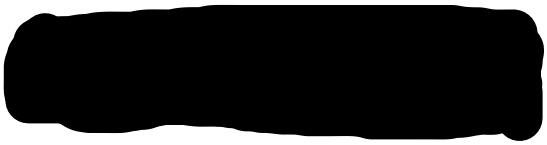
Predicting Ad Clicks for Target Advertising

Milestone: Final Report

Group 45
Prachi Patel
Jignasuben Vekariya



patel.prachi2@northeastern.edu
vekariya.j@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Prachi Patel

Signature of Student 2: J.R.Vekariya

Submission Date: 04/25/2022

Table of Contents

Problem Setting:

Digital marketing through advertising is one of the most effective ways to advertise a product and the online retail industry is one of the biggest markets for advertising products and brands. A customer does not need to necessarily be familiar with your business or brand to find you, as your products can appear whenever the search engine matches them to the user's search keywords.

But often that is not the case and users end up seeing ads of products they are not interested in, and the platform loses out on revenue by showing them ads which will never get a click. Studies shows that showing the correct ad to relevant customer can increase the conversion rate by up to 26% which can be a game changer for many retail companies.

Problem Definition:

The aim of this analysis is to use multiple models and pick the model that best helps assess the relevance of ads to target customers which will lead to eventual increase in the click rate . At its most basic, targeted advertising can just mean that ads are chosen for their relevance to site content, in the assumption that they will then be relevant to the site audience as well. Online advertisers target potential customers based on their browsing traits which would eventually lead to higher profits for the company as well as target an audience that sees what they may be interested in buying. Our intention is to analyze various factors by choosing a model which provides the best performance for the task.

Data Sources:

The dataset is taken from GitHub https://github.com/TriMinhDuong/marketingad-click prediction/blob/master/data/advertising.csv

Data Description:

Here we are considering a dataset that includes features that help us predict which customers are likely to click on an ad. This can assist and guide the marketing team of an organization to target the appropriate target customers to maximize profitability.

The dataset that we are considering has 1019 rows and 10 columns which include features like country, timestamp, time spent on the site, male or female, age of the user, while the ad features include details about the ad topic as well as whether the individual has clicked on the ad or not. These features can then be further used to personalize ads for customers visiting the site to help increase revenue as well as customer satisfaction.

Data Collection and Processing

Data collection and exploration is the initial preprocessing of data in which raw data is collected from its source, reviewed, and cleaned in order to help generate patterns and identify outliers along with identifying the relationships between different variables.

Data Exploration and Cleaning

Data cleaning and exploration involves dealing with raw data and converting it into a form which is easy to access and analyze.

Initially we first load the dataset and then obtain the first five rows in order to get an overview of what the columns are and their respective datatypes.

```
Data columns (total 10 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Daily Time Spent on Site  1018 non-null   float64
 1   Age                       1018 non-null   int64
 2   Area Income               1018 non-null   float64
 3   Daily Internet Usage      1018 non-null   float64
 4   Ad Topic Line             1018 non-null   object
 5   City                      1018 non-null   object
 6   Male                      1018 non-null   int64
 7   Country                   1018 non-null   object
 8   Timestamp                 1018 non-null   object
 9   Clicked on Ad             1014 non-null   float64
dtypes: float64(4), int64(2), object(4)
memory usage: 79.7+ KB
```

Table 1: Information about the columns in the dataset

The Timestamp column is a string and hence we wish to convert it into its appropriate data type which is datetime. Since we aim to clean the data so that it becomes much easier to analyze we need to make sure

that there are no duplicates or null values as that would cause poor reporting and skewed metrics. We notice that there are around 7 duplicate values and 4 null values which need to be removed.

The null values were present in the column clicked on ad. In this case since it is a 0 or 1 i.e., either the user clicked on the ad (1) or did not(0) we cannot interpolate the data as it will lead to skewing, hence in this case it is best to remove the missing values.

Data Exploration and Visualization

Exploring data and displaying it in a visual form is an important tool to help tell us a story, making it easy to understand by highlighting trends and outliers. Removing excess noise, gives us a clear picture and helps enable us to draw coherent conclusions about the data.

Identifying and removing Outliers:

Now that we have removed the duplicates and missing values in our dataset, we want to make sure that there are no outliers present. Outliers can increase bias and reduces the effect of the tests performed on the model.

We obtain the details of the dataset, and we can see that there are extreme values present in the Age column, -25 and 999. Assuming an average individual to be between the age of 0 to 100 years we see the total number of outliers present within the age column to be 3.

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Timestamp	Clicked on Ad
294	56.01	-25	46339.25	127.26	Re-engineered real-time success	Garciamouth	0	Tonga	2016-07-05 18:59	1.0
495	50.60	999	34191.13	129.88	Realigned reciprocal framework	New Daniellefort	1	United States of America	2016-05-03 12:57	1.0
604	57.20	103	57739.03	110.66	Innovative maximized groupware	East Heatherside	0	New Zealand	2016-03-19 11:09	1.0

Table 2: Data showing outliers in ‘Age’ Column

We plot a box plot to get a visual representation of the outlier situation as seen below

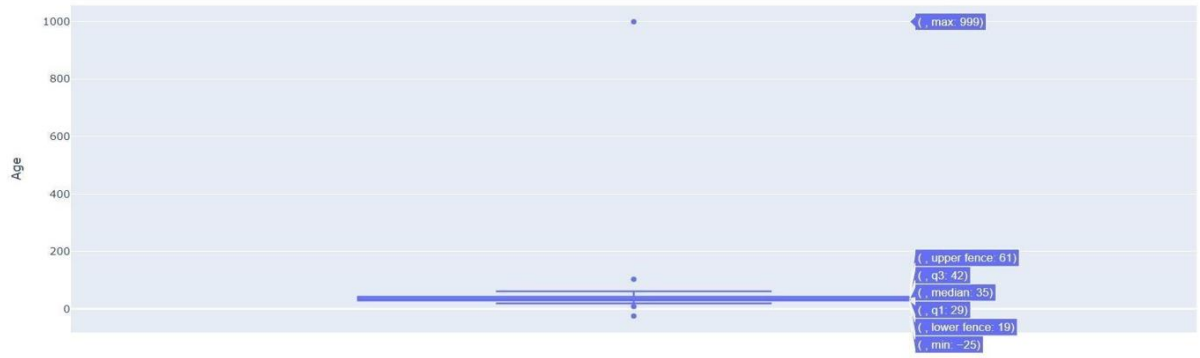


Figure 1: Box Plot 1 showing outliers

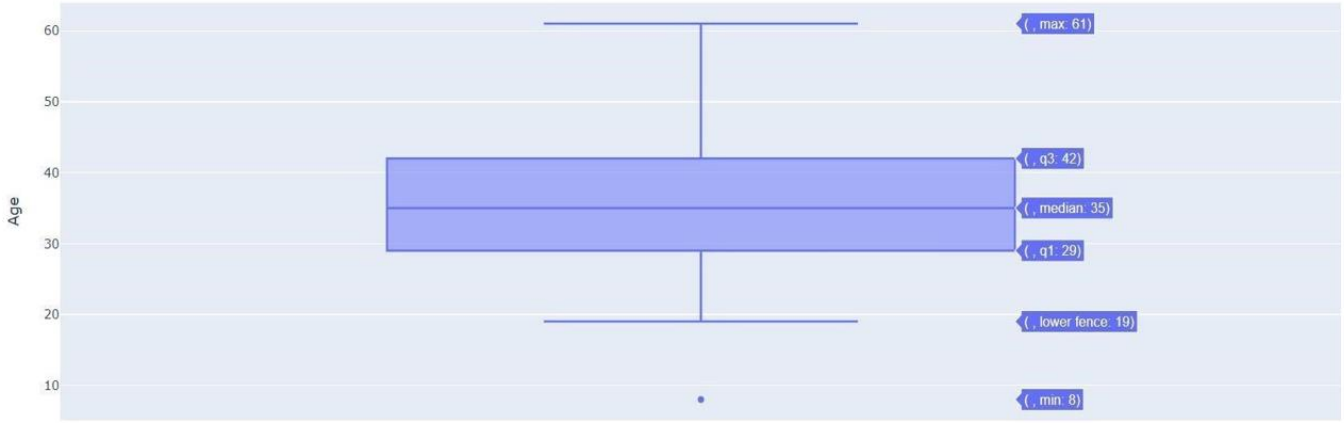


Figure 2: Box Plot 2 after trimming the outliers

By trimming the outliers, we plot the box plot again to verify our results We still however notice one data point below the age of 10. Assuming the average individual with an income or access to funds to be between the ages of 16 to 100 years, we further skim the results to get rid of the outlier.

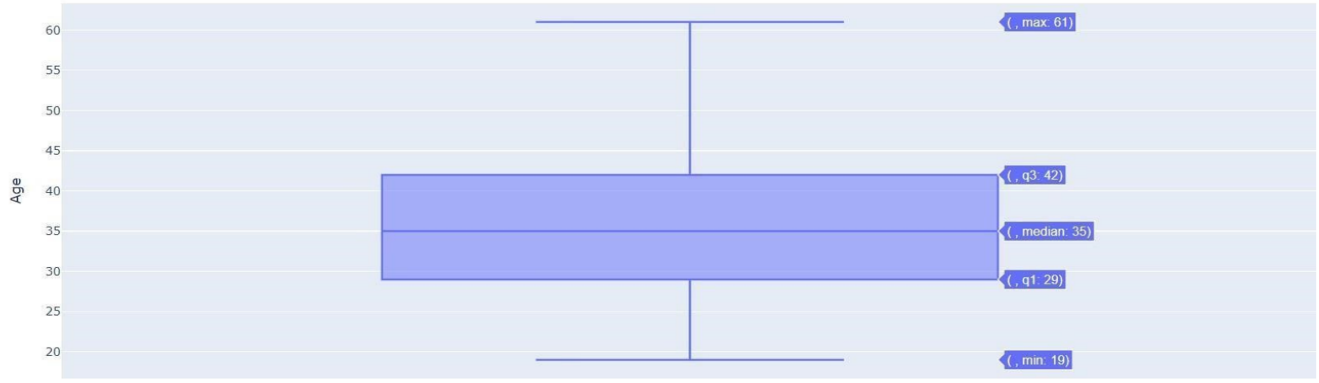


Figure 3: Final Box Plot after removing outliers

Viewing the cleaned dataset

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Male	Clicked on Ad
count	1003.000000	1003.000000	1003.000000	1003.000000	1003.000000	1003.000000
mean	65.005085	36.023928	54955.188166	179.579571	0.482552	0.500499
std	15.856144	8.786979	13434.037845	44.520392	0.499945	0.500249
min	32.600000	19.000000	13996.500000	23.640000	0.000000	0.000000
25%	51.340000	29.000000	46969.130000	138.615000	0.000000	0.000000
50%	68.180000	35.000000	56986.730000	182.650000	0.000000	1.000000
75%	78.555000	42.000000	65441.655000	218.700000	1.000000	1.000000
max	91.430000	61.000000	79484.800000	269.960000	1.000000	1.000000

Table 3: Cleaned Dataset after removing outliers

We want to find the ad which is the most popular based on the number of clicks it has obtained

```
Fundamental zero tolerance solution            2
Operative actuating installation               2
Horizontal hybrid challenge                    2
Cloned 5thgeneration orchestration             1
Reactive impactful challenge                   1
                                              ..
Enhanced zero tolerance Graphic Interface      1
De-engineered tertiary secured line            1
Reverse-engineered well-modulated capability   1
Integrated coherent pricing structure          1
Virtual 5thgeneration emulation                1
```

Table 4: Table showing most popular ads based on number of clicks

From the above observation it looks like 'Fundamental Zero Tolerance solution' was the most popular ad on the website. This could be useful information for the company as it tells them that people visiting their website are most likely interested in purchasing a zero-tolerance solution for their homes or offices.

Exploring the relationship between variables:

Now that we have a fairly clean dataset, we want to know the factors that affect our target variable 'Clicked on Ad'. We initially use the Pearson Coefficient to find the extent of correlation of the other factors.

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Male	Clicked on Ad
Daily Time Spent on Site	1.000000	-0.332547	0.310781	0.510199	-0.018570	-0.748192
Age	-0.332547	1.000000	-0.185785	-0.367615	-0.019215	0.494041
Area Income	0.310781	-0.185785	1.000000	0.342989	-0.002140	-0.477471
Daily Internet Usage	0.510199	-0.367615	0.342989	1.000000	0.017751	-0.777817
Male	-0.018570	-0.019215	-0.002140	0.017751	1.000000	-0.036877
Clicked on Ad	-0.748192	0.494041	-0.477471	-0.777817	-0.036877	1.000000

Table 5: Pearson's Coefficient showing correlation between various factors

As we can see from the table above there is a high negative correlation between the Clicked-on Ad and Daily time spent on site with a value of -0.748804.This indicates that the amount of time that an individual spends on a site doesn't necessarily indicate a higher chance of them clicking on an ad.

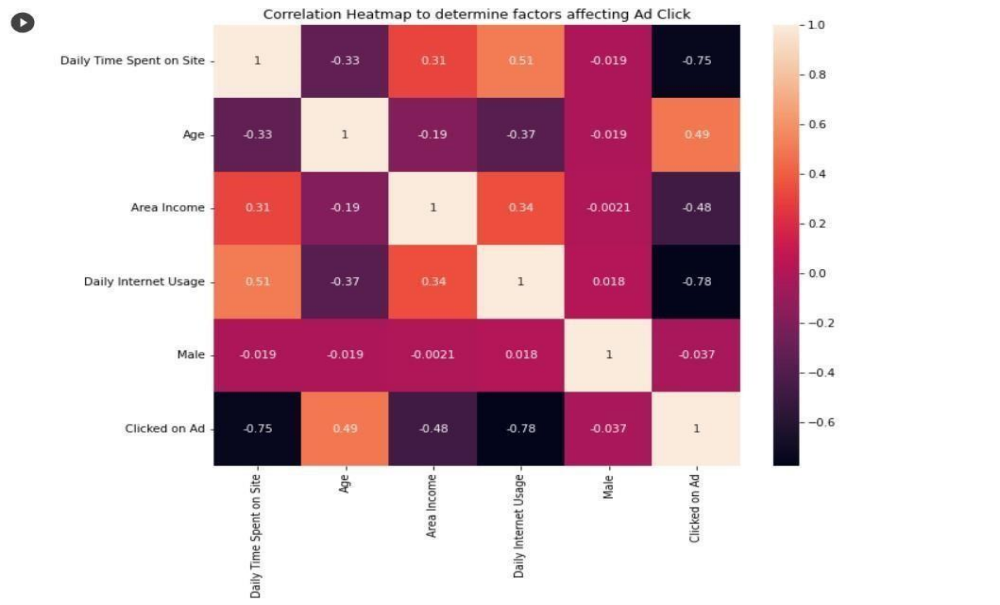


Figure 4 : Heatmap showing the correlation between various factors

A visual representation of the correlation among the various factors is displayed above using a heatmap. The darker the color higher is the negative correlation, as indicated by the values as well and vice versa. As we can see above that there is a high negative correlation between the Daily Internet Usage and Clicked on Ad ,followed by a negative correlation between Area Income and Clicked on Ad. We can also see that there is a positive correlation between daily internet usage and time spent on the site. This can help the company gauge that a person's income or internet usage aren't crucial factors that help determine whether they click on ads displayed on their website, however they can bank on the fact that higher the internet usage of a user ,higher is the possibility of them chancing upon their website and seeing the ad.

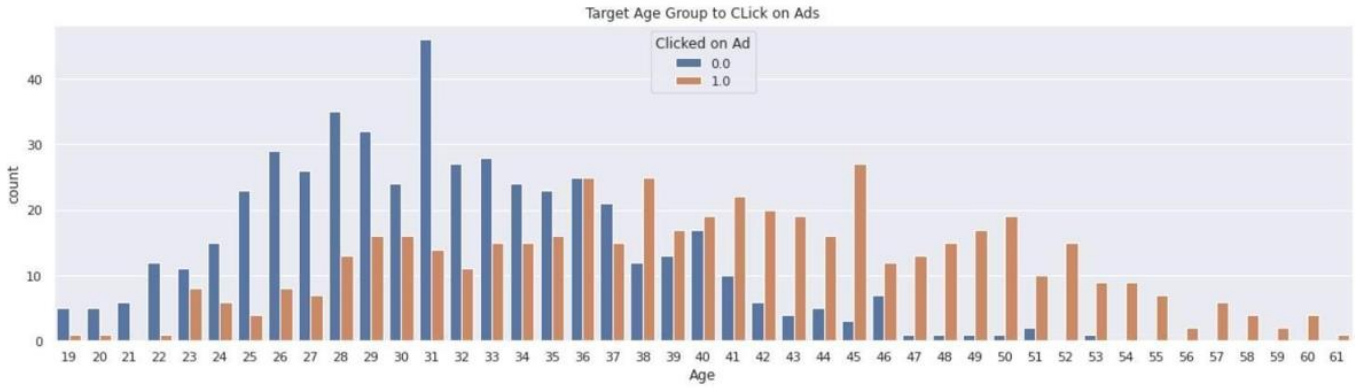


Figure 5: Bar Chart to see the most popular age group to click on ads

From the bar chart we can see the target audience that have a higher chance of clicking on Ads lies between thirties to early fifties. We can see that the extreme age spectrums do not easily click on Ads and hence the marketing strategy of the company can be designed accordingly.

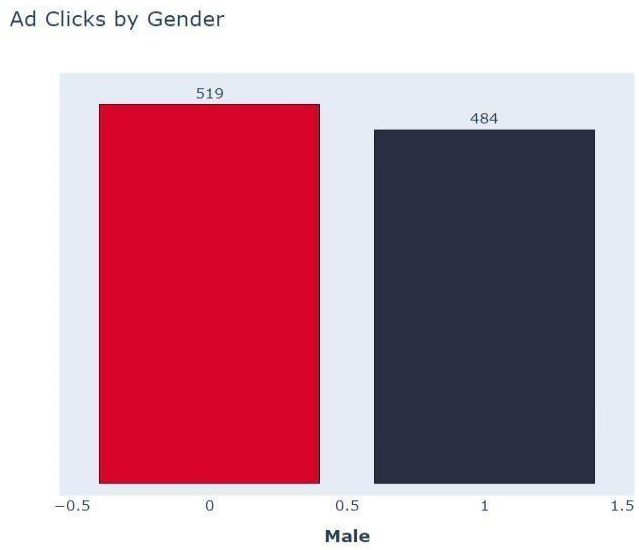


Figure 6: Ad Clicks based on Gender

On comparing the number of clicks based on gender, where 1 indicates Male and 0 indicates female. We can see that more number of females have clicked on the ads in comparison to the males, and hence this helps company owners to help target their female audiences when promoting ads.

Number of Clicks Based on time spent on site

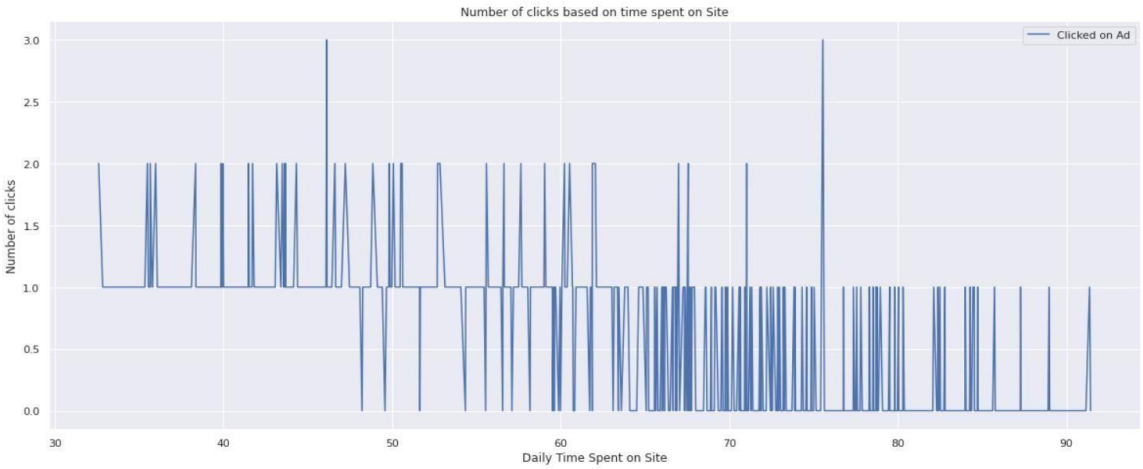


Figure 7: Number of Clicks Based on time spent on site

On visually displaying the data and comparing the number of clicks with the amount of time an average would spend on the site ,we can see that most of the people that clicked on an ad spent about 35 to 55 minutes per day on the site, while those that did not click on ads spent more time on the site.

Most Popular Month Based on Number of Ad Clicks:

[36]

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Timestamp	Clicked on Ad	Date	year	month	day	Time	hour	minute
0	68.95	35	61833.90	256.09	Cloned 5thgeneration orchestration	Wrightburgh	0	Tunisia	2016-03-27 00:53:00	0.0	2016-03-27	2016	3	27	00:53:00	0	53
1	80.23	31	68441.85	193.77	Monitored national standardization	West Jodi	1	Nauru	2016-04-04 01:39:00	0.0	2016-04-04	2016	4	4	01:39:00	1	39
2	69.47	26	59785.94	236.50	Organic bottom-line service-desk	Davidton	0	San Marino	2016-03-13 20:35:00	0.0	2016-03-13	2016	3	13	20:35:00	20	35
3	74.15	29	54806.18	245.89	Triple-buffered reciprocal time-frame	West Terrifurt	1	Italy	2016-01-10 02:31:00	0.0	2016-01-10	2016	1	10	02:31:00	2	31
4	68.37	35	73889.99	225.58	Robust logistical utilization	South Manuel	0	Iceland	2016-06-03 03:36:00	0.0	2016-06-03	2016	6	3	03:36:00	3	36
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1013	72.97	30	71384.57	208.58	Fundamental modular algorithm	Duffystad	1	Lebanon	2016-02-11 21:49:00	1.0	2016-02-11	2016	2	11	21:49:00	21	49
1014	51.30	45	67782.17	134.42	Grass-roots cohesive monitoring	New Darlene	1	Bosnia and Herzegovina	2016-04-22 02:07:00	1.0	2016-04-22	2016	4	22	02:07:00	2	7
1015	51.63	51	42415.72	120.37	Expanded intangible solution	South Jessica	1	Mongolia	2016-02-01 17:24:00	1.0	2016-02-01	2016	2	1	17:24:00	17	24
1016	55.55	19	41920.79	187.95	Proactive bandwidth-monitored policy	West Steven	0	Guatemala	2016-03-24 02:35:00	0.0	2016-03-24	2016	3	24	02:35:00	2	35
1017	45.01	26	29875.80	178.35	Virtual 5thgeneration emulation	Ronniemouth	0	Brazil	2016-06-03 21:43:00	1.0	2016-06-03	2016	6	3	21:43:00	21	43

Table 6: Table showing timestamp split

In order to obtain the most popular month for ad clicks we first split the Timestamp into date,year,month,hour and minute. We then plot a line graph to see the month with the most clicks.

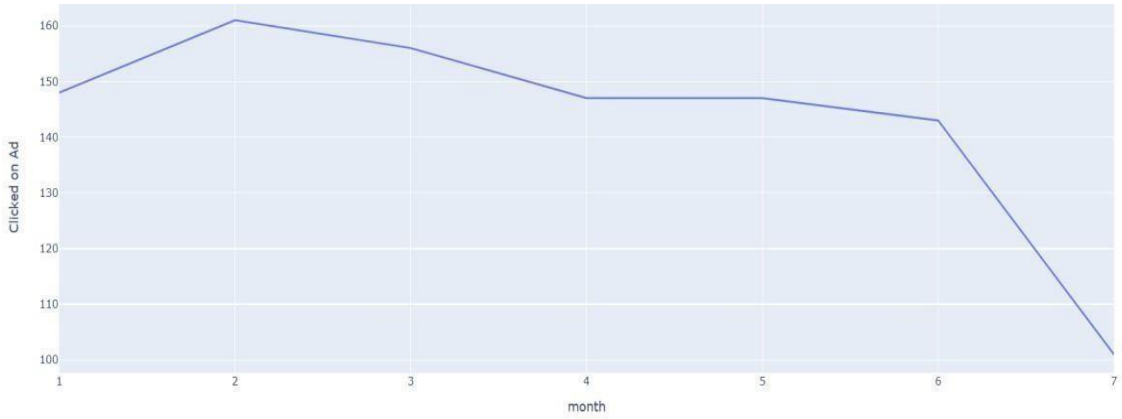


Figure 8: Graph showing the most popular month by maximum ad clicks

The data is available only for the first seven months and we can see that the number of clicks is maximum during the month of February. This can be helpful for the company to market their offers during this time to increase profits based on customer traction.

Highest Number of Ad clicks Based on Day and Time of the Week:

Figure 9: Visuals indicating most popular day of the week and hour for maximum ad clicks.

The above visuals are indicative of the most popular days and times for maximum number of ads clicks throughout the week.0 is indicative of Sundays, and the week begins with Monday-1. We can also see maximum activity during the middle of the week, and weekends with Tuesdays, Saturdays and Wednesdays which seem to be popular days. The time interval that sees a higher number of clicks is between 910am,12pm,6pm and 1 am. These time intervals can be targeted as there is a possibility of people commuting to and from work, during lunch hours and late at night.

Model Exploration and Selection

In this project we plan to implement Logistic Regression, multiple regression, Decision Trees, Random Forest Regression. We need to split the data into training and testing datasets before we can go ahead with any model. We select our target variable as Clicked on Ad and create a random state of 21 to make sure that all methods run in the same state. The models we are implementing is based on accuracy, the higher the accuracy, better the models perform. We split the data into 25% validation set and 75% training dataset.

Logistic Regression:

Logistic regression is a predictive modelling algorithm that is used when the target variable is binary categorical. That is, it can take only two values like 1 or 0. This model can be tried out in our case, as our target variable is 'Clicked on Ad'. Data is fit into the model and is acted upon by a logistic function to predict the target variable.

```
Accuracy Score:
0.9840637450199203
```

Decision Tree:

The essence of decision trees involves splitting the data into different parts and there are some fundamental splitting parameters like Gini index and entropy.

Gini Index	Entropy
Accuracy :	Accuracy :
90.43824701195219	91.63346613545816

Random Forest :

This is usually a two-step procedure where:

1.We build n decision tree estimators which is 100 in our case ,and the trees are built on parameters like depth, tree nodes .

2. We then obtain the average of the tree estimators and that is considered as the final

output

```
Accuracy Score:
0.9920318725099602
```

Multiple Linear Regression:

We use multiple linear regression to predict out output variable or target variable which is 'Clicked on Ad' by using multiple independent variables(age, timestamp etc..).We find the root mean square error and the r2 value to determine how well the model does.

```
r2 socre is 0.7334090360763961
mean_sqrd_error is:  0.06659590467182551
root_mean_squared error of is:  0.2580618233521291
```

Model Selection and Implementation

On exploring the models selected above based on the accuracy scores above we go further in depth considering the Random Forest Classifier **Random Forest Classifier:**

We tune each parameter to explore the changes in out-of-bag score and precision score as well as estimate the possible range of values we could use for tuning the model with multiple parameters at the same time.

Based on various parameters like maximum depth of the trees in the random forest we can obtain more information about the data. The deeper the tree more information or complexity of the data is uncovered. In our case we take the depth range varying from 5 to 14

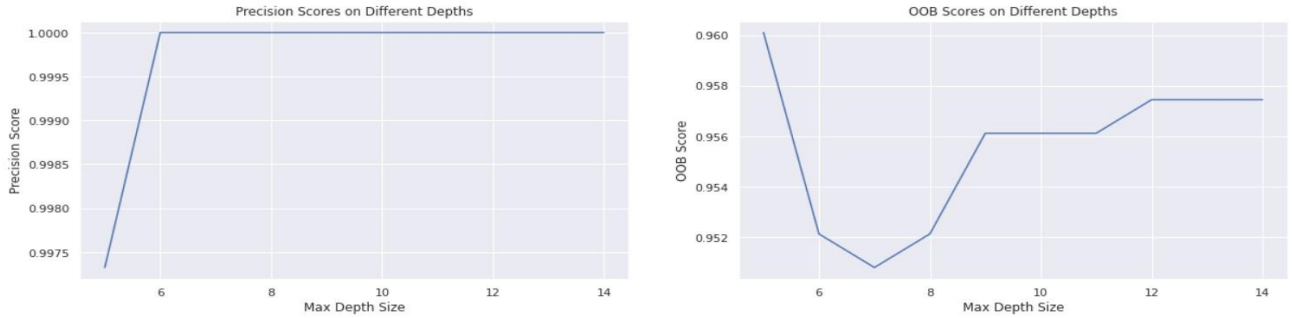


Figure 10: Graphs showing the precision and OOB scores with respect to Max Depth size

Based on the depth size of the trees we can see its variation in terms of precision and OOB score. The OOB (Out of Bag)score provides the coefficient of determination using OOB method, i.e. on 'unseen' out-of-bag data. This score serves as cross-validation loss and keeps a check on overfitting.

Now considering the number of trees as another factor, estimators represent the number of trees in the forest. Usually, the higher the number of trees the better to learn the data. However, adding a lot of trees can slow down the training process considerably, therefore we do a parameter search to find the sweet spot. In our case the optimum number of trees is around 18.

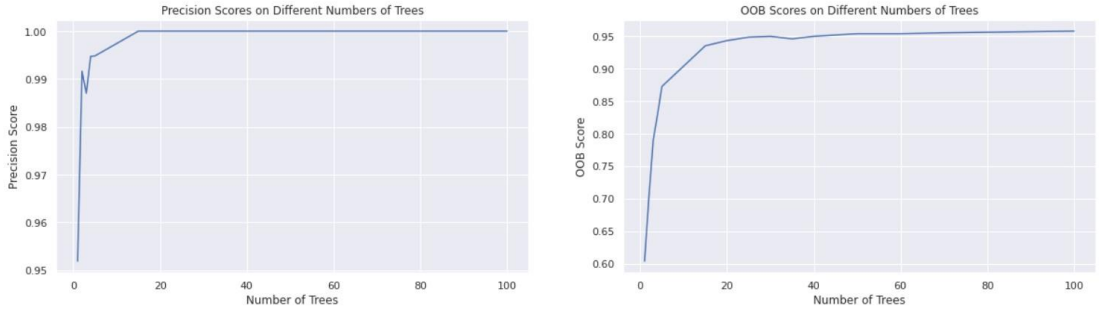


Figure 11: Graphs showing the precision and OOB scores with respect to number of trees

Similarly, we also consider the number of leaves:

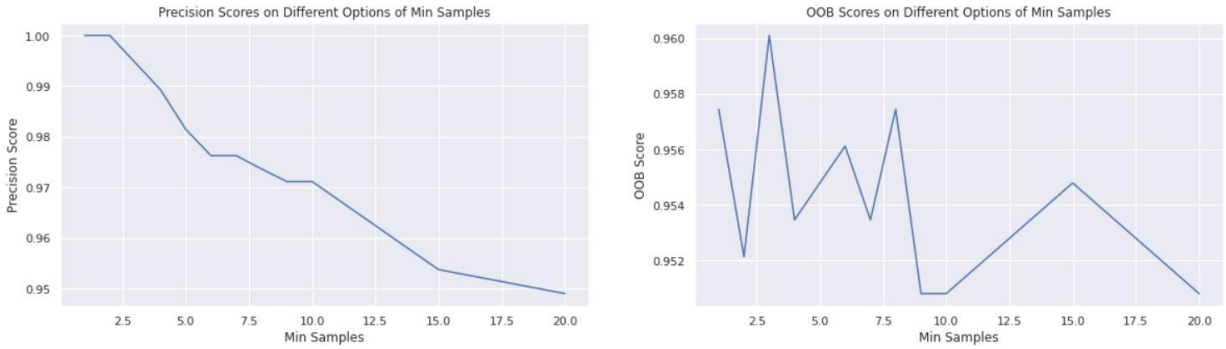


Figure 12: Graphs showing the precision and OOB scores with respect to minimum samples

The best estimators we got from Random Forest Classifier are:

n_estimators = 18,max_depth = 6, min_sample_leaf = 2,max_features = "auto" (features like age, area income, time spent on site, internet usage, gender)and so on.

n_estimators hyperparameter, which is just the number of trees the algorithm builds before taking the maximum voting or taking the averages of predictions. In general, a higher number of trees increases the performance and makes the predictions more stable, but it also slows down the computation.
max_features is the maximum number of features random forest considered for a split
min_sample_leaf determines the minimum number of leaf's required to split an internal node.

OOB score (also called oob sampling), which is a random forest cross-validation method. In this sampling, about one-third of the data is not used to train the model and can be used to evaluate its performance. These samples are called the out-of-bag samples. It's very similar to cross-validation method, but almost no additional computational burden goes along with it.

Performance Evaluation and Interpretation

The performance evaluation and interpretation of the selected Random Forest classifier is done by obtaining the confusion matrix, the precision, evaluation score, F1 and recall.

Factors like internet useage,time spent on site,age,gender,income are some of the factors that are considered and the confusion matrix for the test data tells us whether the ad was actually clicked and correctly predicted (TP),it was actually not clicked, and we predicted negative (TN),it was actually a click but we predicted it to be a non-click (FP) and it was actually a non-click, but we predicted it to be a click(FN).

	Actual Click	Predicted Non -Click
Actual Click	132 (TP)	1 (FP)
Actual Non-Click	1 (FN)	117 (TN)

Table 8: Confusion matrix depicting the click rate

The classification report includes other factors like precision,F1 score and recall

	Precision	Recall	F1	Support
Not Clicked	0.99	0.99	0.99	113
Clicked	0.99	0.99	0.99	118
Accuracy			0.99	251
Macro Average	0.99	0.99	0.99	251
Weighted Average	0.99	0.99	0.99	251

Table 9: Classification report of Random Forest Model

Precision:

The Precision can be explained as the number of classes that we have predicted as clicked, out of which how many are actually clicked(positive). The precision score would be the correctly identified clicks (TP) by the total number of times it was predicted as clicked (TP+FP). The value we obtain is around 99% which is indicative that out of the total ads predicted as clicked 99% of them were clicked.

Recall(Sensitivity):

Recall is determining out of all the clicked(positive) classes how many were predicted correctly, i.e., it is the values that were identified correctly by the total number of values within the class. We get a value of 99% which is indicative that out of all the clicked ads,99% were actually correctly predicted as clicks.

F1 Score:

It is difficult to compare two models with low precision and high recall or vice versa. So, to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean.

The F-measure is the harmonic mean of the precision and recall and in most cases, there is a trade-off between the two. On optimizing to increase one and decrease the other the harmonic mean decreases, however it is ideally greatest when both precision and recall have equal values as seen in our case.

Project Results:

The models used to classify whether a customer has clicked on an ad using relevant factors like internet useage,gender,income,age is some of the factors that are useful in describing the type of audience that clicks(1) on an ad or doesn't(0).The Random Forest classifier gives us the best accuracy of 99% with the recall determining the extent of correct predictions made.

Using the confusion matrix, we can obtain a more accurate result of the funds that need to be invested and the possible losses the company can incur in cases of miscalculation. The Random Forest model helps us reduce the risk of overfitting and considers multiple random factors by creating sub trees and goes down to level of leaf's

and nodes to obtain a higher level of accuracy. The precision and recall of 99% indicates that the out of all the times the click rate was predicted,99% of the times it was done correctly.

Project Impact:

The Advertising industry is extremely competitive, whatever you see sells and hence visibility and relevance is one of the most important aspects of advertising. To be able to grab a user's attention, which could lead to possible sales is something every company wishes for. In this competitive industry it's important to be able to know who your target audience is in order to maximize relevance of the Ad. Narrowing down factors and using the best models to get accurate results would be an important step both for the organization as well as for customers interested in buying a particular product, which results in a more efficient retail experience. It not only helps to save time and money but also increase customer base. In a world where online shopping is taking over, and where everything can be bought with a click of a button it is important for companies to maximize their efforts in personalizing ads for their customers in order for them to be on top of their game. With companies starting to use machine learning and data mining methods to increase profits and make better business decisions, using methods like the Random Forest and logistic regression can help them achieve a clearer picture of the needs of their audience.