., attribute such as device_type seems to be much more reliable for predicting the click probability and is more intuitive. Mobile devices with larger displays such as a tablet can display more ads at a given time than a device like a smartphone which has a smaller panel. Some attributes need to be transformed to a proper format. For example: the “hour” attribute has YYMMDDHH format and should be converted to a more standard format.

Our data is for a period of 10 days in the month of October of the year 2014. This means all the attributes have the same month and year and hence, the hour field needs to be truncated just to give us the day and the hour. We decided to make a new attribute called ‘isWeekDay’ which is 1 in case the day for that row is a week day (Monday to Friday) and is 0 when the day is a weekend (Saturday or Sunday). This is very intuitive and studying the varying CTR over weekdays and weekends did give important results. Hour attribute, therefore, contains only the hour value during which the advertisement was shown. At the end, we separated the date and created a new attribute for the same.

Attributes like C1 and device_type which have categorical data should be converted to string format so that the classifier doesn’t get confused and starts thinking of a mathematical relationship between the values. To reduce the dimensionality of the dataset, we analyzed further to decide which fields have very little significance and can be removed. For example, attribute id has all unique values. It conveys that it cannot be used for classification and will be a bad input for knowledge discovery from the dataset. Similarly, device_id and device_ip should be removed as well. C15 and C16 being the width and height of the advertisement can be used to calculate the attribute advertisement area. It gives us the exact figure conveying how much space will the particular advertisement cover on the web page or the application.

We also ran Random Forest Classifier on all the attributes to obtain variable importance graphs for the attributes, so that we can understand the relative importance between the various attributes and their precedence with respect to predicting whether the advertisement will be clicked or not. For attributes like app_id and device_type which were categorical data and not numerical format, we assigned numerical labels to them using label encoder. Fig. 3 shows the top ten attributes which should be used for classification based on the above-mentioned classifier. We see that hour appears to be of great importance followed by C17 (ad group) and C20 (sponsor id), app_id, date, area, isWeekDay, C1 and device_type. C15 and C16 are the components which are used to derive area.

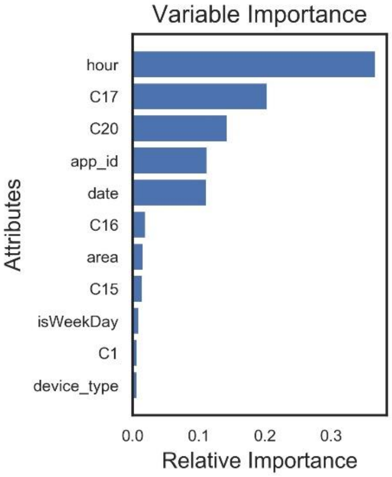


Figure 3: Variable Importance Graph

After we created the isWeekDay, area, date attributes and modified the hour attribute, we proceeded to explore the data to find interesting facts that can be derived from the data as described in the next section.

DATA EXPLORATION

The dataset is in csv format with 24 columns/attributes. Data is ordered chronologically consisting of 10 days of training and 1 day of testing data. To start with the exploration, we decided to understand how many advertisements are clicked out of the total advertisements shown in a general scenario. We took the mean of the of the click attribute of all objects and got 0.1749018. It suggests that 17.49% ads were clicked.

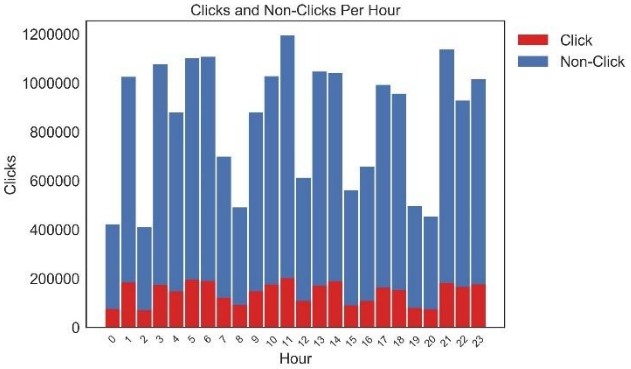


Figure 4: Clicks versus Non-Clicks hourly distribution

In Fig. 4, we plotted the total number of clicks and non-clicks per hour for all rows in the dataset. Red bars represent the total number of clicks and blue shows non-clicks. Together, they account for all the advertisement impressions. We realized that by looking at the total number of clicks, we will not be able to observe the right information. For example, one can say that more advertisements are clicked during the lunch period and early evening but at the same time, we must also consider the relatively high number of total advertisement impressions. To alleviate this problem, we consider CTR or the click through rate which considers normalized data and not absolute values. Fig. 5 shows normalized clicks and non-clicks per hour. We can observe that user engagement with the advertisements is more or less constant throughout the day. This is a good finding as we initially thought there must be some time interval in the whole day where we would see CTR go up significantly. Currently, ads on Facebook, Google and other digital advertising platforms have a time factor which comes in to picture in deciding the cost per click to the customer. Considering the hourly distribution of the normalized click rate is flat, the timing when to display the advertisement should not severely affect user engagement with the advertisement.

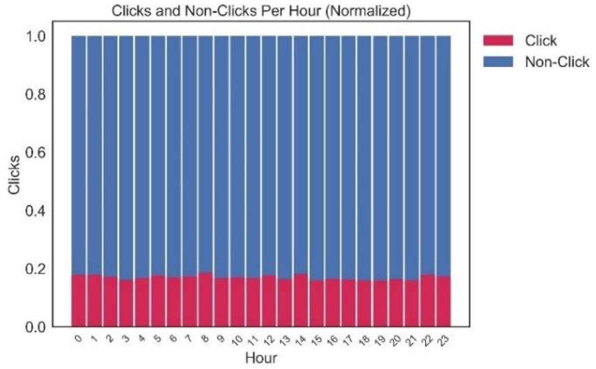


Figure 5: Normalized Clicks vs Non-Clicks hourly distribution

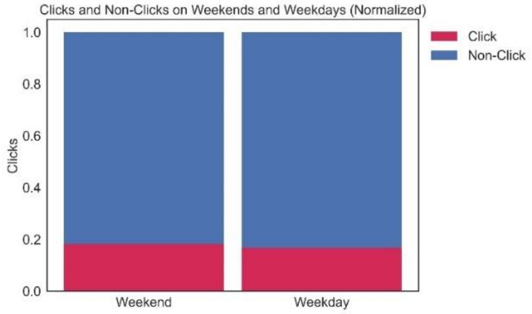


Figure 6: Weekend versus Weekday ad click rate distribution

Fig. 6 was obtained by plotting advertisement impressions versus weekday or weekend and then normalizing the data. Normalization was done so that we can negate the effect of the fact that there are five weekdays and only two weekends. We wanted to test the hypothesis that an advertisement is more likely to be clicked during the weekend as people have relatively more free time to go online. The figure tells us that it doesn't matter if one wants to display a digital advertisement on a weekday or weekend. There is a very marginal difference between the CTR values for the weekend (0.1830) and weekday (0.1690).

We were not very convinced with the outcome of the Fig. 6 and thought that it will be better to plot CTR for all days and then compute the results. Fig. 7 does exactly that and the results that we obtained were in perfect sync with the above results. Initially, we thought that this time, the outcome $CTR_{weekend} > CTR_{weekday}$ could be true but upon careful analysis, when we take the average for all the weekdays and weekend data we yet again obtain the same values i.e. weekend = (0.1830) and weekday = (0.1689).

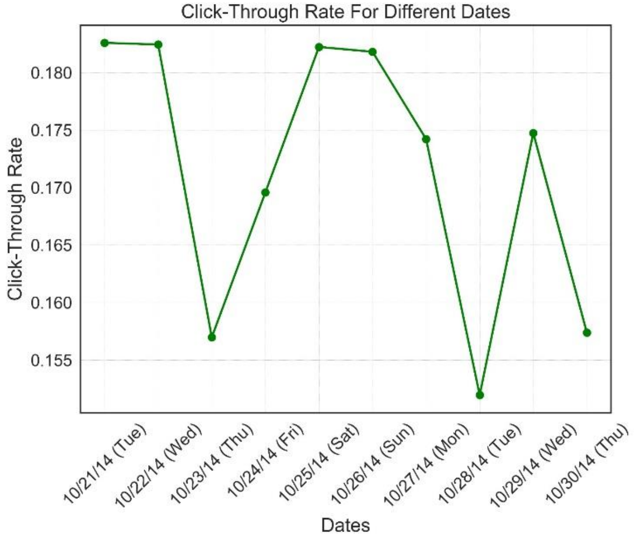


Figure 7: CTR values for each day

We also wanted to study the correlation between various attributes present in the dataset. Thus, we plotted a heatmap shown in Fig. 8 for the correlation between all the attributes of the dataset. It gave us some very important observations. The color red represents a very high degree of correlation, therefore, indicating a mutual connection between the attributes. The color blue represents no correlation which means attributes have no mutual connection between them. C14 (advertisement id) and C17 (advertisement group) appears as red which indicates that they are highly correlated. We should not include both of these attributes as inputs to our classifiers. Also, our newly calculated attribute 'area' is basically composed of C15 (width) and C16 (height). Heatmap tells us that the attribute area is highly correlated with C16 and hence we will only provide either area and C15 as the inputs to the classifier or C15 and C16 but not all the three attributes combined.

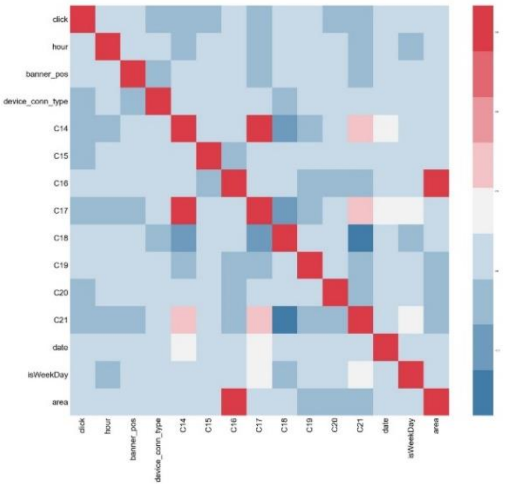


Figure 8: Heatmap of attributes correlation

Fig. 9 and Fig. 10 displays number of times the advertisements appear online which is a direct parameter to predict user activity. As we can see on the weekends, user activity increases during the afternoon period and the distribution is normal with the curve being a bell-shaped curve. For weekdays, the distribution is still normal but with a higher standard deviation. On weekdays, user activity is relatively higher spanning over more hours than weekends.

These plots can help with user activity but cannot be used to gauge user engagement with the digital advertisements. To get that right,

aa

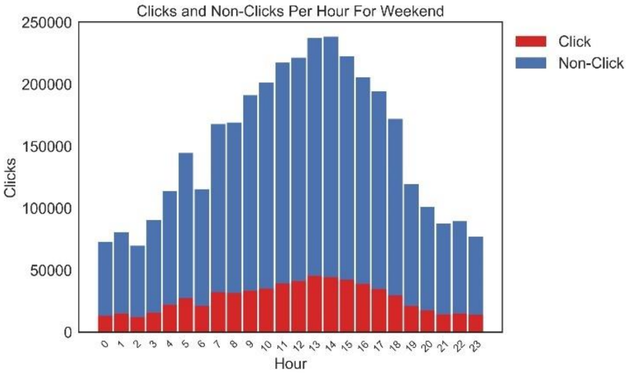


Figure 9: Advertisement Impressions Per Hour - Weekend

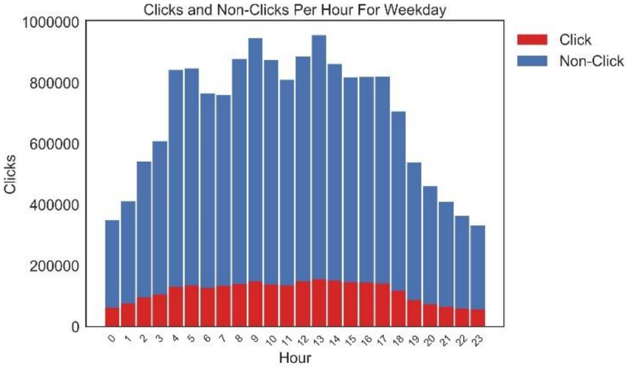


Figure 10: Advertisement Impressions Per Hour - Weekday

we must look at the CTR plots which are obtained after normalizing the data. Fig. 11 and Fig. 12 shows the normalized hourly data from weekend and weekday. As we can see, the click rate doesn't really change over time significantly during the weekend or weekday. This implies that user engagement is not a function of time which strengthens our previous hypothesis that paying more for an advertisement because of some premium and exclusive time frame/slot offered by Facebook, Google and other digital advertising companies is unnecessary as hourly distribution is flat. The cost of putting a digital advertisement online for a customer can be greatly reduced based on this newly acquired knowledge. The digital advertisement's probability of being clicked is more or less dependent on parameters like device_type, connect_type, area and more but not time.

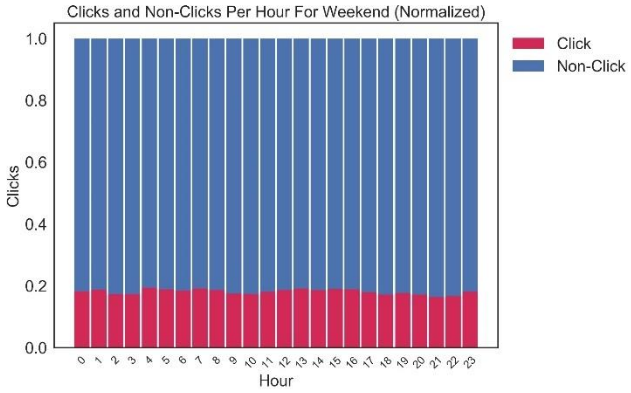


Figure 11: Ad Impressions/Hour – Weekend (Normalized)

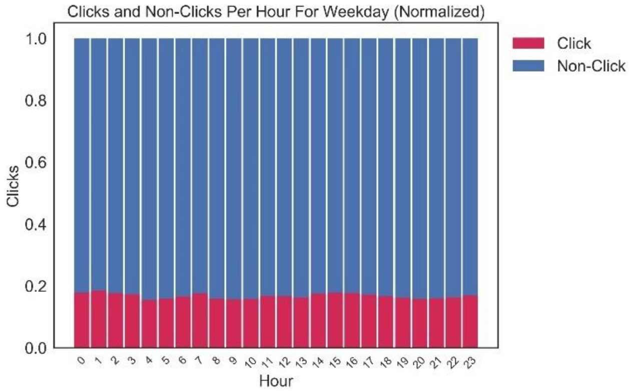


Figure 12: Ad Impressions/Hour – Weekday (Normalized)

We also wanted to research the relationship between the attribute area and CTR for the advertisement. We made a plot for varying areas of different advertisements and corresponding CTR values to understand the relationship between the two attributes.

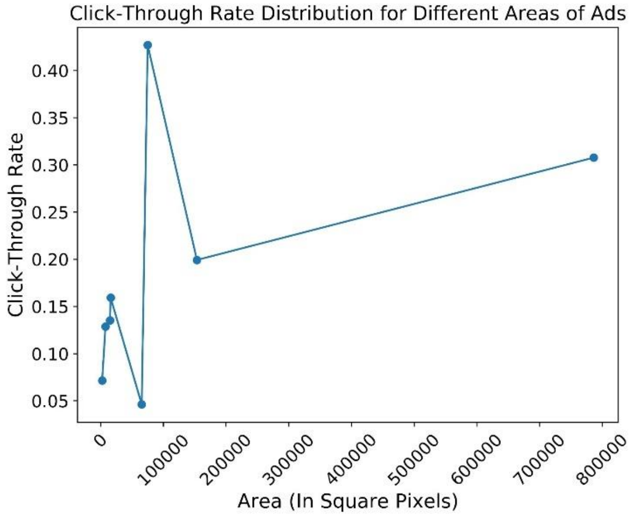


Figure 13: CTR on areal distribution

Fig. 13 shows the graph and one can clearly note the following observations:

- CTR values are low for advertisements that are very small. It makes sense as a small advertisement indicates a very bad probability that the user will click the advertisement.
- CTR values are not very good for large advertisements. This is straight forward as a very large advertisement means that the users will do their best to avoid the advertisement and not click on it as a result. Also, the majority of such ads are spam – which further reduces the chances for the large advertisements being clicked.
- CTR is the highest for advertisements that are neither too large nor too small. Plot gives this estimate to be around 100,000 pixel^2 area. It can be considered something like a 300 pixels x 300 pixels advertisement, which seems like a very ideal size for an advertisement.

The next two sections elaborate on the classifier models that we used, our model evaluation process and results.

MODEL BUILDING

After exploring the dataset, we proceeded to build classifiers to run on the dataset. We are predicting whether the ad was clicked or not which indicates a dichotomous variability (one of the two possible outcomes). So, we require a classification model. We were overviewing various models and decided to go for the following classifiers:

- Decision Tree
- Gradient Boosted Decision Tree
- Logistic Classifier
- Random Forest Classifier

We will be using accuracy and log loss as the metrics to evaluate the performance of these above-mentioned classifiers. To make the system more robust, accurate and to prevent overfitting, we have used cross validation technique. We are using GraphLab by Turi Inc. to implement these classifiers [4].

Decision Tree is a decision-based classifier that uses a tree-like graph structure or model of decisions to predict the possible consequences, which includes chance event outcomes, resource and utility cost [5]. Based on the variable importance graph, we will be using hour, device_type, area, isWeekDay, C1, C15, C20, C17, app_id and date as the features to be used for training the model. Decision trees are simple to understand and interpret. Not only that, decision trees can be combined with other decision techniques. It has some issues too, such as instability and inaccuracy. They also favor attributes with more levels in case of categorical data.

Gradient Boosted Regression is a technique for classification problem which produces a prediction model in the form of an ensemble of weak prediction models. A gradient boosted Decision tree basically has that prediction model in the form of a tree. It predicts a continuous variable which is then further used to predict the class to which the data point belongs [6]. Thus, GBDT can be used to classify the ad click variable to 1 and 0. Based on the variable importance graph, we will be using hour, device_type, area, isWeekDay, C1, C15, C20, C17, app_id and date as the features to be used for training the model. One advantage of this model is that it can reduce overfitting by regularization techniques to prevent degradation of the model's ability to generalize.

Logistic classifier is a model that is used to predict a continuous variable. In our case, we are using a Logistic classifier to predict the probability for the click to be true which is continuous in nature and now, this continuous variable will be used to perform classification [7]. It is a very good choice for a model since it is very fast and does not take a very long time to make predictions. It is much more interpretable than other models which makes it easier to understand how the model is making the predictions. Most important feature of this classifier would be that it outputs very well-calibrated probabilities. We tried multiple attributes to be used as features in training this model and kept iterating till we obtained a good score on our metric. The attributes were hour, C17, app_id, C20, area, C15, isWeekDay and date.

Random forest classifier is an estimator that predicts a fit for a number of decision tree classifiers and uses the averaging technique to improve accuracy and prevent over-fitting. It searches for the best feature among the random subset of features [8]. It can be used for classification as well as regression problems. We are using it for classification whether the advertisement is clicked or not. For this, we selected multiple feature sets involving different attributes short listed from the variable importance graph.

To evaluate we are using metrics namely accuracy, log loss and area under the Receiver Operating Characteristic or ROC curve.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

Where, TP=True Positive, TN=True Negative, FP=False Positive, FN=False Negative

$$Log\ Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum\nolimits_{j=1}^{M} y_{ij}\log p_{ij}$$

Where, N = number of samples, M = number of possible labels

y_{ij} indicates if label j is the correct classification for instance i

p_{ij} indicates the probability of assigning label j to instance i

To make our system more accurate and perform better on a testing set, we perform 10 – fold Cross Validation on our data where we split our data into ten parts and for each iteration, use nine parts as training set and one part as test set [9]. We also tuned hyper parameters of each classifier to get best model performance.

MODEL EVALUATION

The dataset was split into 80% training set and 20% testing set. We had to do this because the test set available on Kaggle did not have ground truth. Our evaluation was done using ten-fold cross validation. We recorded the accuracy and log loss metrics for all the classifiers used. Table 2 summarizes the performance metrics obtained on classification.

Table 2: Model Evaluation – Performance Metrics

Classifier	Accuracy (%)	Log Loss
Logistic Classifier	82.867	0.44748
Random Forest	83.079	0.43049
Decision Tree	81.015	0.57593
GBDT	82.322	0.43599

All of our models have a very low loss compared to the base line model which gave a log loss value of 5.89136. Also, our best classifier gave a better accuracy than baseline accuracy of 82.94%. A higher accuracy means a higher count of predictions where predicted value is equal to the actual value. It is still not a good indicator because it has a yes or no nature and our dataset is skewed towards no click (click = 0). This is why we are also looking at the log loss parameter. Log loss takes into account the uncertainty of our prediction model on how much it varies from the actual label [10]. It gives a more detailed view into the performance of our model. Clearly, random forest classifier performed the best. It has the highest accuracy with the lowest value of log loss among the classifiers which we have tried. Random forest performs implicit feature selection and is very quick to train. Random forest classifier is not very sensitive to some particular hyper-parameters used. The averaging method basically allows the classifier to run several estimators independently and then average their predictions. It is an ensemble method which uses several weak models to give a strong model with reduced variance and hence yields a better result [11].

We also found that for us, both the decision tree as well as the gradient boosted decision tree beat the Logistic classifier model, which we initially thought will perform better as the flexibility of decision trees may lead to overfitting. But since our classification is more of a binary classification with 1 being the advertisement being clicked and 0 conveying the advertisement is not clicked, decision trees did perform better. It was hard for our dataset to be segregated by a single linear boundary and hence Logistic classifier could not score higher [12].

Table 3: Area Under Curve (AUC) for classifiers

Classifier	Area Under Curve (AUC)
Logistic Classifier	0.65966
Random Forest	0.71151
Decision Tree	0.68869
GBDT	0.69644

Fig. 14 shows the ROC curve obtained for the classifiers used. Table 3 displays the area under curve obtained for the various classifiers used. Random forest is the clear winner with the highest area under curve confirming that it has the highest accuracy. Also, gradient boosted decision tree scores higher than the classic decision tree because it builds the model in a stage-wise fashion and then generalizes them by optimization of arbitrary differentiable loss functions.

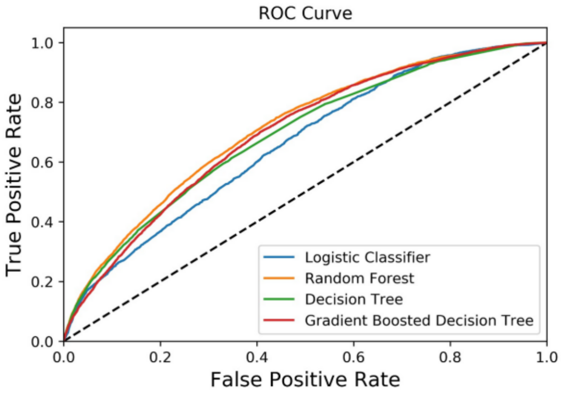


Figure 14: ROC Curve for classifiers used

REAL WORD INSIGHTS

One important aspect in advertising is the publicity or space. It is essential that people can find the company and the product at the right time. In this digital era, a multitude of transactions happen online. Hence, companies who wants to stay at the top of their game must utilize the online advertising platforms. These platforms have varying cost models. Some may charge the user more for putting an advertisement on a specific day than some other day. Some of these advertising platforms may also charge more a specific hour which they label as the peak hour or premium hour.

Another important aspect is conversion. “Conversion” is the prime key in advertising. Some people believe that having more people know about the product is all that matters while some would agree that advertising's sole goal at the end of the day is nothing but a game of conversion. Customers must organically click those ads, reach the company's marketplace and actually make a purchase of their service or product.

Our research work unearths a lot of key findings that could help the business be more efficient with the money invested for their advertising campaign.

Click through Rate is one of the most important parameters for online advertising. CTR or the probability of a digital advertisement being clicked is independent of time. It doesn't matter at what hour you want your advertisement to appear online because CTR is more or less constant throughout the day. So, for those looking for conversion as the key aspect must not waste money in buying a more premium slot on an advertising platform but what about those people who are looking for a larger crowd to see their product and company name? Only in such a case, they must invest in a slot with a better time factor. It will allow a larger population to be made aware of the product but CTR still remains the same so the businesses should not expect a higher percentage of sales for different hours even if the absolute numbers are higher. One must also consider the extra money that goes in the premium slot.

What about displaying an advertisement on a weekday versus a weekend? In this case, CTR nearly remains the same but same rules apply. Conversion rate and user engagement ratio will remain constant so paying more for advertisements being displayed on a weekend is only worth if instead of conversion, publicity is of a higher concern. People come online for a larger duration on weekdays than on weekends so for those who want their product to be seen by a larger number of people, must exploit this factor.

The advertisement itself is the biggest factor if it will be clicked or not. The area it occupies on the online source is of chief concern. A very large advertisement will be ignored by users who may think it is a scam or might just avoid interacting with it. A very small advertisement will be ineffective at capturing the audience or appealing to them. A good fit for dimensions is must if we want the people online to click the advertisements. An advertisement, having an area of about 100,000 square pixels is optimum. It can have both the dimensions around 300. It should certainly help increase the CTR as per our research and exploration. Thus, the above findings can help small scale to medium scale businesses advertise in a cost-effective manner

REFERENCES

[1] StackAdapt Blog. (2018). *How Data Science is Revolutionizing Digital Advertising [Interview] - StackAdapt Blog*. [online] Available at: https://blog.stackadapt.com/what-is-data-science-and-how-is-it-revolutionizing-the-digital-advertising-industry/

[2] Click-Through Rate Prediction | Kaggle: 2018. [online] *https://www.kaggle.com/c/avazu-ctr-prediction/data*.

[3] Pyle, D., 1999. *Data Preparation for Data Mining*. Morgan Kaufmann Publishers, Los Altos, California.

[4] Turi Create Intelligence – Graph Lab 2018, https://github.com/apple/turicreate

[5] *Kamiński, B.; Jakubczyk, M.; Szufel, P. (2017). "A framework for sensitivity analysis of decision trees". Central European Journal of Operations Research.*

[6] Friedman, J. H. "Greedy Function Approximation: A Gradient Boosting Machine." (February 1999)

[7] *David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press. p. 128.*

[8] Ensemble methods | 2018 Bagging meta estimators http://scikit-learn.org/stable/modules/ensemble.html

[9] 10-fold cross validation | Properties https://www.openml.org/a/estimation-procedures/

[10] Log Loss - http://wiki.fast.ai/index.php/Log_Loss

[11] Machine Learning and metrics https://blog.statsbot.co/ensemble-learning-d1dcd548e936?gi=e3202aed0954

[12] What is logistic regressor? https://www.statisticssolutions.com/what-is-logistic-regression/

Click-Through Rate Prediction

DATA ANALYTICS (CS 5525)

ASHISH MALPANI (ashish76@vt.edu)

VIDUR KAKAR (vidurkakar@vt.edu)

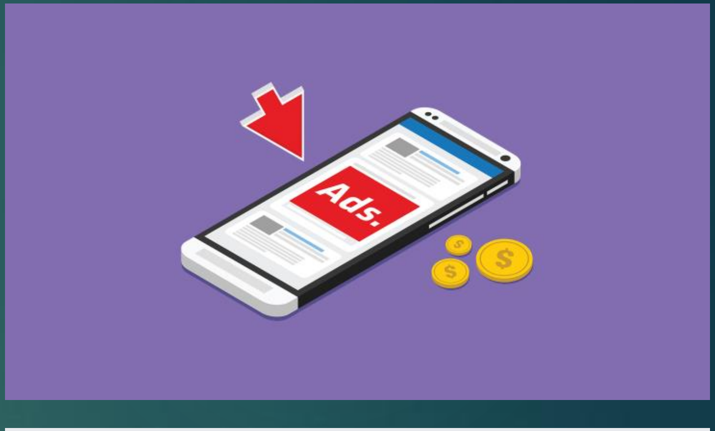
Problem Statement

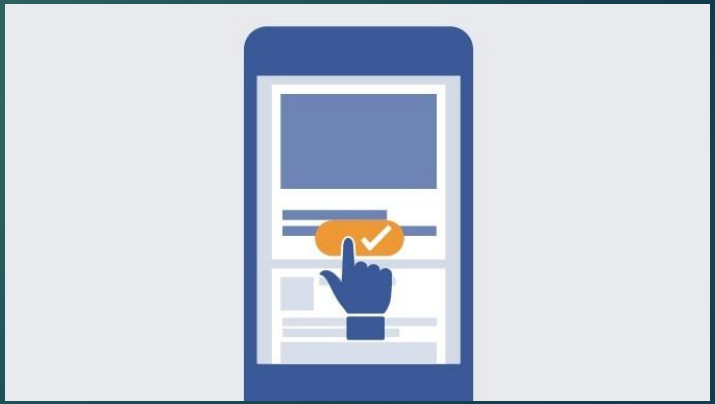
- “To predict whether a mobile advertisement will be clicked buy the user or not.”

- Further, we calculate the Click Through Rate (CTR) which is defined as the following –

$$Click\ Through\ Rate\ = \frac{Number\ of\ Advertisements\ Clicked}{Total\ number\ of\ Advertisments\ Published}$$

- For online advertising, the CTR is a very important metric to measure the advertisement performance and estimate the revenue that will be earned from the advertisement





Data Description

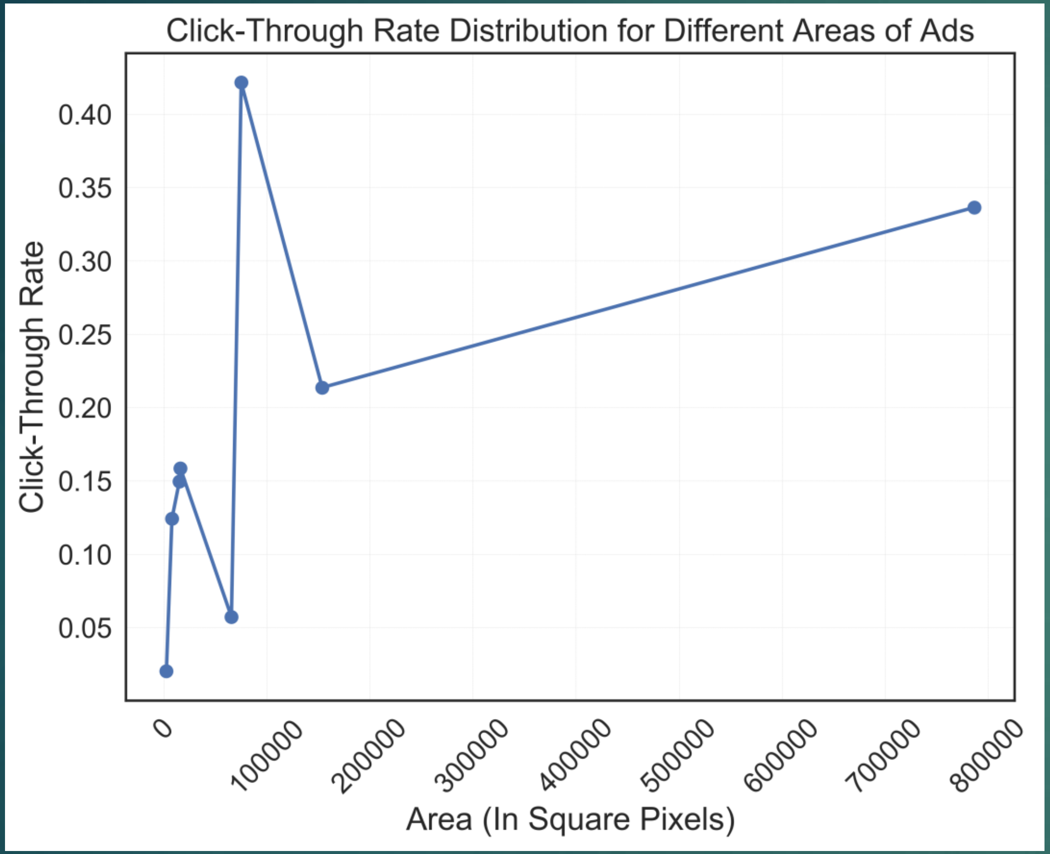
Attribute	Type	Description	Unique values	Data type
Id	Nominal	Ad Identifier	All unique	String
Click	Nominal	0 = non-click, 1=click	2	Integer
Hour	Interval	YYMMDD HH format	240	Integer
C1	Nominal	Anonymized data	7	Integer
banner_pos	Nominal	Position of the ad	7	Integer
site_id	Nominal	Id of the website	4,729	String
site_domain	Nominal	Domain of the website	7,755	String
site_category	Nominal	Category of the website	26	String
Date	**Interval**	**Date of Record**	**10**	**String**

Attribute	Type	Description	Unique values	Data type
app_domain	Nominal	Application domain	556	String
app_category	Nominal	Application category	36	String
device_id	Nominal	Id of the device	2,695,591	String
device_ip	Nominal	IP of the device	6,737,543	String
device_model	Nominal	Model of the device	8,254	String
device_type	Nominal	Type of the device	5	Integer
device_conn_type	Nominal	Connection type	4	Integer
C14-C21	Ordinal	Anonymized Data	-	Integer
Area	**Nominal**	**Area of Ad**	**8**	**Integer**
isItWeekday	**Nominal**	**Record of Weekday or Weekend**	**2**	**Integer**

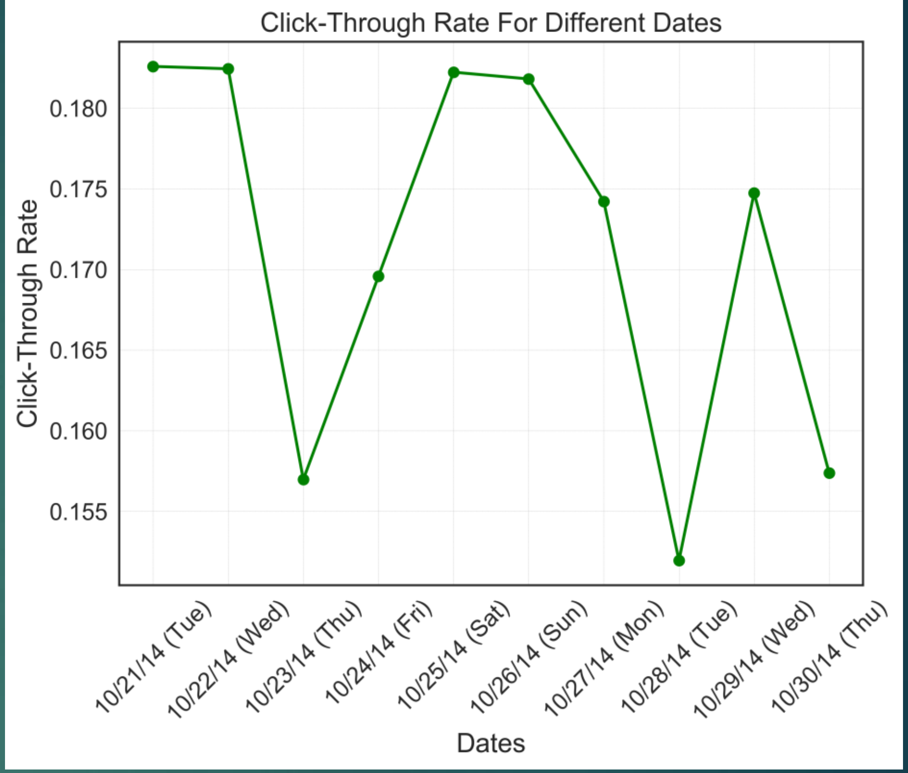
The attributes in bold have been generated by us.

Data Preprocessing

The following attributed were generated by us -



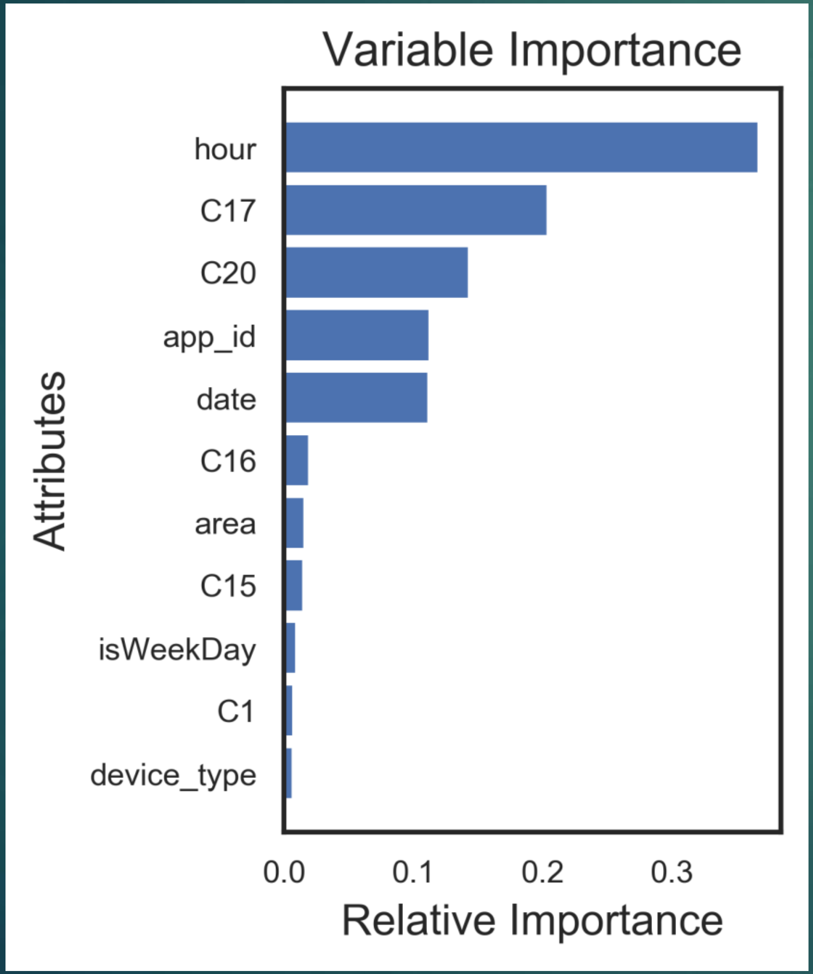
Area: Generated by multiplying height and width of Ad. to get an insight on how CTR varies with different Ad. height



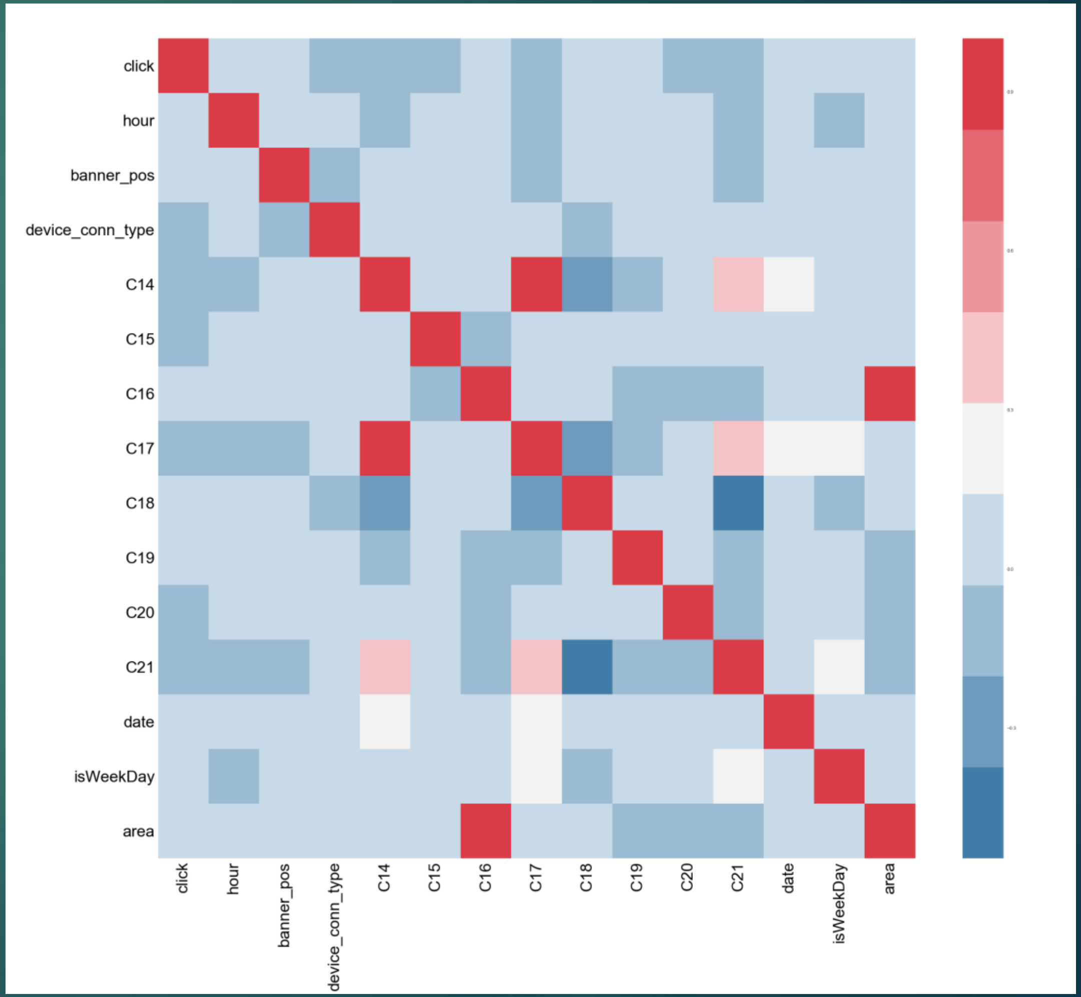
Date: Generated to get an insight on how CTR varies every day which is then used to determine the factors that cause that variation

isWeekDay: To categorize whether the ad was clicked on a weekday or a weekend. This was used to find insights such as CTR is marginally higher on weekends

Variable Importance Graph - We calculated the Variable Importance and plotted the top 10 attributes on graph to figure out which attributes are important and should be used by the classifiers

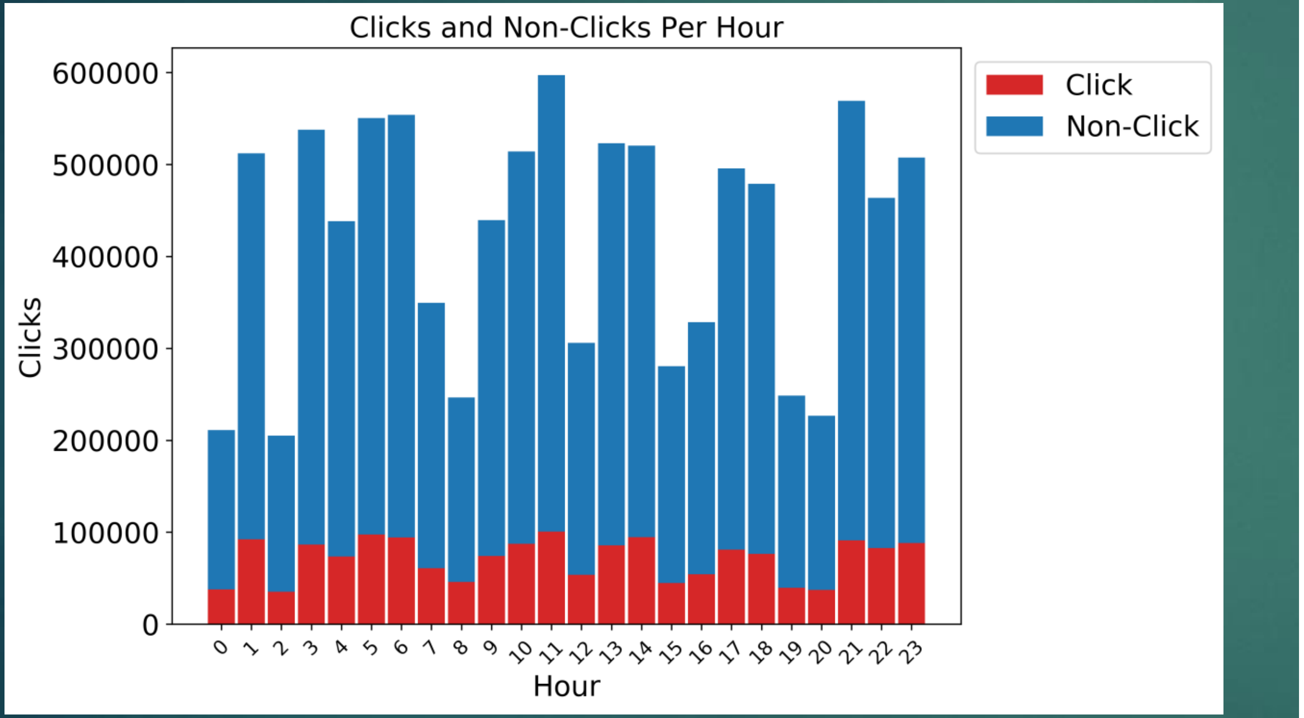


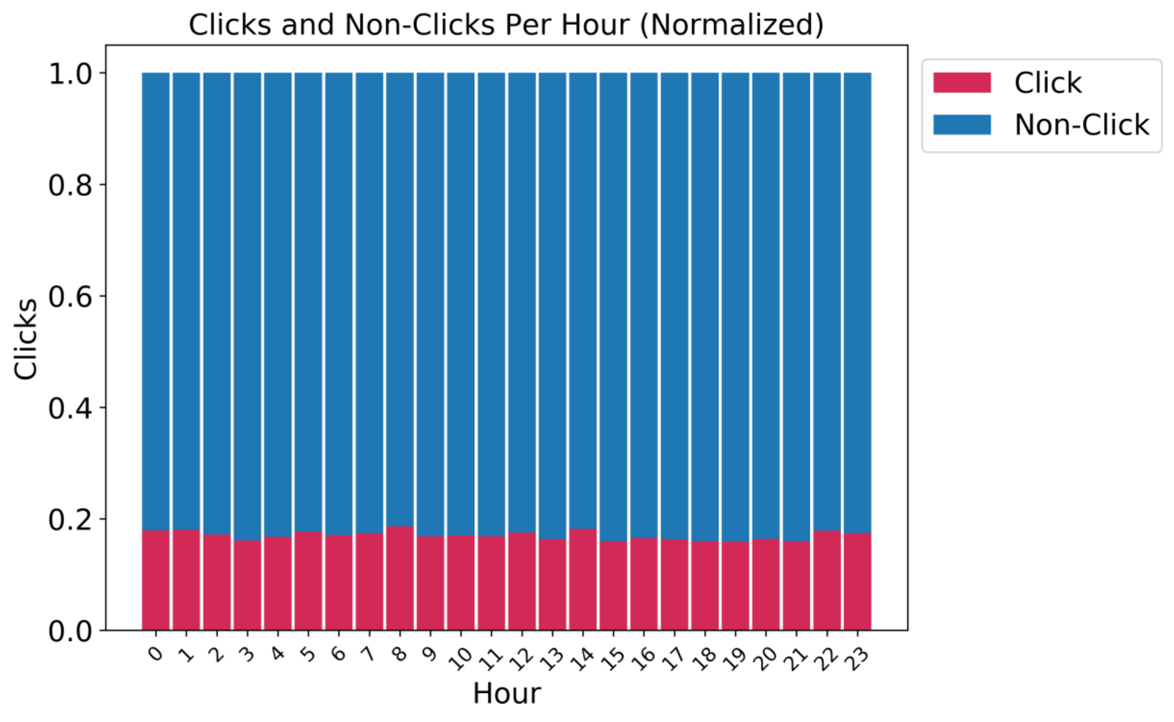
Correlation Matrix - A correlation matrix was plotted to figure out the attributes which were highly correlated and should not be used in the classifier models together.



Clicks by the Hour

- The number of ad impressions and ad clicks vary throughout the date as shown in the graph below on the left side.
- But on the other hand, if we normalize the data and plot the Normalized clicks and Non-Clicks, per hour, we get to know that there is not much variation in the Click Through Rate per hour.





Model Building

- Here, we are predicting “whether the ad will be clicked or not”, which is a dichotomous variable. It has only two values - 0 or 1.
- Thus, we will be using Classifier Models to predict the Click Through Rate (CTR). We are also using the models to predict the probability if an ad being clicked.
- The following Classifier models were used to predict the CTR
  - Logistic Classifier
  - Decision Tree
  - Gradient Boosted Decision Tree (GBDT)
  - Random Forest Classifier

- The models were evaluated based on the following metrics –
- Log Loss- Used to quantify the accuracy of the classifier by penalizing the false classification.

$$\text{Log Loss} = -\frac{1}{\text{N}}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log p_{ij}$$

- Accuracy – Used to measure how accurately our classifier classifies the data to it’s right class label.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

- ROC Curve- ROC curve is also used to measure the classifier performance by calculating the area under the curve in the ROC graph.
- Feature Set – Various feature sets were made from the Variable Importance Graph and tried to on the classification models to narrow down on the feature set that yields the best classification result.

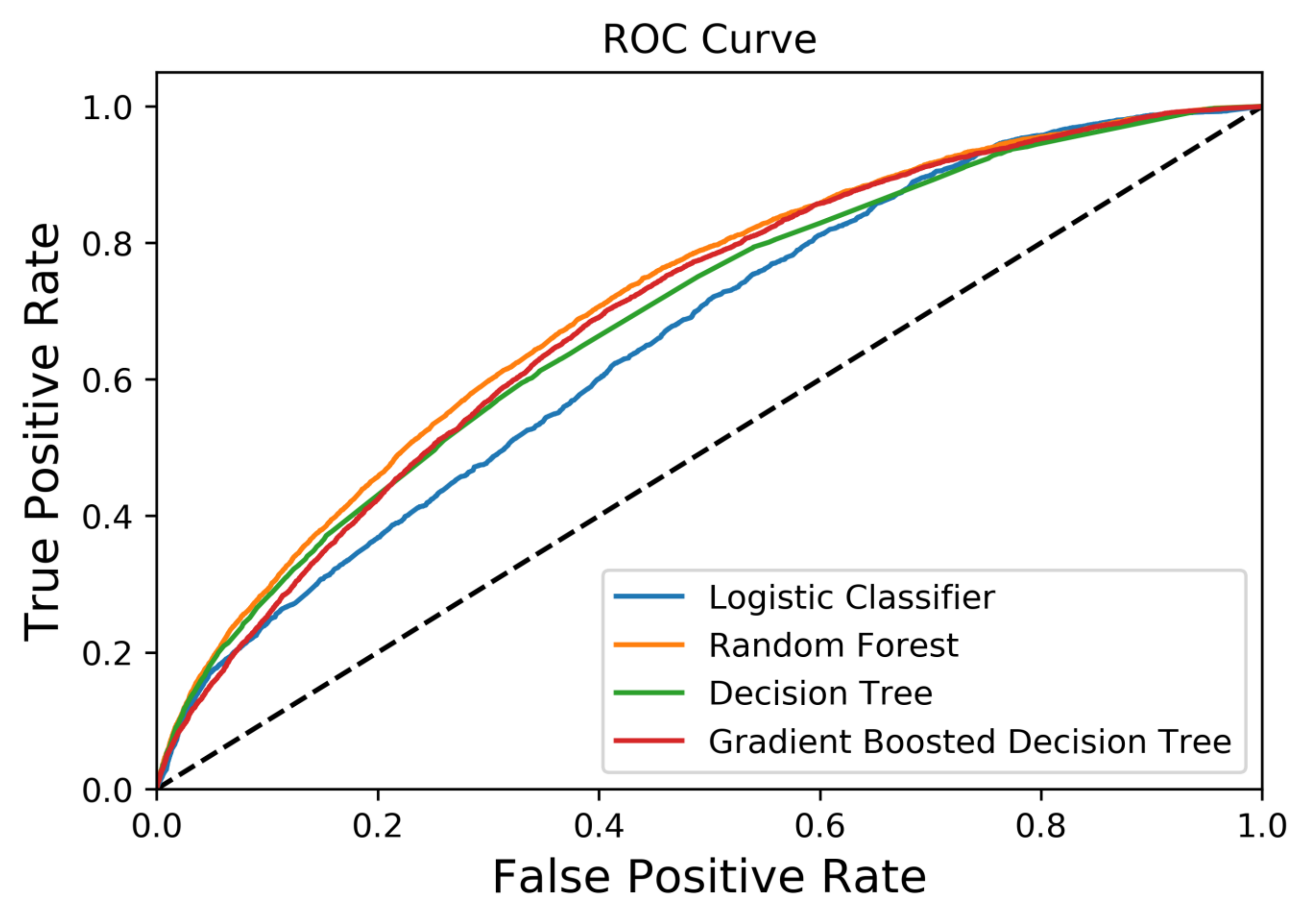
Model Evaluation

Classifier	Accuracy (%)	Log Loss
Logistic Classifiern	82.867	0.44748
Random Forest	83.079	0.43049
Decision Tree	81.015	0.57593
GBDT	82.322	0.43599

From our values obtained for Log Loss and Accuracy, we figured out the following was the order of Best Classifier in descending order –

1. Random Forest – Best classifier as it is an ensemble method yielding best results.
2. Gradient Boosted Decision Tree
3. Logistic Regression
4. Decision Tree – Worst performance as tends to overfit for the given dataset

ROC Curve



Classifier	**Area Under Curve (AUC)**
Logistic Regression	0.65966
Random Forest	0.71151
Decision Tree	0.68869
GBDT	0.69644

- Our area under the ROC curve also indicated that Random Forest is the best classifier used, followed by GBDT, Decision tree and Logistic Regression.

Real World Insights

- CTR is independent of day of the week i.e. $CTR_{weekday} \approx CTR_{weekend}$.
- CTR is independent of the time of the day. Though, the advertisement impressions do increase during the afternoon or lunch period, CTR remains unchanged. This is good for increasing publicity but doesn't increase user engagement with the advertisement which is the key point of the advertising campaign.

Real World Insights (continued)

- The size of the advertisement is a very important criteria for an online being to click the ad.
- If the advertisement is too small, the advertisement may be ignored and fail to capture audience.
- If it is too big, it may appear like spam or might distract the audience.
- A right size or for the advertisement is crucial. It should be somewhere around 300 pixels by 300 pixels or have an area near 100,000 square pixels.