Paper Title* (use style: paper title)

Jayaweera S. T.
Dept. of Electrical and Information
Engineering
Faculty of Engineering
University of Ruhuna
Galle, Sri Lanka
jayaweera_st_e22@engug.ruh.ac.lk

Gunathilake A. P. B.
Dept. of Electrical and Information
Engineering
Faculty of Engineering
University of Ruhuna
Galle, Sri Lanka
gunathilake_apb_e22@engug.ruh.ac.lk

***Abstract*— In the modern world, predictive models are necessary to growth efficiency in advertising. Digital advertising plays a major function in modern-day marketing. Prompting the want for predictive models to enhance efficiency is a crucial factor. In this undertaking, data science and advertising had been merged to leverage system learning to activate that want. To supply personalized and relevant content material to their target market, predictive analytics are important. In this undertaking, the take a look at especially specializes in logistic regression and choice trees models to forecast ad clicks. Transparent and flexible version exploration is guided by means of ethical considerations.**

Keywords*— *Machine Learning, Predictive Modeling, Predictive Analytics, Logistic Regression, Decision Trees, Online Ads, Ad Click Prediction, Ad Click-through Rate

I. INTRODUCTION

In the modern world marketing, digital advertising is an essential factor. For that reason, advanced predictive models are required. Here, machine Learning is utilized to forecast the user engagement with day-to-day online ads. That is the main objective of this project. For this purpose, both the data science and marketing are used. It is achievable to predict the click of an add through machine learning algorithms. Logistic Regression and Decision Tree algorithms are the algorithms used. The application of these algorithms to explore the user behavior is discussed here. As mentioned in the abstract, ethical consideration is of utmost importance. In the digital marketing context, this ethical behavior is important to maintain when dealing with user data. The ultimate objective of the project is to provide advertisers with powerful tools for efficient and effective marketing. It will benefit in targeted content delivery and tactical decision-making within the digital ecosystem.

II. METHODOLOGY

(A) Data:

The available dataset used in this project is originating from the Kaggle website. The dataset consists of various features. These features are related to user engagement with online ads. They are, ‘Daily Time Spent on Site', 'Age', 'Area Income', 'Daily Internet Usage', 'Ad Topic Line', 'City', 'Male', 'Country', 'Timestamp'. In addition, the dataset has feature called 'Clicked on Ad'. In this project, 'Clicked on Ad' is considered as the resulting feature. The output variable, 'Clicked on Ad,' says whether a user clicks on an ad. The dataset has 1000 data points with no missing values. Each entry has a unique instance of user interaction with digital ads. The diverse range of features provides a rich source for training predictive models.

(B) Pre-Processing:

In this project, a comprehensive pre-processing pipeline was implemented.

In this methodology, first of all the necessary dataset file and libraries are imported. The dataset file was named as 'advertising.csv'. It was ensured that the dataset was already clean. It had no missing data or duplicated data. Then, several steps were followed to prepare the data for the machine learning models. The steps include outlier handling, timestamps transforming, feature engineering. For categorical data handing, One-hot encoding was used. As it appeared to be complicated, label encoding was used. And the correlation between the features were also analyzed. For predicting ad clicks, Logistic Regression and Decision Tree algorithms were chosen. They are well known for being straightforward and effective. At the end, fine tuning was done to increase the accuracy of the models.

IMPORTED LIBRARIES:

At the beginning of the project, necessary Python libraries were imported. They are, pandas, numpy, seaborn, scikit-learn. ‘pandas’ library is used for data manipulation and analysis. ‘numpy’ library is used for numerical operation computing. ‘seaborn’ library is used for data visualization built on Matplotlib. ‘scikit-learn’ library is used for machine learning functionalities.

Dataset (advertising.csv):

The dataset is obtained from the website known as Kaggle. The dataset is a structured collection of information related to online advertising. The dataset contains various 10 features including the targeting feature. The other features are considered to predict the targeted feature.

Checking and Handling Missing Data:

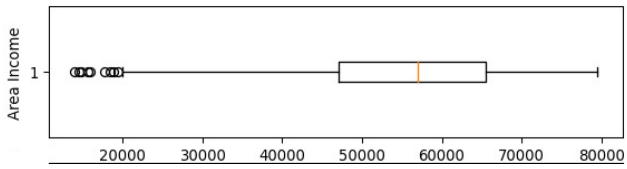
At the first instance, missing values were checked using ‘pandas’ library. In this dataset, there were no missing values. By that means, it was ensured that the dataset was complete. Next, it was necessary to identify and address the duplicate values.

Checking and Handling Duplicate Data:

At second, the dataset was checked for duplicate data points. This process was also done using the ‘pandas’ library. In this dataset, there were no duplicate values. By that means, it was ensured that the dataset is consisting unique data points. Therefore, it was clear that the dataset is clean.

Checking Outliers through Data Exploration and Visualization:

After checking for the cleanliness of the dataset, box plots were drawn. Box plots were drawn for each numerical feature in the dataset. A box plot is a special type of diagram that shows the quartiles in a box and the line extending from the lowest to the highest value.



Dropping Outliers:

By examining the box plots, some outliers were identified. An outlier is a single data point that goes far outside the average value of a group of statistics.

Particularly in the 'Area Income' feature, outliers were handled. Bounds based on the interquartile range (IQR) were established for this. This helped maintain the reliability of the dataset for subsequent analysis.

Splitting Timestamps:

The 'Timestamp' feature is a combination of two features. They are date and time. Therefore, 'Timestamp' feature was transformed into several features. After the transformation, 'Year,' 'Month,' 'Day,' 'Hour,' 'Minute,' and 'Second' features were extracted. This enhanced the dataset by incorporating time-related information.

Feature Engineering:

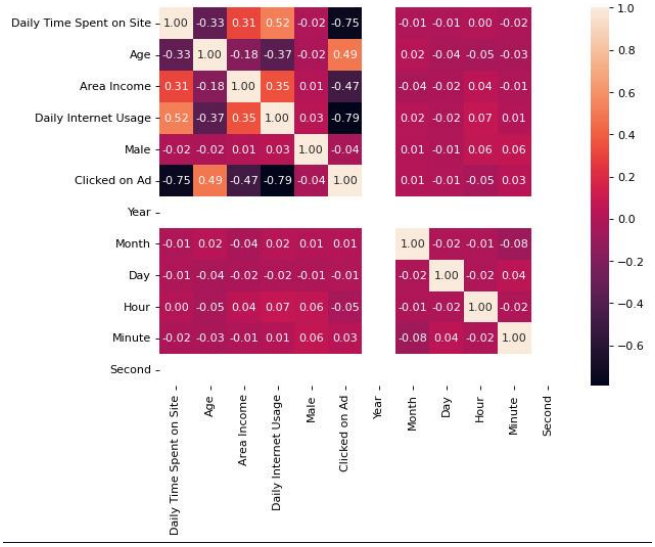
Several features were engineered during the process. After the extraction, 'Timestamp' feature was dropped. In addition, 'Year,' 'Second,' and 'Ad Topic Line' columns were subsequently dropped. 'Year' and 'Second' were dropped as the data points were same all over the columns. 'Ad Topic Line' were dropped for the simplification of the dataset manipulation.

Label Encoding:

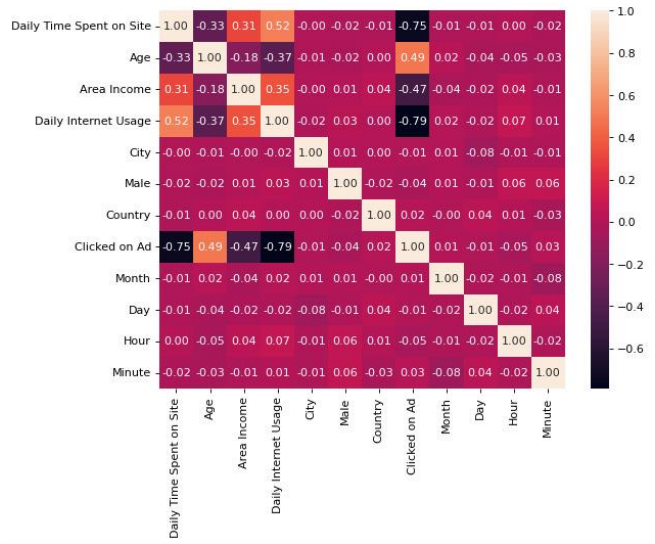
Then, the categorical data were addressed. There had to be applied an encoding method. One-hot encoding was complicated as the number of categorical data was huge. Therefore, label encoding was applied for the 'City' and 'Country' features. For this purpose, scikit-learn's Label Encoder was used. It converted categorical variables into a numerical format. So, the dataset was suitable for machine learning algorithms.

Correlation Analysis:

After by considering all the numerical values, a correlation matrix heatmap was generated. It was helpful to analyze relationships between numeric features. This analysis helped with the decision to drop certain features. Highly correlated features usually perform well on the training data. But they do not perform well on new and unseen data.



Therefore, dropping highly correlated features enhances model efficiency. This reduces the risk of overfitting and making the model more robust.



Normalization:

Normalization in machine learning is the process of translating data into the range [0, 1]. In this project, Normalization was implemented through the MinMaxScaler from scikit-learn. This ensures that all features were on a similar scale. It prevents dominance by specific features during model training. This method is helpful to improve the performance of models.

(C) Algorithms:

In this machine learning project, two machine learning algorithms were used. Namely, Logistic Regression and Decision Tree algorithms were trained for prediction of ad clicks. Logistic Regression is a simple and effective algorithm for binary classification tasks. It models the probability of an ad click based on input features. Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. Decision Tree is selected for its ability to capture complex relationships in the data. It makes decisions by recursively splitting the dataset into subsets based on feature conditions. This algorithm is simple to understand and to interpret. Both algorithms are implemented using the scikit-learn library in Python. Logistic Regression is particularly suited for cases where a predictions and classification problems are involved. Decision Trees can handle non-linear relationships and interactions between features. They are commonly used in operations research and operations management. They are also used to solve classification problems, regression problems or as a method to predict continuous outcomes from unforeseen data. The choice of these algorithms is based on their suitability for the classification.

(D) Implementation:

In the implementation, scikit-learn library in Python was utilized for the two models. The code follows a structured approach. As mentioned previously, it starts with data preprocessing. And then steps into the stage of training the machine learning models. The dataset was then split into training and testing sets for model evaluation. For Logistic Regression, a grid search was conducted to find the optimal hyperparameters. Considered hyperparameters were regularization and penalty terms. Similarly, grid search was conducted for the Decision Tree as well. Considered hyperparameters were criteria for splitting, maximum depth, and minimum samples per leaf. The chosen hyperparameters were determined based on their impact on model accuracy. The resulting models were evaluated using accuracy, precision, recall, and F1-score.

Logistic Regression Algorithm:
https://www.geeksforgeeks.org/understanding-logistic-regression/

Decision Tree Algorithm:
https://www.geeksforgeeks.org/decision-tree/

Post-Processing:

In the post-processing phase, hyperparameter tuning was employed to enhance the performance of the models. For Logistic Regression, a Grid Search approach was implemented to systematically explore different combinations of hyperparameters, such as regularization strength (C) and penalty terms. This process aimed to identify the optimal set of hyperparameters that maximizes the model's accuracy. Similarly, for the Decision Tree model, a Grid Search was conducted to fine-tune hyperparameters like the splitting criterion, maximum depth, minimum samples per leaf, and others. The selected hyperparameters were determined based on their impact on model accuracy during cross-validation. Hyperparameter tuning is a crucial step to optimize model performance and generalization, ensuring the models perform well on unseen data.

Results
Logistic Regression Results:

The Logistic Regression model exhibited considerably good performance in ad-click prediction. During the evaluation, the model achieved an accuracy of approximately 76.88%. The classification report further informed the model's precision, recall, and F1-score metrics The classification report highlighted improvements with a weighted average precision of 78%. The confusion matrix showed a balanced nature of the model's predictions, with 90 true negatives, 63 true positives, 13 false positives, and 33 false negatives. These results indicate a satisfactory ability to predict who click the ads and those who don't. Considering the overall results, the effectiveness of Logistic Regression is satisfactory.

Decision Tree Results*:*

The Decision Tree model also exhibited reasonable performance in ad-click prediction. But it was low in comparison to the Logistic Regression model. During the evaluation, the model achieved an accuracy of approximately 67.34%. The classification report further outlined the model's precision, recall, and F1-score metrics. It indicated that the model exhibited balanced performance with lower overall accuracy. The model demonstrated macro average precision with a weighted average precision of 67%. The confusion matrix showed a balanced nature of the model's predictions, with 72 true negatives, 62 true positives, 31 false positives, and 34 false negatives. Therefore, the Decision Tree model provided reasonable predictions. With that results, it was suggested for further refinement to enhance performance. And it was noted that Decision Tree algorithm well performs for larger datasets. In overall, these results indicate a very reasonably satisfactory ability to predict who click the ads and those who don't.

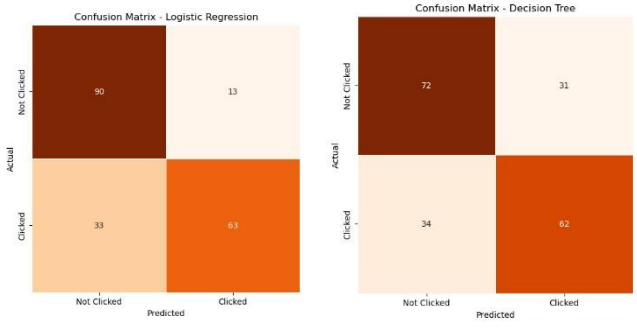
Discussion:

The Logistic Regression and Decision Tree models were implemented in this project to do the prediction. However Logistic Regression outperformed in all the accuracy, precision, and recall. Logistic Regression's better generalization ability makes it suitable for this task. Ethical aspects were taken into consideration, emphasizing responsible data usage in digital advertising. The choice between these models depends on the specific requirements and priorities. These findings highlight the importance of choosing models aligned with specific application needs.

Ethical Aspects:

Ethics play a critical and paramount role in the context of predictive modeling. This research recognizes the ethical implications of using machine learning in digital advertising. As personal data is used, privacy concerns are highly considered. Finding the right balance between customizing content and safeguarding user privacy is crucial. Transparency and fairness in algorithmic decision-making are also vital ethical considerations. We are dealing with a world of predictive analytics for digital advertising. So, it is importance to adhere to ethical standards, ensuring user trust, and prioritizing data protection. It is crucial to highlight the significance of upholding moral principles digital advertising field.

Conclusion:



This project successfully navigated the intersection of data science and marketing. It utilized machine learning to predict user engagement with online ads. For this utilization, Logistic Regression and Decision Tree models were employed. The comprehensive analysis included data exploration, preprocessing, algorithm implementation, and post-processing with hyperparameter tuning. A robust performance was ensured by subjecting the models to cross-validation. For advertisers looking for effective, ethical, and transparent methods in digital advertising, the findings offer valuable information. This research emphasizes the significance of responsible data usage and ethical considerations in leveraging machine learning for digital marketing.

REFERENCES

[1] "Data Preprocessing in Machine Learning: A Beginner's Guide." Accessed: Jan. 21, 2024. [Online]. Available: https://www.simplilearn.com/data-preprocessing-in-machine-learning article

[2] guest_blog, "10 Techniques to Solve Imbalanced Classes in Machine Learning (Updated 2024)," Analytics Vidhya. Accessed: Jan. 21, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal with-class-imbalance-in-machine-learning/

[3] N. Tanwar and K. F. Rahman, "Machine Learning in liver disease diagnosis: Current progress and future opportunities," IOP Conf. Ser. Mater. Sci. Eng., vol. 1022, no. 1, p. 012029, Jan