Predicting Advertisement Clicks

Abstract

Advertisements are important in society; they can inform people of an issue and influence decisions. Businesses and corporations spend massive amounts of money trying to promote themselves in many different forms of media successfully. In this report, we will be analyzing an advertising dataset of a marketing agency to develop a machine learning algorithm that predicts if a user will click on an online advertisement.

1. Business problem

1.1 Objective:

Develop a machine learning model with high accuracy in predicting whether a user will click on an advertisement. The primary objective is to maximize the model's predictive performance. Identify the most influential features or factors that contribute to ad clicks. Understanding these factors can provide valuable insights for advertisers to optimize their ad campaigns.

1.2 Challenges:

The dataset may exhibit class imbalance, where one class (e.g., clicks) is significantly more prevalent than the other (e.g., no-clicks). This imbalance can lead to biased models that favor the majority class and difficulty in accurately predicting the minority class.

1.3 Real World Impact:

The dataset may exhibit class imbalance, where one class (e.g., clicks) is significantly more prevalent than the other (e.g., no-clicks). This imbalance can lead to biased models that favor the majority class and difficulty in accurately predicting the minority class.

2. Dataset

2.1 Datasets:

Source of the dataset is from Kaggle:

https://www.kaggle.com/datasets/gabrielsantello/advertisement-click-on-ad

2.2 Data Fields:

- Daily Time Spent on Site
- Age
- Area Income: Avg income of the geographical area of the user
- Daily Internet Usage
- Ad Topic Line
- City
- Male
- Country
- Timestamp
- Clicked on Ad (0 or 1)

2.3 Data Understanding & Tools:

Data comes from a Kaggle competition so it can be downloaded directly for the solution but if we want to productionize the live data we might have to make a data pipeline for the same. Cloud solutions and SQL queries for data pipelines are very commonly seen in companies which can be used effectively. For this particular instance we can use Pandas and Numpy libraries to process the data as we have data in CSV format.

3. Key metric (KPI)

3.1 Business Metric:

In a machine learning project focused on predicting whether a person clicks on an advertisement, there are several business metrics that you may consider to evaluate the performance and impact of the model. These metrics help assess the effectiveness of the advertising campaign and the return on investment (ROI) from ad clicks.

3.2 Available metrics:

There are multiple metrics to choose from:

1. Accuracy
2. Precision
3. Recall (Sensitivity)
4. Classification Report

5. Confusion Matrix

3.3 Metric selection and Reasoning:

The metric which we will use for this problem is Classification Report.

Comprehensive Summary: The Classification Report provides a comprehensive summary of key classification metrics, including precision, recall, F1 score, and support (the number of occurrences of each class in the dataset). This holistic view allows for a thorough assessment of the model's performance across multiple dimensions.

Insight into Class Imbalance: In scenarios where the dataset exhibits class imbalance (e.g., more instances of one class than the other), the Classification Report helps in understanding how the model performs for each class individually. It reveals whether the model's predictions are biased towards the majority class or if it can effectively identify instances of the minority class.

Balanced Evaluation: Unlike simple accuracy, which can be misleading in imbalanced datasets, the Classification Report considers metrics such as precision, recall, and F1 score, which provide a balanced evaluation of the model's performance across both positive and negative classes.

Interpretability: The metrics included in the Classification Report (precision, recall, and F1 score) are intuitive and easy to interpret. They offer insights into the model's ability to make correct predictions (precision), capture true positive instances (recall), and strike a balance between the two (F1 score).

Actionable Insights: By analyzing the Classification Report, stakeholders can identify areas for model improvement and take actionable steps to address any shortcomings. For example, if the model exhibits low recall for the positive class, efforts can be directed towards improving the model's sensitivity to positive instances.

Overall, the Classification Report provides a robust and informative framework for evaluating the performance of a logistic regression model, offering insights that can inform decision-making and model refinement processes in advertising campaigns.

4. Real-world challenges and constraints:

Data Quality and Quantity:

Challenge: Availability of high-quality data with sufficient quantity is essential for building accurate predictive models. However, obtaining clean and comprehensive datasets can be challenging due to data collection limitations, privacy concerns, and data biases.

Constraint: Limited access to relevant data sources, incomplete data, or data that is not representative of the target population can constrain the model's predictive power.

Imbalanced Data Distribution:

Challenge: Imbalanced datasets, where the number of positive and negative instances (clicks and non-clicks) is skewed, can lead to biased models and poor generalization.

Constraint: Dealing with class imbalance requires specialized techniques such as resampling methods, cost-sensitive learning, or algorithmic adjustments, which may add complexity to the modeling process.

Feature Engineering and Selection:

Challenge: Identifying and selecting informative features from raw data that effectively capture user behavior and preferences is critical for model performance. However, feature engineering requires domain expertise and careful consideration of feature relevance, redundancy, and interpretability.

Constraint: Limited availability of domain experts or resources for feature engineering may restrict the model's ability to leverage valuable insights from the data.

6.References:

1. https://www.kaggle.com/code/imprime/logistic-regression-with-ad-clickdataset/notebook

2. https://www.kaggle.com/datasets/gabrielsantello/advertisement-click-on-ad

3. https://medium.com/@sinethpathirana/predicting-advertisement-clicksc60fc46cb121

4. https://encyclopedia.pub/entry/41132

EDA and Feature Extraction

Exploratory Data Analysis (EDA):

Target Variable Analysis: Examine the distribution of the target variable (click vs. noclick) to understand the class balance. Imbalanced classes may require special treatment during modeling.

Feature Analysis: Analyze the distribution and properties of individual features (e.g., demographics, browsing behavior, ad attributes). Use summary statistics, histograms, box

plots, and correlation matrices to understand feature distributions, identify outliers, and detect patterns.

Feature Interactions: Investigate relationships between features and the target variable. Explore how different features interact with each other and how they may influence ad clicks.

Temporal Trends: If the dataset contains timestamps, analyze temporal trends in ad clicks. Explore patterns related to time of day, day of week, or seasonality to identify temporal dependencies.

Visualization: Visualize relationships between variables using scatter plots, heatmaps, and time series plots. Visual inspection can reveal insights and guide feature selection.

Feature Extraction:

Demographic Features: Extract demographic information such as age, gender, location, and device type from user data. These features can provide insights into the audience's characteristics and preferences.

Behavioral Features: Capture user behavior features such as past click history, browsing patterns, session duration, and engagement metrics. Behavioral features can indicate user intent and propensity to click on ads.

Contextual Features: Include contextual features such as ad type, ad position, ad content, and time of ad impression. Contextual features provide information about the ad context and its relevance to the user.

Content-Based Features: Extract features related to ad content, such as keywords, ad keywords, and sentiment analysis of ad text. Content-based features can help capture the relevance and appeal of the ad to the user.

Interaction Features: Create interaction features by combining multiple features or performing feature transformations. For example, combine demographic and behavioral features to capture user segments with specific behaviors.

1.1 Datasets: -

These datasets contain respective columns and the descriptions are as follows:-

1. **Daily Time Spent on Site –** Number of hours spent by users on a website.
2. **Age** - Age of the user.
3. **Area Income –** Income of the people Living in the area.
4. **Daily internet usage –** the Daily amount of time a user uses the site.
5. **Male –** Gender of the User.
6. **Clicked on Ad –** Whether or not a person clicks on a particular Ad.

1.2 Libraries :-

We have used multiple libraries to perform EDA and Machine learning :

Importing Libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [ ]:
```
from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import log_loss
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
```

In [56]:
```
from sklearn.ensemble import RandomForestClassifier
```

Description of these libraries are as follows:-

1. Google Drive to Google Colab Mounting this library is used so that we can files present on our google drive.

2. Pandas for Data frame operations

3. NumPy for Numeric operations

4. Datetime for Date & time operations

5. Matplotlib and Seaborn are Data Visualization libraries

6. Logistic regression is the model on which we will be predicting our data.

7. Random Forest Classifier is another model on which we will be predicting our data on and then we will evaluate which of the 2 performs better.

EDA: -

We will start with the understanding of the data.

```
In [32]: df.head()
```

Out[32]:

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Male	Clicked on Ad	Month	Day	Hour
0	4.247781	3.583519	11.032223	5.549426	0	0	3	27	0
1	4.397285	3.465736	11.133754	5.271819	1	0	4	4	1
2	4.255187	3.295837	10.998543	5.470168	0	0	3	13	20
3	4.319486	3.401197	10.911576	5.508943	1	0	1	10	2
4	4.239454	3.583519	11.210346	5.423098	0	0	6	3	3

```
In [4]: df.describe()
```

Out[4]:

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Male	Clicked on Ad
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000
mean	65.000200	36.009000	55000.000080	180.000100	0.481000	0.50000
std	15.853615	8.785562	13414.634022	43.902339	0.499889	0.50025
min	32.600000	19.000000	13996.500000	104.780000	0.000000	0.00000
25%	51.360000	29.000000	47031.802500	138.830000	0.000000	0.00000
50%	68.215000	35.000000	57012.300000	183.130000	0.000000	0.50000
75%	78.547500	42.000000	65470.635000	218.792500	1.000000	1.00000
max	91.430000	61.000000	79484.800000	269.960000	1.000000	1.00000

Finding NA and Null values: - There are no null values in the dataset .

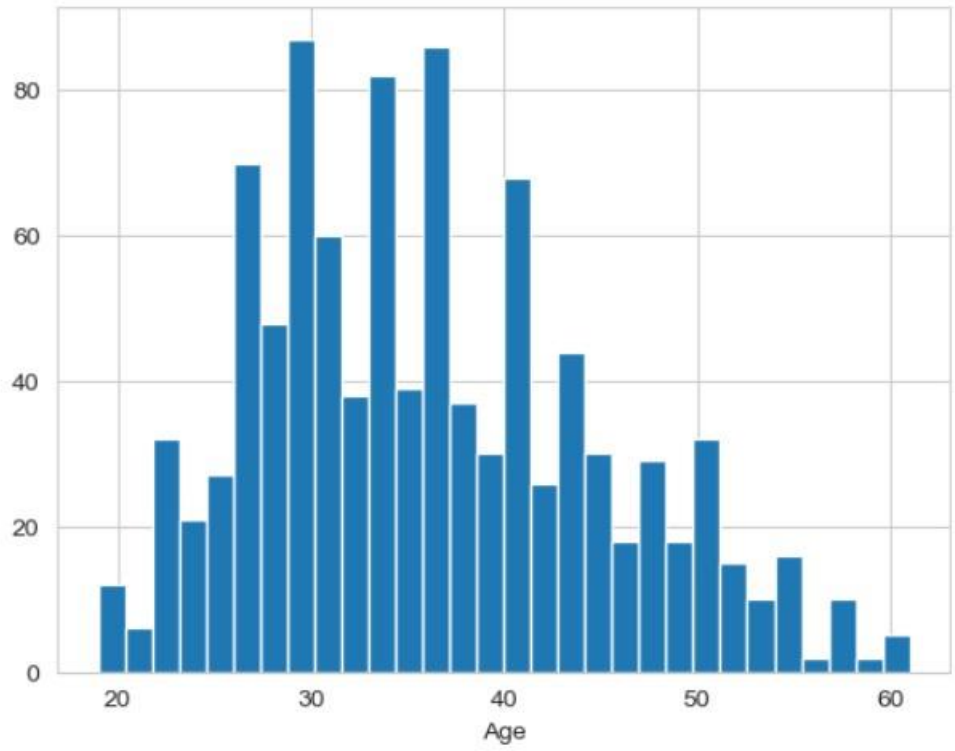
```
In [7]: df.isnull().sum()

Out[7]: Daily Time Spent on Site    0
        Age                         0
        Area Income                 0
        Daily Internet Usage        0
        Ad Topic Line               0
        City                        0
        Male                        0
        Country                     0
        Timestamp                   0
        Clicked on Ad               0
        dtype: int64
```

3. Visualization :-

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Male	Clicked on Ad
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000
mean	65.000200	36.009000	55000.000080	180.000100	0.481000	0.50000
std	15.853615	8.785562	13414.634022	43.902339	0.499889	0.50025
min	32.600000	19.000000	13996.500000	104.780000	0.000000	0.00000
25%	51.360000	29.000000	47031.802500	138.830000	0.000000	0.00000
50%	68.215000	35.000000	57012.300000	183.130000	0.000000	0.50000
75%	78.547500	42.000000	65470.635000	218.792500	1.000000	1.00000
max	91.430000	61.000000	79484.800000	269.960000	1.000000	1.00000

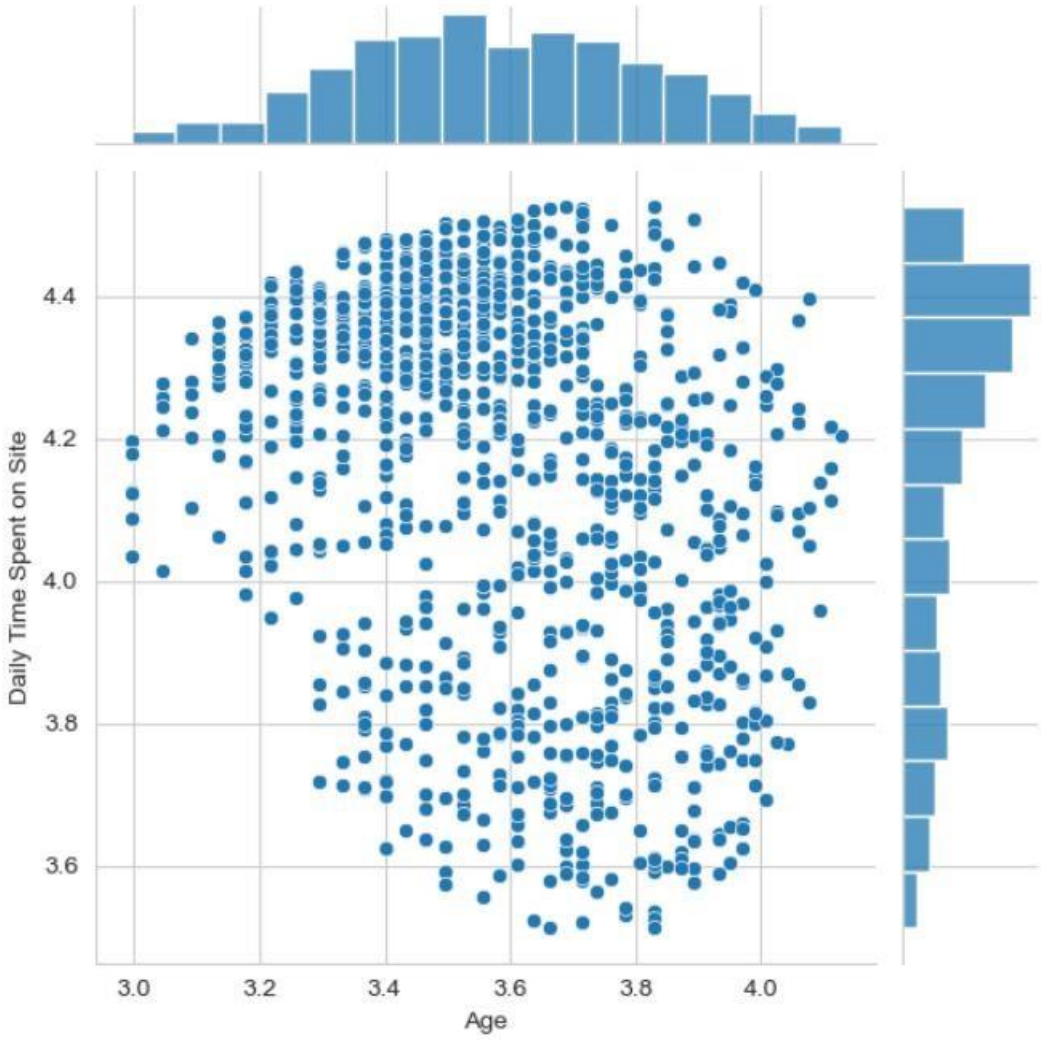
The table displayed above provides a summary of descriptive statistics derived from our dataset. A notable feature of interest is the Area Income. The smallest area income in this dataset is $13, 996.50 and the maximum area income is $79, 484.80. This provides us with an insight into the distribution of socioeconomic

statuses of all the site visitors. Based on the numbers observed under Daily Time Spent on Site, the minimum duration spent on the site is 32 minutes, while 91 minutes is recorded as the highest. Intuitively, we can speculate that this website is somewhat popular among people. In addition, the mean age of the visitors is 36 years old. The minimum age of the visitors is 19 years, while the maximum age is 61 years. This is a sign that this site attracts or targets young to middle-aged adults. However, it is also evident that the site did not favor one gender, as there is a similar percentage of both men and women visiting the site.



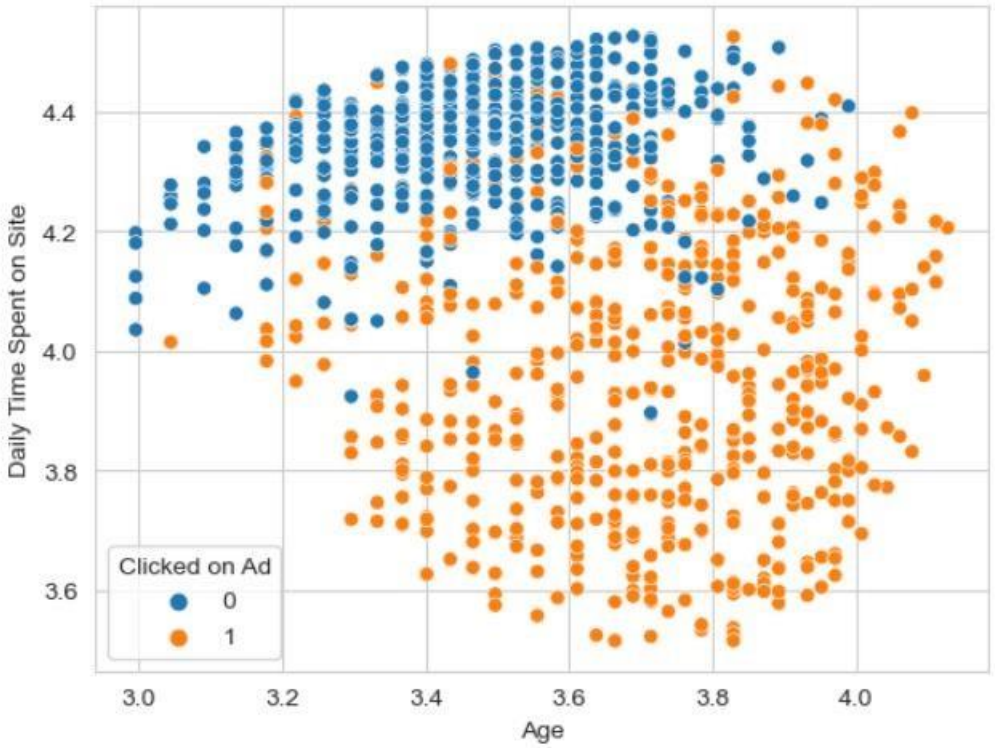
The above count plot created by seaborn displays the number of clicks sorted by age. The graphs imply that approximately a great number of people clicking on advertisements are in the age group of 29–50 years. The common consensus on

the number of people using electronic devices such as personal computers and smartphones in the age groups over 50 is low, hence the number of clicks will also be less.

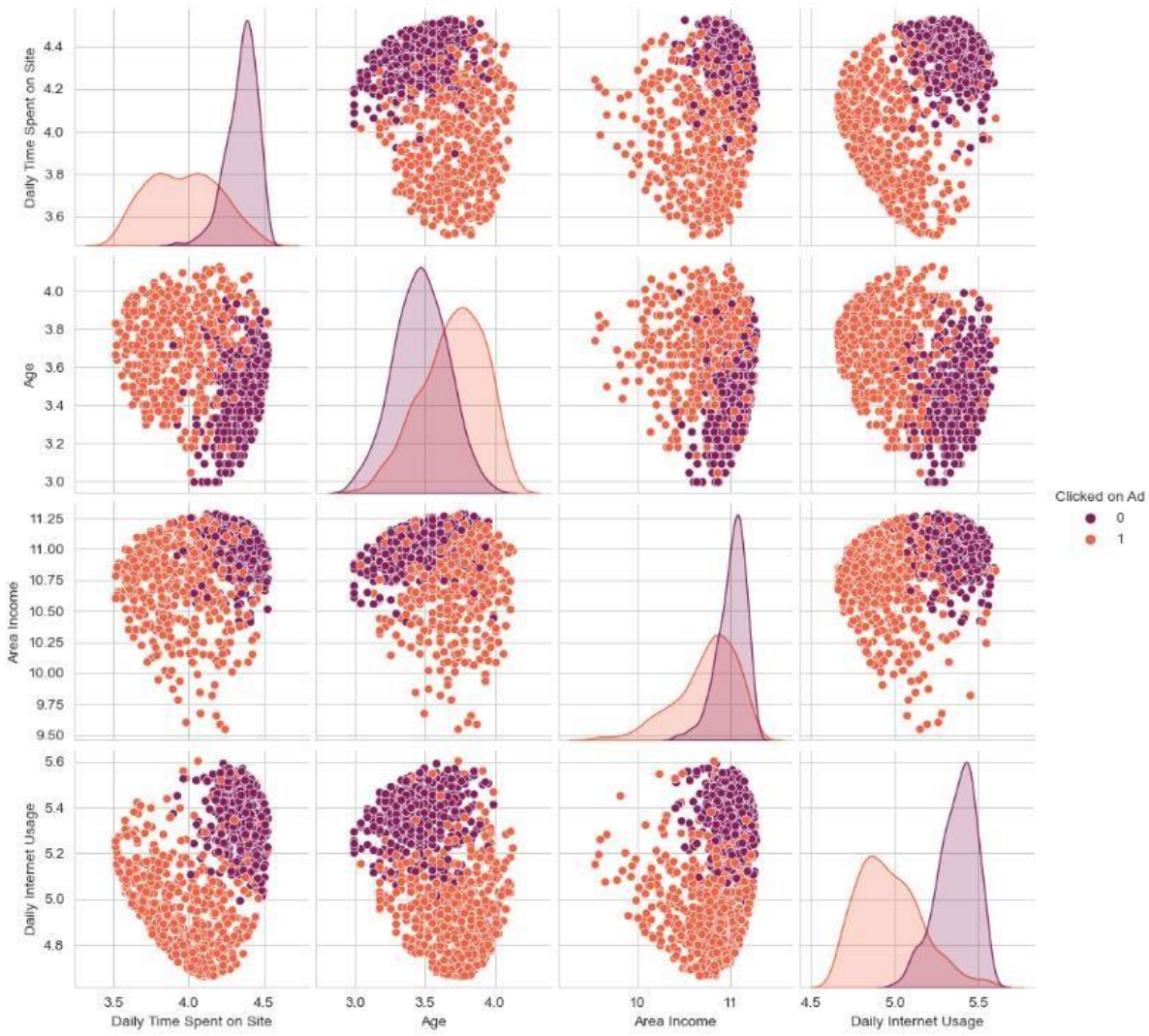


In the above scatterplot we can see that there's a strong correlation between the "Daily Time Spent on Site" and "Age" column.

We can see that more people aged between 30 to 40 are spending more time on site daily.



We can see that more people aged between 20 to 40 are spending more time on site daily but less chances of them to click on the ads.



The pair plot displayed above helps us visualize trends and abnormalities in our data. Only numerical variables are displayed here for further analysis.

There appears to be some correlation between Age, Daily Internet Usage, Daily Time Spent, and Area Income on Site. As for the plots with Sex and Month, these plots do not seem to provide us any intuition as to how we can use them to predict the target variable of Ad Clicks.

Data Analysis

Model Selection

For this classification project, I have chosen to utilize two machine learning models: Logistic Regression and Random Forest Classifier. Each of these models offers distinct advantages and is well-suited for different aspects of our project. Below is a brief introduction to each model and their significance in our project:

Logistic Regression:

Logistic Regression is a popular linear model used for binary classification tasks. Despite its name, it's not used for regression but for classification. It estimates the probability that a given input belongs to a particular class. The model works by fitting a logistic curve to the training data, which allows it to predict the probability of the occurrence of a binary outcome.

Significance:

1. Interpretability: Logistic Regression provides easily interpretable results, making it valuable for understanding the relationship between features and the target variable.

2. Efficiency: It's computationally efficient, especially with large datasets, making it suitable for real-time applications or when computational resources are limited.

3. Feature Importance: Logistic Regression can quantify the importance of each feature in predicting the target variable, aiding in feature selection and model understanding.

Random Forest Classifier:

Random Forest Classifier is an ensemble learning method based on decision trees. It constructs multiple decision trees during training and outputs the mode of the classes predicted by individual trees. Each tree is trained on a random subset of the training data and a random subset of features, adding randomness to the model.

Significance:

1. Robustness: Random Forest is robust to overfitting and noise in the data due to the ensemble of trees and feature randomness, making it suitable for complex datasets with high dimensionality.

2. Accuracy: It often yields high accuracy in classification tasks, even without extensive hyperparameter tuning.

3. Feature Importance: Random Forest can provide estimates of feature importance, allowing us to identify the most relevant features for classification.

We get the following results after fitting the Logistic Regression model:-

	precision	recall	f1-score	support
0	0.92	0.99	0.95	162
1	0.99	0.92	0.95	168
accuracy			0.95	330
macro avg	0.95	0.95	0.95	330
weighted avg	0.95	0.95	0.95	330

```
array([[160,   2],
       [ 14, 154]], dtype=int64)
```

Below are the results for Random Forest Classifier Model:-

```
Accuracy: 0.9545454545454546

Confusion Matrix:
 [[156   6]
 [  9 159]]

Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.96      0.95       162
           1       0.96      0.95      0.95       168

    accuracy                           0.95       330
   macro avg       0.95      0.95      0.95       330
weighted avg       0.95      0.95      0.95       330

log loss: 1.638347881323507
```

Evaluating the results of the 2 Models:

Looking at the provided classification reports for Logistic Regression and Random Forest models, both models seem to perform similarly in terms of accuracy and overall F1-score.

However, there are slight differences in precision and recall values.

Considering precision, recall, and F1-score for both classes, it appears that the Random Forest model has slightly higher precision and recall for both classes compared to Logistic Regression. Therefore, based on the provided metrics, the Random Forest model may be considered slightly better for this specific dataset.

Classification Report:

Precision: For class 0 (not clicking on the ad), the precision was 95%, and for class 1 (clicking on the ad), it was 96%. This indicates that the model has a high precision for both classes, meaning that when it predicts an instance as positive (clicking on the ad), it is correct around 95-96% of the time.

Recall: For class 0, the recall was 96%, and for class 1, it was 95%. This suggests that the model effectively captures the majority of positive instances for both classes.

F1-score: The F1-score for both classes was 95%, indicating a good balance between precision and recall for both classes.

Confusion Matrix:

The users that are predicted to click on commercials and the actually clicked users were 156, the people who were predicted not to click on the commercials and actually did not click on them were 159.

The people who were predicted to click on commercials and actually did not click on them are 6, and the users who were not predicted to click on the commercials and actually clicked on them are 9.

We have only a few mislabeled points which is not bad from the given size of the dataset.

Log Loss Report:

Log loss values typically range from 0 to positive infinity. A log loss of 1.6383 indicates that, on average, the model's predicted probabilities diverge from the true labels by approximately 1.6383 bits per sample.

Conclusion

In conclusion, the Random Forest model performed well across all evaluation metrics, demonstrating high accuracy, precision, recall, and F1-score for both classes. With an accuracy of 95% and balanced performance across multiple metrics, the Random Forest model appears to be effective in predicting whether a user will click on an ad based on the given features.

Being able to accurately predict 98% of user ad clicks based on the sample population is an impressive accomplishment. However, that means about roughly 2% of predictions are wrong, which can result in some financial loss. More specifically we are addressing potential false positives, where the model believes they will click but does not. Furthermore, being able to generalize our algorithm to a much broader population will prove to be difficult for a variety of reasons. Firstly, the entire dataset is based on a group of people visiting a specific website and comparing the results to people who visit other sites may be unwise, due to potential unseen variables influencing results related to website browsing.

A future study should analyze the type of device the consumer uses (i.e. laptop, smartphone, tablet etc.). Doing this may allow the model to discover new areas to invest in for businesses and companies. Furthermore, working with geographical categorical variables such as city and country should be done to examine factors that may influence the likelihood of someone clicking on an advertisement.