Improving Advertising Return on Investment (ROI) through Data Analysis: A Study on Ad Click Prediction

Business Problem

Online advertising is crucial for reaching target audiences and promoting business growth in the current digital era (Bharadkar et al., 2021). Advertisers invest significant resources in developing and distributing advertisements, and it is crucial to comprehend which advertisements are effective at generating user engagement. This project's primary business challenge is to predict whether a user will click on an ad (Yes or No) based on various factors, such as demographic data, online behavior, and ad-specific details. The insights obtained from this prediction will assist advertisers in optimizing their advertising strategies and allocating resources more effectively.

Datasets

The dataset, named "advertising.csv," from Kaggle comprises a thorough compilation of data related to individuals who participated in a survey. The dataset includes a variety of factors, including demographic data, online behavior, and ad-specific. The dataset consists of 1000 observations with 10 variables.

Data Description

Daily.Time.Spent.on.Site: The amount of time a user spends on the website (numeric).

Age: The age of the user (integer).

Area.Income: The user's annual income in their geographical area (numeric).

Daily.Internet.Usage: The amount of time a user spends on the internet daily (numeric).

Ad.Topic.Line: The title or topic of the advertisement (character).

City: The city where the user is located (character).

Male: Gender of the user (0 for female, 1 for male; integer).

Country: The user's country (character).

Timestamp: The timestamp when the user interacted with the ad (character).

Clicked.on.Ad: The target variable to be predicted (0 for No, 1 for Yes; integer).

Research Question

What factors influence a user's decision to click on an advertisement, and can we accurately forecast ad clicks using these factors?

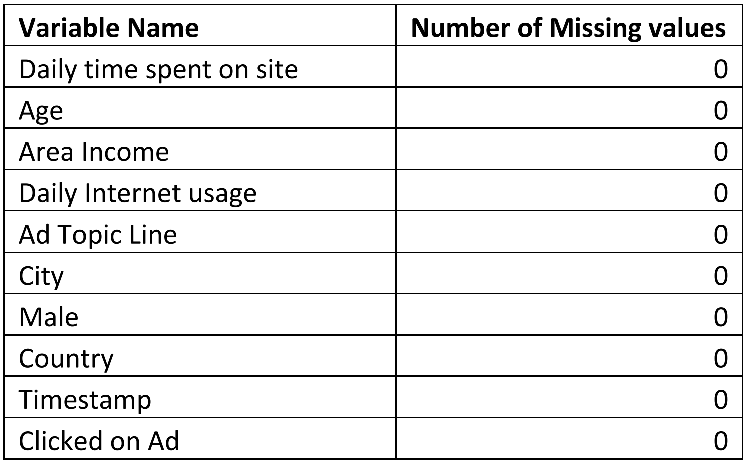
Methods

Data Preprocessing

Dealing with Missing Values

The data was checked for missing values and no missing values were reported.

Table 1: Table of Missing Values



Variable Name	Number of Missing values
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

The process of encoding categorical variables

Categorical variables are of considerable importance in the analysis of socioeconomic data. In order to render them compatible with machine learning techniques, we employed the technique of one-hot encoding. This methodology transforms categorical variables into a numerical representation by generating binary columns for each category. By employing this approach, the categorical data's integrity is retained while ensuring its compliance with the prediction models.

Standardization/Normalization

The numerical features present in the dataset underwent either standardization or normalization. The standardization process guarantees that numerical qualities possess a mean value of zero and a standard deviation of one. On the other hand, normalization involves scaling numbers to a predetermined range, such as the interval [0, 1]. Ensuring fair comparisons of numerical features is of utmost importance, as it prevents traits with greater scales from exerting undue influence over the modeling process.

Model Construction

The investigation of algorithms

This project's scope encompassed an examination of many machine learning algorithms to determine the optimal model for forecasting the likelihood of clicking an ad. The algorithms under consideration were:

1) Logistic Regression is a fundamental (Cokluk, 2010) linear model commonly employed for binary classification applications.

2) Random Forests, a versatile (Cornelius & Shanthini, 2023) ensemble method, can effectively capture intricate correlations within the dataset.

3) Decision Tree which make decisions by recursively splitting the data based on the most informative features, aiming to maximize information gain or minimize impurity.

4) KNN, an algorithm that classifies or predicts a data point by considering the class labels of its k-nearest neighbors in the feature space.

The investigation of various algorithms facilitated the identification of the model that exhibited the highest level of performance in terms of predictive accuracy, precision, recall, and F1-score.

Parameter Tuning

Parameter tuning, referred to as hyperparameter tuning, entails the systematic selection of optimal hyperparameters for a machine learning algorithm. Hyperparameters refer to predetermined settings not derived from the data but established before the training process. These settings possess the potential to (Zaki et al., 2021) exert a substantial influence on the model's performance. The parameter tuning process entails systematically exploring various combinations of hyperparameters and assessing the model's performance using cross-validation.

The objective is to identify the hyperparameters that yield optimal model performance on the validation dataset.

Feature Selection

Selecting relevant features from a given dataset is called feature selection. The process of feature selection holds significant importance in the study. In order to ascertain the characteristics that substantially contribute to the model's predictive capability, we employed feature importance scores. The scores provide valuable insights into the relative impact of different characteristics on the model's output. Features that possess greater significance scores were deemed to be more pertinent in the prediction of whether a user will click on an ad (Yes or No).

Selecting Relevant Features

Identifying key characteristics is a crucial task, and it is equally imperative to carefully choose a subset of features that optimizes predicted accuracy while limiting the potential for introducing bias. Five best features were selected from the dataset using this method.

Table 2: Table of 5 best-selected features

Variable Name	***Variable Type***
Daily Time spent on site	*Float64*
Age	*Float64*
Area Income	*Float64*
Daily Internet Usage	*Float64*
Country_Ethiopia	*Float64*

Feature Importance

Feature importance is frequently employed in machine learning to ascertain the individual contributions or levels of significance of features (also known as variables or attributes) within a prediction model. Understanding the features that exert the most influence on the predictions made by the model is beneficial. Age, daily time spent on site, daily internet usage and area income are the (see **Appendix A**) most important features in predicting clicking ad.

Model Evaluation

In order to appropriately evaluate the performance of our models, the dataset was divided into two distinct subsets: a training set and a testing set. The training set was utilized to train the model, while the testing set was set aside to evaluate its performance on data that had not been previously encountered (Vrigazova, 2021). This methodology enables us to assess the model's capacity to extrapolate to novel instances.

Performance metrics are quantitative measures used to evaluate and assess the performance of a system, process, or individual. These metrics provide objective data. The models were assessed using a variety of performance indicators, which encompassed:

1) Accuracy refers to measuring accurately anticipated cases about the total number of cases.

2) Precision refers to the capacity to categorize instances as positive examples accurately.

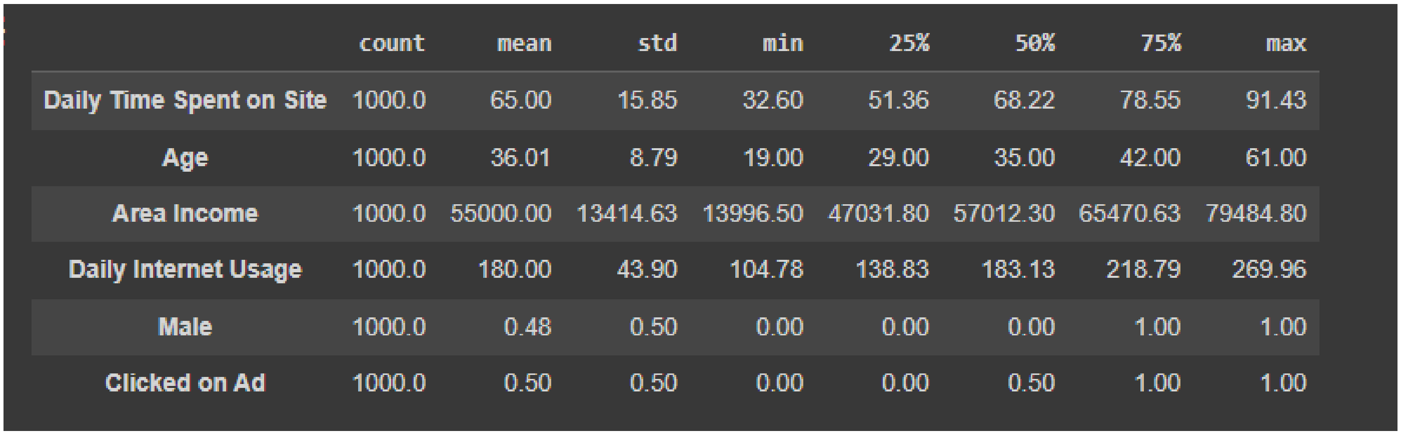
3) Recall refers to the cognitive capacity to accurately identify and retrieve all pertinent cases or instances relevant to a certain context or situation.

4) The F1-score is a metric that quantifies the balance between precision and recall by calculating their harmonic mean.

The metrics thoroughly evaluate the models' predictive capacities, allowing us to make educated judgments regarding their appropriateness for our depression prediction task.

Analysis

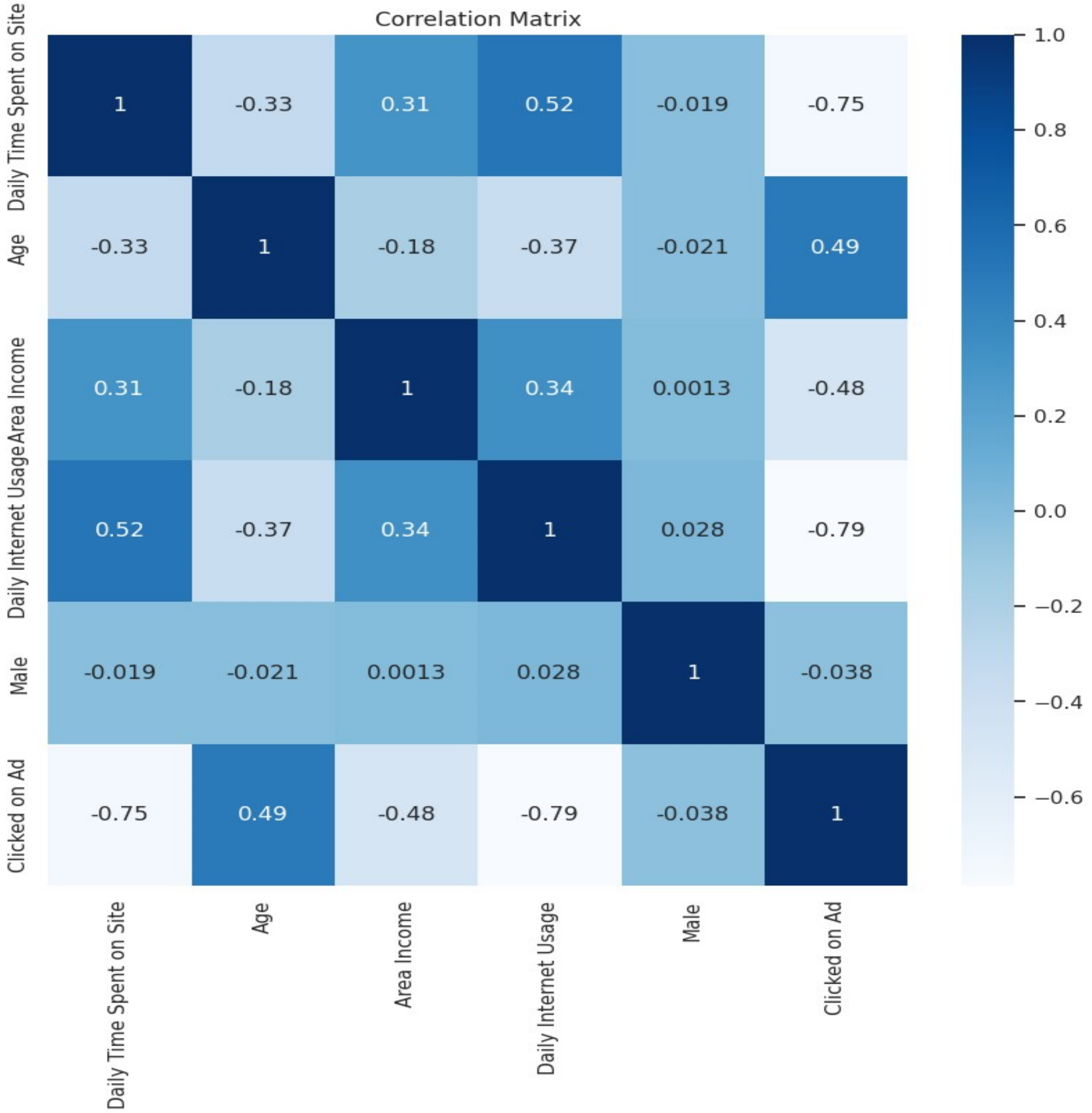
a) Descriptive statistics



	count	mean	std	min	25%	50%	75%	max
Daily Time Spent on Site	1000.0	65.00	15.85	32.60	51.36	68.22	78.55	91.43
Age	1000.0	36.01	8.79	19.00	29.00	35.00	42.00	61.00
Area Income	1000.0	55000.00	13414.63	13996.50	47031.80	57012.30	65470.63	79484.80
Daily Internet Usage	1000.0	180.00	43.90	104.78	138.83	183.13	218.79	269.96
Male	1000.0	0.48	0.50	0.00	0.00	0.00	1.00	1.00
Clicked on Ad	1000.0	0.50	0.50	0.00	0.00	0.50	1.00	1.00

The table above gives a summary of descriptive statistics including mean, standard deviation, minimum and maximum of all the numeric variables.

b) ***Correlation Analysis***



Positive values (greater than 0) indicate a positive correlation between the variables. This means that as one variable increases, the other tends to increase as well. For example, "Daily Time Spent on Site" and "Daily Internet Usage" have a positive correlation of approximately

0.52, suggesting that as daily time spent on the site increases, daily internet usage also tends to increase.

Negative values (less than 0) indicate a negative correlation between the variables. This means that as one variable increases, the other tends to decrease. For instance, "Area Income" and "Clicked on Ad" have a negative correlation of approximately -0.48, indicating that high income earners are less likely to click on ads in this dataset.

Daily Time Spent on Site vs. Clicked on Ad (-0.75):

This correlation is quite strong and negative, indicating a significant inverse relationship.

People who spend more time on the site are less likely to click on ads.

Age vs. Clicked on Ad (0.49):

This is a positive correlation, though not extremely strong.

Generally, as the age of users increases, they are more likely to click on ads.

Area Income vs. Clicked on Ad (-0.48):

There is a moderate negative correlation.

Users with higher area income are less likely to click on ads.

Daily Internet Usage vs. Clicked on Ad (-0.79):

This is a strong negative correlation.

Users who spend more time on the internet are less likely to click on ads.

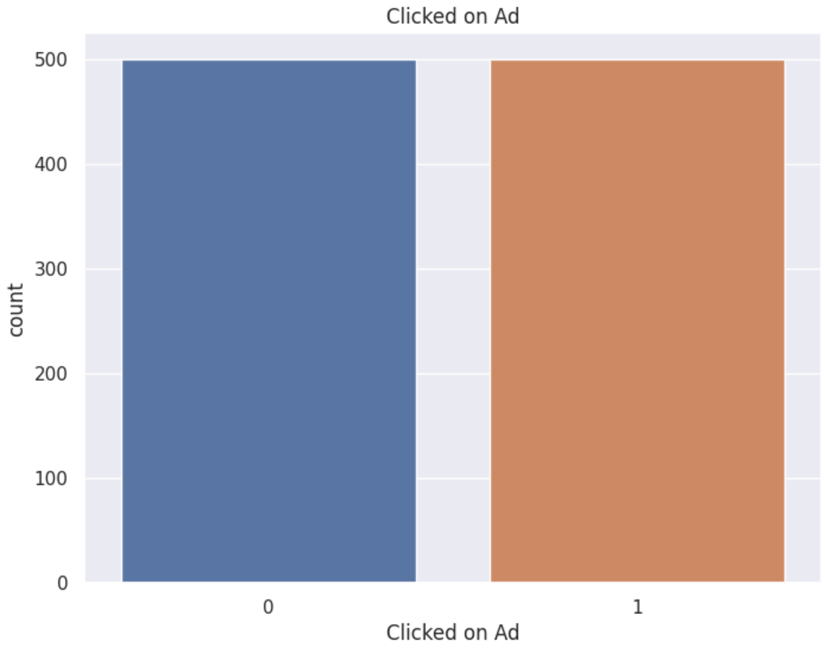
Male vs. Clicked on Ad (-0.038):

There is a very weak negative correlation.

Being male or female doesn't strongly influence the likelihood of clicking on ads.

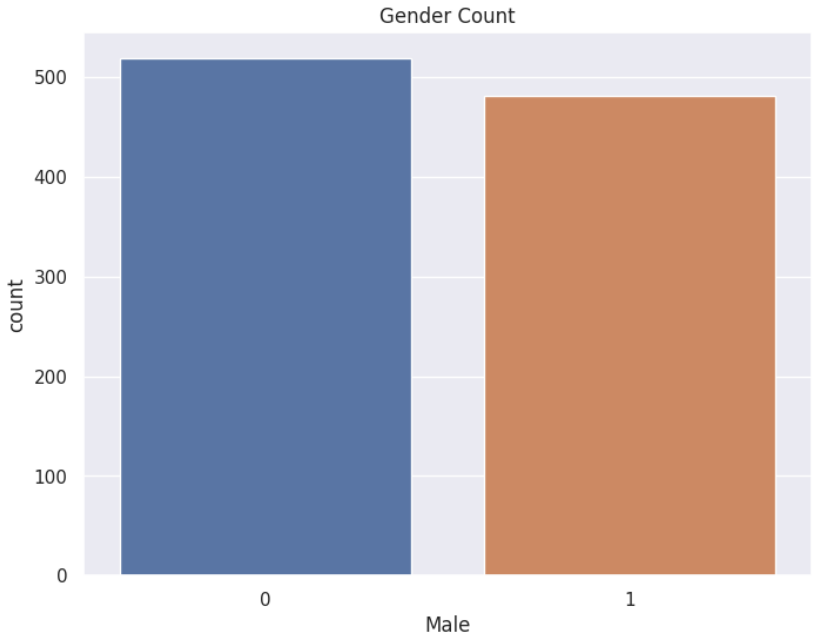
c) Exploratory Data Analysis

Figure 1: Distribution of Clicked on Ad



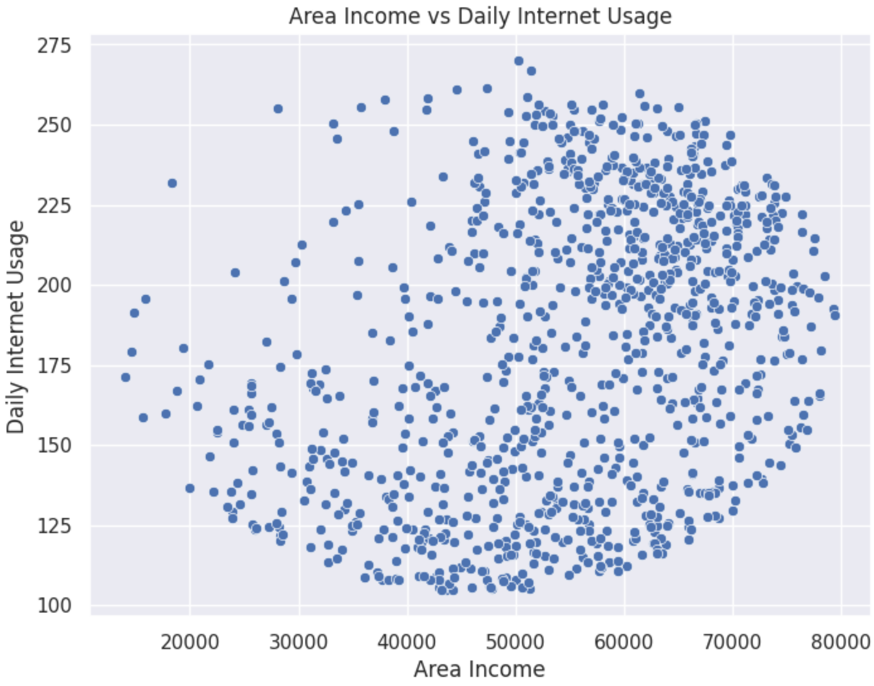
The distribution of participants who clicked on ad and those who did not click on ad are equal.

Figure 2: Distribution of Gender



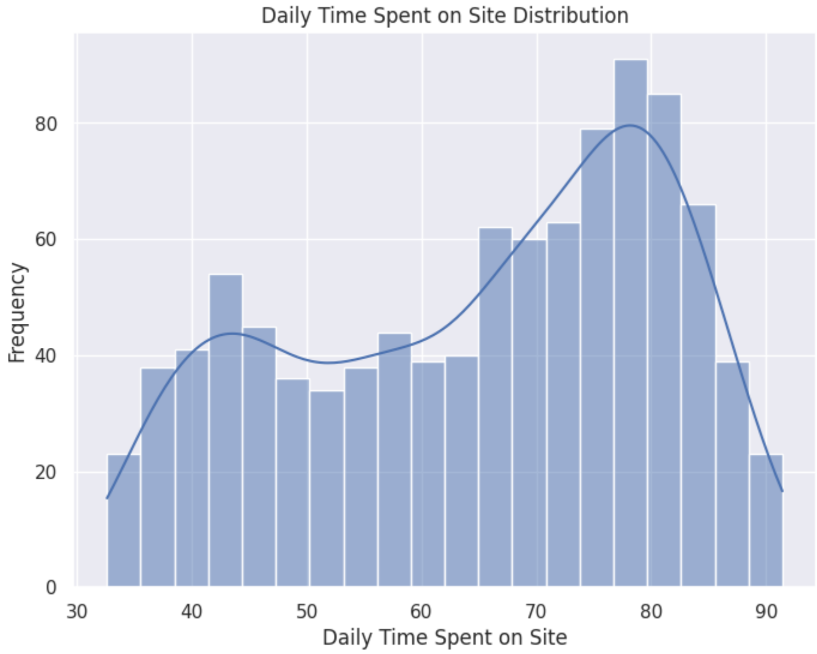
The distribution of participants who are male is less than those who are female.

Figure 3: Distribution of Area income against daily internet usage



The distribution of Area income against daily internet usage is nonlinear**.**

Figure 4: Distribution of Daily time spent on site



The distribution of daily time spent on site is not normally distributed. It skewed to the right.

d) Model Evaluation

Logistic Regression

	Before Tuning	**After Tuning**
Accuracy	0.975	0.975
Precision	0.989	0.989
Recall	0.960	0.960
F1 score	0.974	0.974

The performance metrics remain the same after tuning.

Random Forest

	Before Tuning	After Tuning
Accuracy	0.960	0.960
Precision	0.950	0.969
Recall	0.970	0.950
F1 score	0.960	0.959

Accuracy remains the same, but precision, recall, and F1 score have improved after tuning.

KNN

	Before Tuning	After Tuning
Accuracy	0.955	0.970
Precision	0.989	0.989
Recall	0.920	0.950
F1 score	0.953	0.969

Accuracy, recall and F1 score improved after tuning.

Decision Tree

	Before Tuning	After Tuning
Accuracy	0.960	0.960
Precision	0.960	0.960
Recall	0.960	0.960
F1 score	0.960	0.960

The performance metrics remain the same after tuning.

e) Model Comparison

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	98	99	96	97
Random Forest	96	95	97	96
KNN	96	99	92	95
Decision Tree	96	96	96	96

Based on these metrics:

Logistic Regression has the highest accuracy and precision, making it the top performer in terms of classifying the positive class ("Clicked on Ad"). It also has a strong F1 score and recall.

Random Forest is also strong, with high recall and an overall balanced performance. It performs slightly lower than logistic regression in precision.

K-Nearest Neighbors (KNN) has high precision and accuracy, but its recall is lower. This means it's good at correctly classifying positive cases but may miss some of them.

Decision Tree performs similarly to Random Forest in this case, with balanced performance across all metrics.

Conclusion

The research conducted an in-depth analysis of the performance of four machine learning models: Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and Decision Tree, within the context of predicting online advertisement clicks. The study's goal was to assess their efficacy based on demographic and behavioral features of users.

The findings revealed distinct strengths and areas of improvement for each model, offering valuable insights for practical application.

Logistic Regression exhibited outstanding performance, particularly in terms of precision and accuracy. It excelled in correctly identifying users likely to click on ads, making it a strong choice for applications where reducing false positives is essential.

Random Forest demonstrated a balanced performance, with a high recall rate and competitive F1 score. Its ensemble-based approach displayed robustness in handling diverse data patterns, making it a versatile choice suitable for scenarios requiring adaptability.

K-Nearest Neighbors (KNN) excelled in precision and accuracy, although its recall rate was comparatively lower, potentially leading to missed opportunities to identify users who would click on ads. KNN may be favored when precision is paramount, and the cost of false positives is significant.

Decision Tree, while not topping any specific metric, showcased balanced performance across all evaluation criteria. Its simplicity and interpretability rendered it a pragmatic choice, especially when model transparency and ease of comprehension are priorities.

It is important to recognize that model selection should be guided by the specific goals and constraints of the application. Considerations such as computational efficiency, interpretability, and the relative importance of different performance metrics should inform the model selection process.

Additionally, it is crucial to emphasize that the model's effectiveness was evaluated based on the dataset and features employed in this research. Real-world applications may necessitate further feature engineering and parameter tuning to optimize model performance. Regular model evaluation and adaptation are vital for sustaining effectiveness in practical use cases.

In conclusion, the outcomes of this research offer valuable insights into the performance of diverse machine learning models for online advertisement click prediction. By comprehending the strengths and weaknesses of these models, practitioners can make informed decisions when selecting the most suitable model for their specific application, ultimately enhancing the efficiency and effectiveness of their marketing campaigns.

Assumptions

During the course of constructing and evaluating our prediction models, a number of assumptions were made.

1) The dataset presented exhibits characteristics that are indicative of the larger population, and the survey data collected is deemed to be precise and dependable.

2) The features employed for the purpose of prediction are pertinent to the discernment of depression.

3) The partitioning of the data into training and testing sets was conducted in a random manner, adhering to established guidelines for evaluating machine learning models.

4) The fairness criteria employed for bias evaluation successfully capture potential biases inherent in the model predictions.

Limitations

The present study also possesses certain limitations:

1) It is possible that the dataset may not comprehensively encompass all pertinent aspects that contribute to the development of depression, and the inclusion of further variables has the potential to improve the efficacy of the model.

2) The potential limitation of the models' generalizability to larger populations may arise from the quantity of the dataset.

3) The issues encountered in our study were to the metrics of precision and recall, suggesting that our models may not now possess the requisite level of suitability for making crucial decisions in a clinical environment within the real world.

Challenges

Potential issues that may arise in the course of this project include:

i. Dealing with unbalanced data if there are substantially more instances of one class than the other (e.g., more "No" than "Yes" clicks).

ii. Effectively managing and processing timestamp data.

iii. Reducing bias in datasets and model predictions.

Considerations of ethics regarding data privacy and consent

Future uses/Additional Applications

There is need for additional refinement to enhance their applicability across a wider range of contexts.

1) Personalized interventions involve the customization of therapies to align with the unique risk profiles of individuals.

2) Population-level analysis involves the expansion of our existing models to examine the patterns and trends of depression within broader populations.

3) The integration of real-time data into various systems and processes has become increasingly prevalent in contemporary society. Leveraging contemporary data sources to enhance predictive capabilities.

Recommendations

The following recommendations are suggested for this analysis:

1) Examine supplementary functionalities: To improve the effectiveness of the model, it is recommended to incorporate supplementary socio-economic and psychological variables.

2) Conduct a comprehensive examination of sophisticated algorithms. Conducting experiments with advanced machine learning techniques in order to enhance the accuracy of predictions.

3) Perform external validation: It is imperative to assess the generalizability of our models by subjecting them to validation on diverse datasets.

4) Engage in collaborative efforts with mental health professionals: It is advisable to enlist the expertise of professionals in order to obtain specialized knowledge and enhance the comprehensibility of the model.

Implementation plan

To implement the following recommendations, the following plan is proposed:

1) Data augmentation: Procure and incorporate supplementary data sources that are pertinent, hence broadening the range of features.

2) The objective of this study is to conduct algorithm experimentation in the field of machine learning, specifically focusing on advanced algorithms such as gradient boosting and neural networks. The primary aim is to optimize the performance of these algorithms by fine-tuning their hyper parameters.

3) External validation involves the process of validating models using datasets that are external to the ones used for model development. This approach allows for the assessment of model performance in various demographic and cultural situations.

4) Collaboration entails engaging with mental health practitioners and researchers in order to acquire valuable insights pertaining to the field and enhance the therapeutic applicability of our models.

Ethical Considerations

Privacy and Confidentiality

The dataset will only be utilized for research objectives and will not be disclosed to external entities. To safeguard the identities of the respondents, the data will undergo anonymization. The process will entail eliminating any personally identifiable information (PII) from the dataset, including but not limited to names, addresses, and contact details.

Transparency

In order to ensure transparency, it is imperative to thoroughly document the data sources, preprocessing stages, and model choices employed in the analysis. This documentation serves to provide a comprehensive record of the information utilized, the methods applied to prepare the data, and the specific models selected for the analysis. By maintaining this level of transparency, researchers and stakeholders may better understand the decisions made throughout the process and assess the validity and reliability of the results.

The process of mitigating bias

Evaluating and mitigating potential biases in the data or models will be conducted. This will be accomplished by the utilization of a diverse range of methodologies, including:

i. Data cleaning encompasses the process of discovering and eliminating mistakes or inconsistencies present within the dataset.

ii. Feature selection is a crucial step in data analysis, as it entails carefully selecting a subset of features pertinent to the specific task at hand. The objective of feature selection is to identify and retain only those relevant elements while ensuring that the selected features do not introduce bias into the analysis.

iii. The model evaluation process encompasses the assessment of biases within the models through the utilization of methodologies such as fairness metrics and discrepancy analysis.

Informed Consent

Before collecting their data, the researchers will ensure that the survey participants have consented. The goal of the study, the data collection process, and the data utilization will be communicated to the participants. The participants will also have the option to revoke their consent at any time.

References

Goyal, A., Bhong, S., Kumbhare, P., & Bharadkar, R. (2021). The new era of digital marketing: a literature review. *PalArch's Journal of Archaeology of Egypt/Egyptology*, *18*(10), 728-741.

Cokluk, O. (2010). Logistic Regression: Concept and Application. *Educational Sciences: Theory and Practice*, *10*(3), 1397-1407.

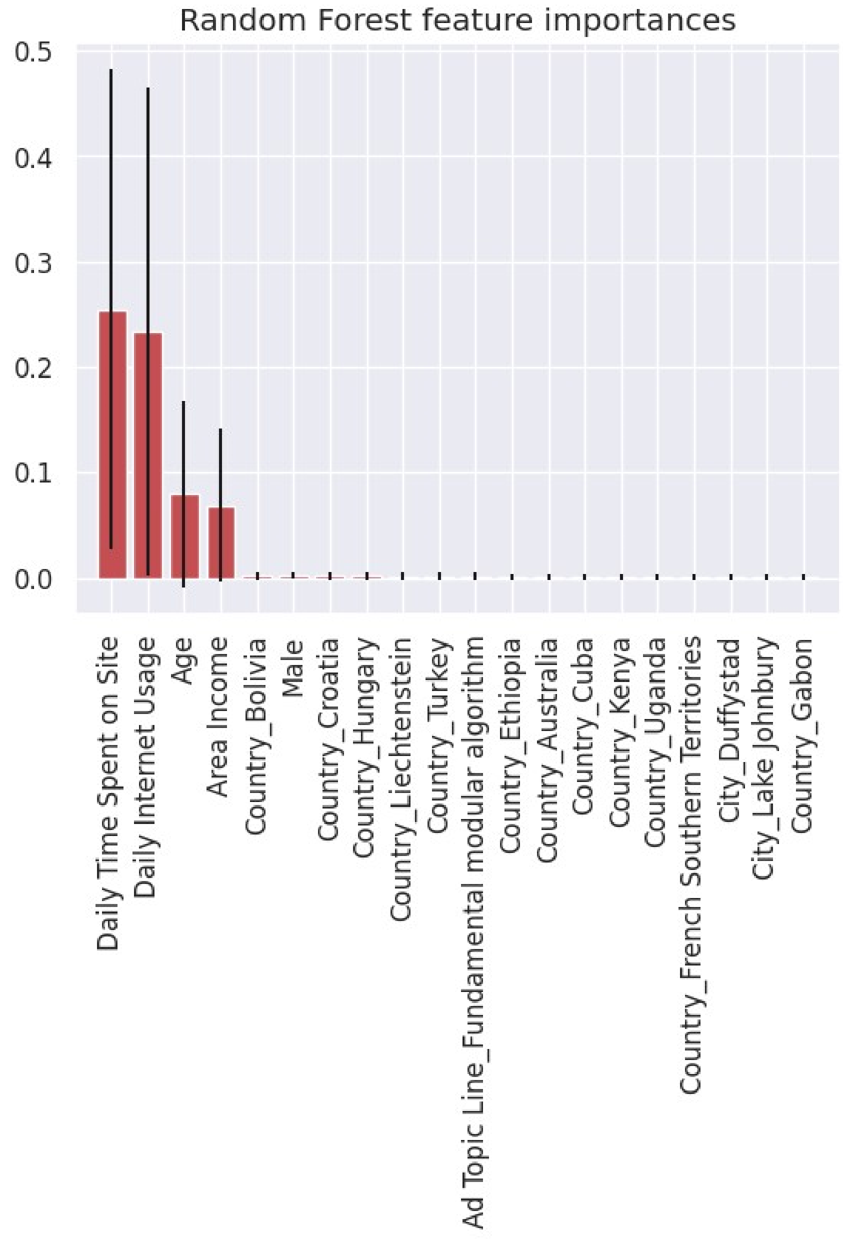
Cornelius, K., & Shanthini, B. (2023). Air Quality Data Analysis and Prediction Using Modified Differential Evolution-Random Forest Algorithm. *Journal of Survey in Fisheries Sciences*, pp. 3067–3078.

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021, November). Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. In *Informatics* (Vol. 8, No. 4, p. 79). MDPI.

Vrigazova, B. (2021). The proportion for splitting data into training and test sets for the bootstrap in classification problems. *Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy*, *12*(1), 228-242.

Appendix

Appendix A


Random Forest feature importances
Daily Time Spent on Site
Daily Internet Usage
Age
Area Income
Country_Bolivia
Male
Country_Croatia
Country_Hungary
Country_Liechtenstein
Country_Turkey
Ad Topic Line_Fundamental modular algorithm
Country_Ethiopia
Country_Australia
Country_Cuba
Country_Kenya
Country_Uganda
Country_French Southern Territories
City_Duffystad
City_Lake Johnbury
Country_Gabon