1. Abstract

In the competitive landscape of digital marketing, predicting ad clicks is essential for optimizing campaign effectiveness and maximizing ROI. This project, "Predicting Ad Clicks on Advertisement Website," applies machine learning techniques on user demographic and behavioral data to uncover key insights. Models such as Logistic Regression, Support Vector Machine (SVM), and Gradient Boosting showcase their unique strengths in classification. Logistic Regression provides robust accuracy and interpretability, SVM offers precise boundary definitions, and Gradient Boosting captures complex patterns. Comparative analysis highlights the value of feature scaling, with Logistic Regression emerging as the top performer.

The study's findings suggest actionable insights for targeted ad delivery, helping businesses better allocate resources, reduce ad spending, and increase engagement. While limited by available features, this study recommends future work on deeper behavioral metrics and advanced models to improve prediction accuracy further. This project provides a framework for enhancing targeted marketing, contributing to efficient and effective advertising strategies.

1. Why *Click Sense*?

2.1. Introduction:

In the fast-paced field of digital advertising, the ability to predict user engagement through ad clicks is essential for optimizing marketing effectiveness and increasing return on investment (ROI). The competition for users' attention online has made it more important than ever for companies to target their ads to the right audience efficiently. By accurately predicting ad clicks, businesses can streamline their ad spending, reach potential customers more effectively, and enhance user experience with relevant advertisements. This project, "Predicting Ad Clicks on Advertisement Website," was selected to address these industry needs using machine learning to make data-driven predictions.

2.2. Reason for choosing the project:

The idea for our project, *Click Sense*, originated from a personal experience that highlighted a common challenge. While searching for solutions to a coding problem, I encountered numerous online ads. One ad was particularly compelling, almost leading me to make a purchase. This experience sparked a realization: advertisements play a significant role in influencing user behavior. Recognizing the relevance of this phenomenon to our field of study, I proposed the idea to my teammates. After thorough discussion and understanding the broader importance of analyzing and optimizing ad interactions, we decided to pursue it as our project. That moment marked the beginning of *Click Sense*.

2.3. Problem Statement:

The primary challenge of this project aims to improve the effectiveness of online advertising campaigns by predicting user behavior—specifically, whether a user will click on an advertisement or not. This is crucial for optimizing ad-targeting strategies, allocating resources efficiently, and increasing overall return on investment (ROI) for the advertising budget.

2.4. Data Description:

Attributes	Sample Data	Description
Age	35	Customer age in years
Area Income	68441.85	Avg. Income of geographical area of consumer
Daily Internet Usage	193.77	Avg. minutes a day consumer is on the internet
Male	1	Whether or not consumer was male
Daily time spent on site	68.95	Consumer time on site in minutes
Time stamp	2016-03-27 00:53:11	Time at which consumer clicked on Ad or closed window
Ad Topic Line	Robust logistical utilization	Headline of the advertisement
City	Wright burgh	City of consumer
Country	Tunisia	Country of consumer
Clicked on Ad or Not	1	0 or 1 indicated clicking on Ad

2. Methodology

3.1. Sample:

We looked through a few data sets relevant to advertisement ads and found one from Kaggle. There are 1000 records. The dataset contains 9 variables for each individual and one additional column that indicates whether the individual will click on the ad or not.

3.2. Explore:

We searched for the missing values and discovered zero missing values in the dataset.

3.3. Modify:

We searched the dataset for unnecessary columns and Ad Topic Line, Country, Timestamp, and city are unique identifiers for that have no bearing on forecast. So, those are deleted.

3.4. Standardization:

We standardized the data because all the variables are on different scales. So, the features with larger magnitudes may disproportionately influence the model, leading to biased results and reducing the overall performance of the model.

3.5. Model:

For predicting ad clicks, we selected Logistic Regression, Support Vector Machine (SVM), and Gradient Boosting models. Each model provided unique strengths, allowing us to compare their performance in identifying ad click patterns.

3.5.1. Logistic Regression:

Logistic Regression was chosen as our primary model due to its interpretability and relatively low false positive rate. Reducing false positives is crucial in advertising, as overestimating clicks can lead to inefficient ad spending. To enhance accuracy, we applied standardization for standardizing feature scales to improve class distinction. Logistic Regression also provided insights into feature significance, making it valuable for understanding user behavior.

3.5.2. Support Vector Machine (SVM)

SVM was implemented for its capacity to classify data by creating a decision boundary that maximizes the margin between clicked and non-clicked users. SVM achieved high accuracy but had a higher false positive rate than Logistic Regression, guiding our preference for Logistic Regression in scenarios requiring high precision. Nevertheless, SVM's ability to capture complex relationships made it a valuable comparative model.

3.5.3. Gradient Boosting

Gradient Boosting was employed to capture non-linear relationships between features, leveraging an ensemble of decision trees to improve prediction accuracy. This model excelled in capturing complex patterns, though its computational intensity limited its suitability for real-time applications compared to Logistic Regression. Gradient Boosting's performance provided a robust benchmark.

3. Findings

4.1. Results

Model	Precision	Accuracy	Confusion Matrix
Logistic Regression	98.5 %	97.25 %	[188 3] [8 201]
Support Vector Machine	98.02 %	96.5 %	[187 4] [10 199]
Gradient Boosting	96.56 %	95.25 %	[184 7] [12 197]

- Based on the results it can be inferred that Logistic Regression is performing well in terms of Precision and accuracy.
- Moreover, the false positive rate for logistic regression is low compared to other models. So, we decided to move ahead with Logistic Regression.

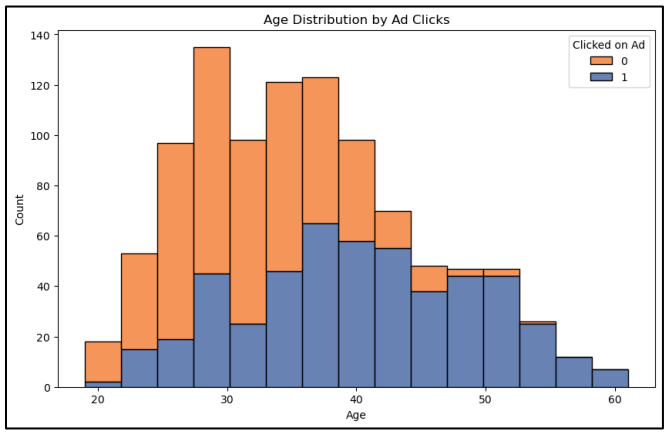
4.2. Feature Selection:

Feature	Odds Ratio
Daily Time Spent on Site	0.939854
Age	1.301788
Area Income	0.999984
Daily Internet Usage	0.976033
Male	1.001782

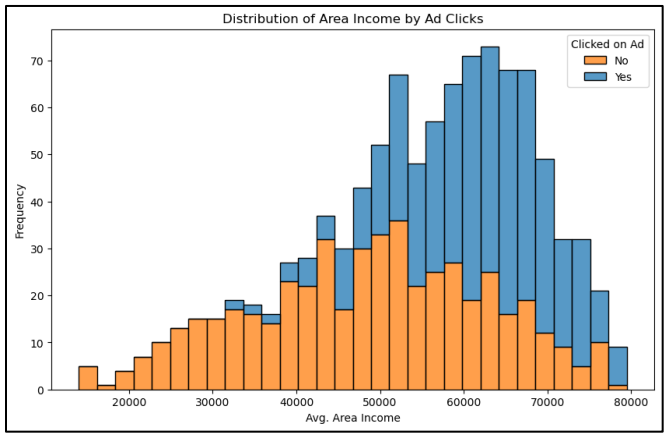
- Daily Time Spent on Site – Odds Ratio: 0.94
  A slight negative impact. For every additional minute spent on the site, the likelihood of clicking decreases slightly.
- Age – Odds Ratio: 1.30
  A significant positive impact. Older users are more likely to click on ads.
- Area Income – Odds Ratio: 0.99
  A minimal negative impact, indicating that income doesn't play a strong role in ad-clicking behavior.
- Daily Internet Usage – Odds Ratio: 0.98
  Similar to time spent, there is a slight decrease in the likelihood of clicking as internet usage increases.
- Male – Odds Ratio: 1.00
  Neutral impact, showing gender doesn't strongly influence ad clicks in this dataset.

4.3. Exploratory Data Analysis:

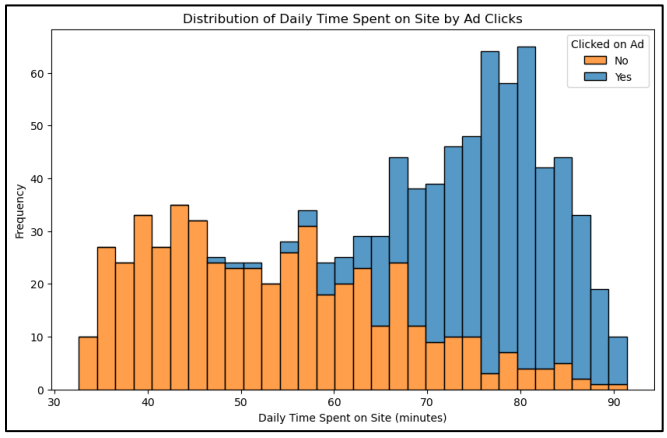
Age vs Clicked on Ad or not:



Area Income vs Clicked on Ad or not:



Daily time spent on site vs Clicked on Ad or not:



- From the EDA it can be observed that as Age of person increases the probability of that person clicking ad increases.
- Also, as Income of an individual increased his purchasing power increases then the probability of clicking ad increases.
- If an individual spends more time on site, the chances of clicking ad increases.

4. Challenges

- Data Preparation and Scaling: One of the significant challenges was preparing and scaling the data, particularly for models like SVM and Logistic Regression. Without proper scaling, the initial results showed poor performance, with Logistic Regression achieving low accuracy prior to standardization.

- Balancing Model Performance: Each model had its strengths and limitations. While SVM performed well in terms of recall, it had a slightly higher false positive rate, which could lead to inefficient ad spending by misclassifying non-clickers as clickers. Balancing these trade-offs was essential in selecting the best model, as we needed a model that provided both precision and interpretability while minimizing unnecessary ad spending.

- Data Limitations

5.1. Solutions and Adjustments

To address these challenges, we made the following adjustments:

- Feature Selection and Scaling: We used correlation analysis to carefully select the features with the most predictive power and standardized the data to improve model performance. After standardization, Logistic Regression's precision improved significantly, highlighting the importance of scaling for this model. This scaling adjustment was also crucial for SVM to enhance its accuracy and reduce false positives.

- Model Comparison and Selection: By comparing the models across various metrics, we determined that Logistic Regression provided the best balance of precision, interpretability, and a low false positive rate. Its improved performance post-standardization made it an ideal choice for a practical ad-click prediction model that aligns with our project goals.

5.2. Lessons Learned

Reflecting on the project, we gained valuable insights:

Importance of Standardization: The improvement in Logistic Regression's precision post-standardization underscored the critical role of scaling in machine learning models, particularly for models sensitive to feature magnitudes.

Potential for Cross-Validation: In hindsight, incorporating cross-validation could have improved the robustness of our model by ensuring consistency across multiple training and testing splits.

Value of Simple, Interpretable Models: We learned that simpler models like Logistic Regression can be highly effective for binary classification tasks when the data is well-prepared, without the need for complex algorithms. This reinforced the importance of focusing on data quality and preprocessing as much as on model selection.

5. Conclusions

We divided the age into 4 categories like 21-30, 31-40, 41-50, 51-60.

- Our analysis highlighted **age** as the deciding feature influencing ad-click likelihood. Older users demonstrated a higher propensity to click on ads, possibly due to different browsing behaviors or preferences compared to younger users who may have higher exposure and resistance to online advertisements. This insight suggests that targeting older demographic segments could yield more engagement, making age-targeted marketing an effective strategy for businesses.
- Whereas from EDA it can be inferred that Daily time spent on site, Area Income and Age shows a positive relation with clicking the Ad.

Based on the above conclusions we want to focus on 4 factors on.

6.1. Future Scope:

Advanced Personalization: We can take a step further by implementing dynamic content that adapts based on user interests. By doing so, we can provide users with more relevant ads, which would likely boost both engagement and conversion rates.

A/B Testing and Model Refinement: A/B testing is a powerful tool that can help refine the effectiveness of ad creatives, placement, and timing. By continuously testing and fine-tuning our approach, we can ensure that ads are optimized for maximum impact and ROI.

Expanding Customer Segments: Our current model has already provided insights into age-based ad strategies. But there's much more we can explore. Diving deeper into behavioral, geographic, and cultural data will help us uncover untapped customer segments that we can target more effectively.

Predictive Insights for ROI: By leveraging predictive analytics, we can forecast ad performance and dynamically allocate advertising budgets. This will allow businesses to make data-driven decisions that maximize returns on their ad spending.

6. References

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