Predicting Click-Through Rate (CTR) Based on User Demographics and Online Behavior

yuetiany@uchicago.edu

2025-02-24

> **!** Important
>
> **Github Repo** https://github.com/yuetianyuan/30100_final_project.git

Table of contents

1. Indroduction

Click-Through Rate (CTR) had been a critical measure in digital marketing and online advertising, measuring whether users who engage with an advertisement click on it. Understanding the factors that influence CTR and modeling CTR prediction are essential for optimizing ad targeting and improving engagement strategies. This study aims to answer the research question: "How do users demographics and online behavior influence Click-Through Rate (CTR)?"Specifically how do age, internet usage, and income levels influence whether users engage with online ads? By analyzing user characteristics such as age, daily internet usage, and income, this study seeks to uncover key behavioral patterns that drive engagement with online advertisements.

2. Background and Literature Review

Predicting Click-Through Rate (CTR) has been a crucial area of research in computational advertising, aiming to enhance ad targeting, user engagement, and conversion rates. Traditional statistical models, machine learning techniques, and deep learning approaches have all been explored to improve CTR prediction accuracy.

Machine learning models have significantly advanced CTR prediction in recent years. One study, "Research on Advertising Click-Through Rate Prediction Model Based on Taobao Big Data" by Chen (2023) compares various models, including Logistic Regression, Random Forest, Gradient Boosting Decision Trees (GBDT), and LightGBM, for CTR prediction (Chen 2023). The study emphasizes that ensemble learning methods like GBDT and LightGBM outperform traditional statistical models due to their ability to capture complex interactions between user demographics and behavioral features. Furthermore, the research explores a Stacking-based fusion model and a BP neural network, which introduces deep learning techniques for improving prediction accuracy. These findings suggest that CTR models must balance interpretability and predictive power, selecting algorithms that align with both computational efficiency and marketing needs. Another study also try to explain CTR from users' previous behaviors. Qin et al. (2020), in their study "User Behavior Retrieval for Click-Through Rate Prediction", highlight the importance of historical user behavior retrieval in CTR modeling. Their research argues that user behavior follows long-term dependencies and periodic patterns, which can significantly influence CTR predictions (Qin et al. 2020). Instead of feeding a complete history of user behavior into predictive models, the study introduces the User Behavior Retrieval (UBR4CTR) framework, which selects only the most relevant past behaviors. This technique improves model efficiency while retaining crucial behavioral signals. Qin et al.'s findings suggest that traditional machine learning approaches may struggle to fully leverage sequential user data, necessitating specialized retrieval-based frameworks for enhanced CTR predictions.

Beyond traditional and ensemble learning models, deep learning frameworks have emerged as powerful tools for CTR prediction. Guorui Zhou et al. (2019) introduce the Deep Interest

Evolution Network (DIEN), a model designed to capture dynamic user interest evolution over time. Their study argues that user interests are not static but evolve due to external factors (e.g., market trends, advertisements) and internal cognitive changes (e.g., shifting preferences, new needs) (Zhou et al. 2019). Unlike static models, DIEN incorporates Gated Recurrent Units (GRUs) to extract temporal patterns from user behavior and an Attention Update Gate mechanism to model interest evolution dynamically. The results demonstrate that CTR models benefit from architectures that explicitly capture sequential dependencies and adapt to changing user behaviors. These insights suggest that future CTR prediction frameworks may require deep learning architectures that integrate memory mechanisms and adaptive interest modeling to enhance predictive accuracy.

Existing research highlights the complexity of CTR prediction, emphasizing the role of user behavior retrieval, dynamic interest evolution, and ensemble learning methods in improving model accuracy. While machine learning approaches like Random Forest and GBDT have demonstrated strong performance in handling structured data, deep learning methods like DIEN provide a more sophisticated framework for capturing user interest dynamics over time. This study builds on these insights by integrating demographic, behavioral, and temporal features into a Random Forest model while evaluating Lasso Regression and OLS Regression for feature selection and interpretability. By comparing traditional, ensemble, and regression-based approaches, this research tries to provide a comprehensive understanding of how user demographics and online behavior impact CTR, contributing to more marketing implications.

3. Data Description

The dataset used in this study was obtained on Kaggle and contains 10,000 user interactions with advertisements. The data includes a variety of demographic, behavioral, and temporal attributes, allowing for a thorough analysis of factors influencing CTR. The key features include Age, Daily Internet Usage, Area Income, and Timestamp, along with the binary outcome variable, Clicked on Ad, indicating whether a user engaged with an ad. To capture temporal patterns, additional features were extracted, including Time of Day (Morning, Afternoon, Evening, Night) and Weekend Indicator (0 for weekday, 1 for weekend). These temporal features allow for an assessment of whether time-based factors influence CTR.

4. Data Limitations and Challenges

A notable challenge in CTR prediction is data imbalance. Real-world CTR datasets often have extremely low click rates (<1%), requiring resampling techniques or cost-sensitive learning to improve model performance. However, this dataset appears balanced, making it easier to train models but less representative of real-world CTR trends. Another potential bias stems from income disparities—users from higher-income regions may engage with ads differently from lower-income users. Additionally, time-dependent effects such as variations in engagement between weekdays and weekends may introduce patterns that affect CTR predictions.

Addressing these limitations requires careful consideration of feature engineering and model selection.

Daily.Time.Spent.on.Site	Age	Area.Income	Daily.Internet.Usage	Ad.Topic.Line	City	Gender	Country	Timestamp	Clicked.on.Ad
62.26	32	69481.85	172.83	Decentralized real-time circuit	Lisafort	Male	Svalbard & Jan Mayen Islands	2016-06-09 21:43:05	0
41.73	31	61840.26	207.17	Optional full-range projection	West Angelabury	Male	Singapore	2016-01-16 17:56:05	0
44.40	30	57877.15	172.83	Total 5thgeneration standardization	Reyesfurt	Female	Guadeloupe	2016-06-29 10:50:45	0
59.88	28	56180.93	207.17	Balanced empowering success	New Michael	Female	Zambia	2016-06-21 14:32:32	0
49.21	30	54324.73	201.58	Total 5thgeneration standardization	West Richard	Female	Qatar	2016-07-21 10:54:35	1
51.30	26	51463.17	131.68	Focused multi-state workforce	Port Maria	Female	Cameroon	2016-05-15 13:18:34	0

```
# Convert categorical variables
df$Clicked.on.Ad <- as.factor(df$Clicked.on.Ad)

# Summary statistics
summary(df)
```

```
 Daily.Time.Spent.on.Site      Age          Area.Income    Daily.Internet.Usage
 Min.   :32.60            Min.   :19.00   Min.   :13996   Min.   :105.2
 1st Qu.:48.86            1st Qu.:29.00   1st Qu.:44052   1st Qu.:140.2
 Median :59.59            Median :35.00   Median :56181   Median :178.9
 Mean   :61.66            Mean   :35.94   Mean   :53840   Mean   :177.8
 3rd Qu.:76.58            3rd Qu.:42.00   3rd Qu.:61840   3rd Qu.:212.7
 Max.   :90.97            Max.   :60.00   Max.   :79332   Max.   :270.0
 Ad.Topic.Line          City              Gender            Country
 Length:10000       Length:10000       Length:10000       Length:10000
```

```
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
  Timestamp          Clicked.on.Ad
 Length:10000       0:5083
 Class :character   1:4917
 Mode  :character
```

```
# Check for missing values
colSums(is.na(df))
```

```
Daily.Time.Spent.on.Site                      Age                Area.Income
                       0                        0                          0
    Daily.Internet.Usage           Ad.Topic.Line                       City
                       0                        0                          0
                  Gender                  Country                  Timestamp
                       0                        0                          0
           Clicked.on.Ad
                       0
```

5. Feature Engineering

To enhance predictive power, several feature transformations and interactions were applied. First, Gender and Cilck on ad variables were converted to categorical varibles for later analysis. The interaction term Age_InternetUsage was created to assess whether the effect of age on CTR is influenced by online activity levels. Temporal features were extracted from timestamps to categorize interactions into different time periods (Time of Day and Weekend Indicator). These variables were prepared in better format for further analysis, and later were included in model training to improve the representation of user behavior and demographic influences on CTR.

```
# Convert categorical variables if needed
df$Gender <- as.factor(df$Gender)
df$Clicked.on.Ad <- as.factor(df$Clicked.on.Ad)

library(lubridate)
```

```r
# Convert timestamp to datetime
df$Timestamp <- as.POSIXct(df$Timestamp, format="%Y-%m-%d %H:%M:%S")

# Extract Hour of the Day
df$Hour <- hour(df$Timestamp)

# Create Time of Day Categories
df$Time_of_Day <- case_when(
  df$Hour >= 6 & df$Hour < 12 ~ "Morning",
  df$Hour >= 12 & df$Hour < 18 ~ "Afternoon",
  df$Hour >= 18 & df$Hour < 24 ~ "Evening",
  TRUE ~ "Night"
)

# Convert to factor
df$Time_of_Day <- as.factor(df$Time_of_Day)

# Extract Weekend Indicator
df$Weekend <- ifelse(weekdays(df$Timestamp) %in% c("Saturday", "Sunday"), 1, 0)
```

6. Exploratory Data Analysis (EDA)

The exploratory data analysis process was conducted to gain a depper insights of our data sights, and access the distribution of variables in our data set. EDA revealed several key behavioral trends in CTR engagement. The first step in EDA involved checking for missing values and ensuring data consistency. No missing values were detected in the dataset, indicating that all features were fully available for analysis. Summary statistics were created to examine the central tendencies and dispersion of numerical variables such as Age, Daily Internet Usage, and Area Income. This analysis revealed that most users were within the age range of 20 to 60, daily internet usage varied widely from less than an hour to over 300 minutes per day, and area income was skewed toward higher values, necessitating a log transformation to normalize the distribution. To better understand how user demographics and users' behaviors affect CTR, several visualizations were created. First, a heatmap of feature correlations was created to understand the correlation between variables, showing that Age and CTR had a moderate positive correlation, while Daily Internet Usage and CTR had a slight negative correlation. A scatter plot of Daily Internet Usage against CTR revealed that users who spent more time online tended to have lower engagement with ads. This pattern suggested the presence of ad fatigue, where users who are frequently exposed to online ads may become desensitized, leading to lower CTR. Another important visualization was the bar chart of Time of Day against CTR. This plot indicated that CTR was slightly higher in the evening compared to morning hours, suggesting that users might be more engaged with advertisements after work hours. A

density plot of Age grouped by Clicked on Ad status was plotted to observe CTR variation across different age groups. This plot showed that younger users (<30) had lower CTR, while middle-aged users (30–50) exhibited the highest engagement.

```
# Count unique values in categorical variables
df %>%
  summarise(
    Unique_Genders = n_distinct(Gender),
    Unique_Countries = n_distinct(Country),
    Unique_Cities = n_distinct(City)
  )
```

```
  Unique_Genders Unique_Countries Unique_Cities
1              2              207           521
```

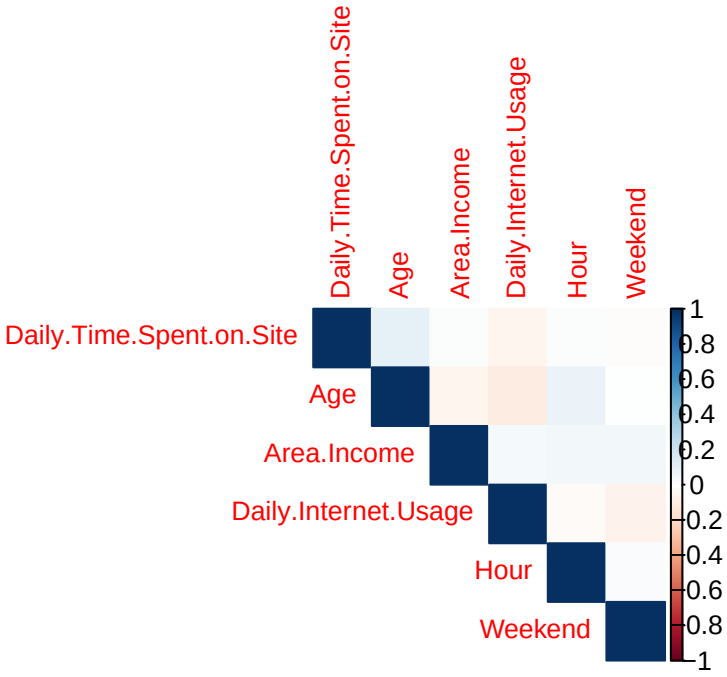
```
glimpse(df)
```

```
Rows: 10,000
Columns: 13
$ Daily.Time.Spent.on.Site <dbl> 62.26, 41.73, 44.40, 59.88, 49.21, 51.30, 66.~
$ Age                      <dbl> 32, 31, 30, 28, 30, 26, 43, 26, 33, 51, 29, 3~
$ Area.Income              <dbl> 69481.85, 61840.26, 57877.15, 56180.93, 54324~
$ Daily.Internet.Usage     <dbl> 172.83, 207.17, 172.83, 207.17, 201.58, 131.6~
$ Ad.Topic.Line            <chr> "Decentralized real-time circuit", "Optional ~
$ City                     <chr> "Lisafort", "West Angelabury", "Reyesfurt", "~
$ Gender                   <fct> Male, Male, Female, Female, Female, Female, M~
$ Country                  <chr> "Svalbard & Jan Mayen Islands", "Singapore", ~
$ Timestamp                <dttm> 2016-06-09 21:43:05, 2016-01-16 17:56:05, 20~
$ Clicked.on.Ad            <fct> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, ~
$ Hour                     <int> 21, 17, 10, 14, 10, 13, 21, 6, 6, 5, 0, 18, 6~
$ Time_of_Day              <fct> Evening, Afternoon, Morning, Afternoon, Morni~
$ Weekend                  <dbl> 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, ~
```

```
# Summary Statistics
summary(df)
```

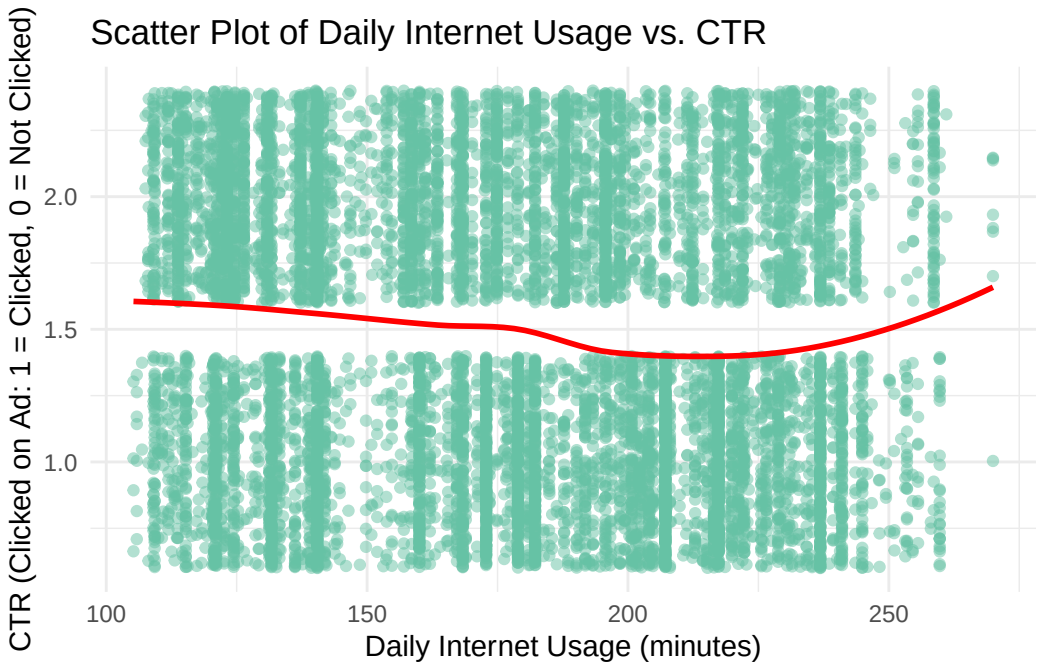
```
 Daily.Time.Spent.on.Site      Age         Area.Income    Daily.Internet.Usage
 Min.   :32.60            Min.   :19.00   Min.   :13996   Min.   :105.2
 1st Qu.:48.86            1st Qu.:29.00   1st Qu.:44052   1st Qu.:140.2
 Median :59.59            Median :35.00   Median :56181   Median :178.9
```

```
Mean   :61.66            Mean   :35.94   Mean   :53840   Mean   :177.8
3rd Qu.:76.58            3rd Qu.:42.00   3rd Qu.:61840   3rd Qu.:212.7
Max.   :90.97            Max.   :60.00   Max.   :79332   Max.   :270.0
Ad.Topic.Line          City              Gender       Country
Length:10000       Length:10000       Female:5376   Length:10000
Class :character   Class :character   Male  :4624   Class :character
Mode  :character   Mode  :character                 Mode  :character



  Timestamp                       Clicked.on.Ad      Hour
Min.   :2016-01-01 02:52:10.00   0:5083        Min.   : 0.00
1st Qu.:2016-02-26 01:18:44.00   1:4917        1st Qu.: 6.00
Median :2016-04-04 22:00:15.00                 Median :13.00
Mean   :2016-04-10 17:16:41.53                 Mean   :12.34
3rd Qu.:2016-06-02 21:02:22.00                 3rd Qu.:19.00
Max.   :2016-07-23 11:46:28.00                 Max.   :23.00
   Time_of_Day      Weekend
Afternoon:2482   Min.   :0.000
Evening  :3020   1st Qu.:0.000
Morning  :2250   Median :0.000
Night    :2248   Mean   :0.289
                 3rd Qu.:1.000
                 Max.   :1.000
```



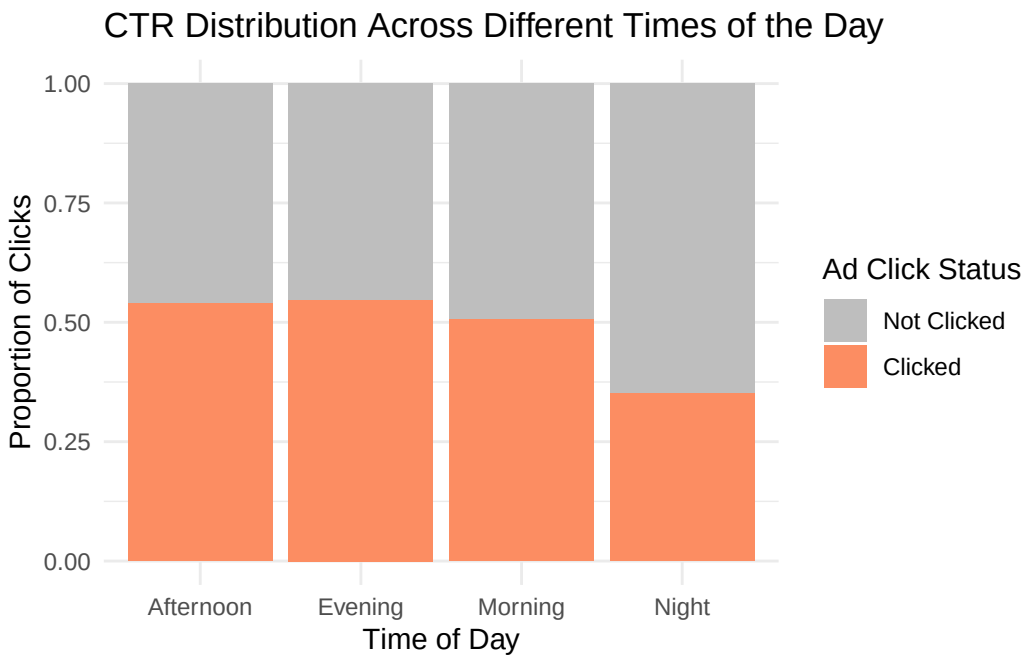

```
#Scatter Plot: Daily Internet Usage vs. CTR
ggplot(df, aes(x = Daily.Internet.Usage, y = as.numeric(Clicked.on.Ad))) +
  geom_jitter(alpha = 0.5, color = "#66C2A5") +  # Adds scatter points with slight jitter for
  geom_smooth(method = "loess", color = "red", se = FALSE) +  # Add smooth trend line
  labs(title = "Scatter Plot of Daily Internet Usage vs. CTR",
       x = "Daily Internet Usage (minutes)",
       y = "CTR (Clicked on Ad: 1 = Clicked, 0 = Not Clicked)") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

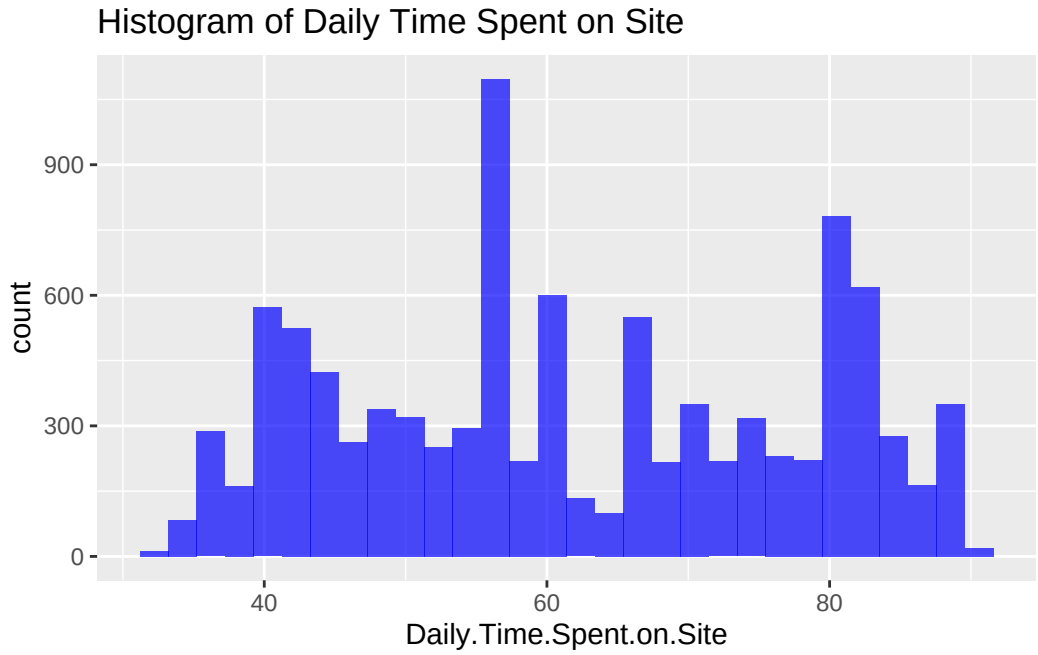


Users with higher internet usage show a decreasing trend in CTR.

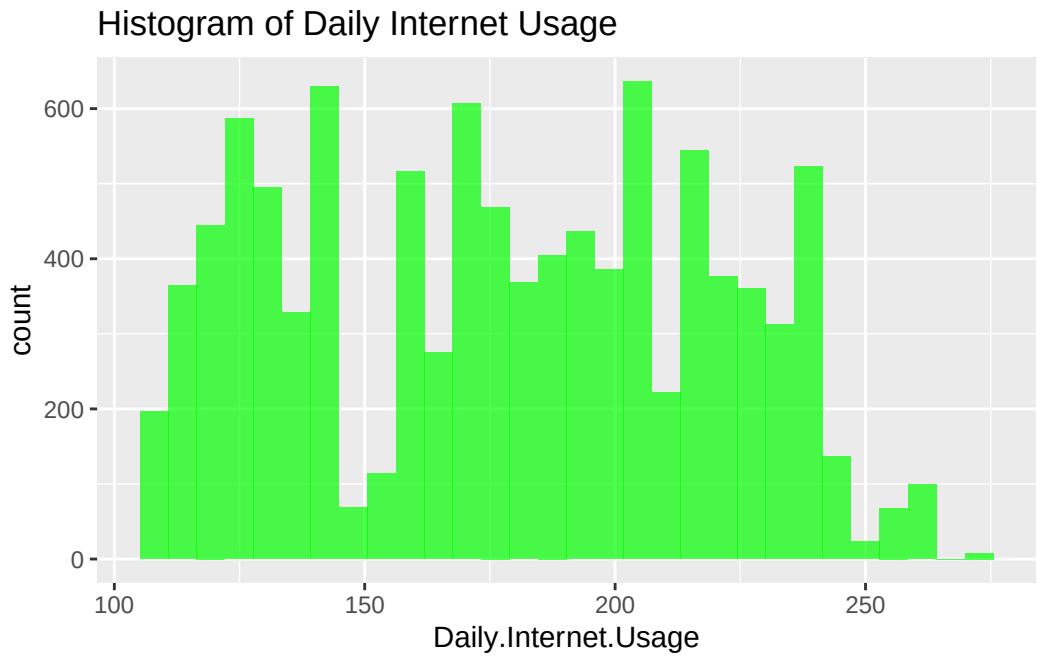
```
ggplot(df, aes(x = Time_of_Day, fill = as.factor(Clicked.on.Ad))) +
  geom_bar(position = "fill") +  # Proportional CTR distribution across time of day
  scale_fill_manual(values = c("0" = "gray", "1" = "#FC8D62"), labels = c("Not Clicked", "Cl
  labs(title = "CTR Distribution Across Different Times of the Day",
       x = "Time of Day",
       y = "Proportion of Clicks",
       fill = "Ad Click Status") +
  theme_minimal()
```



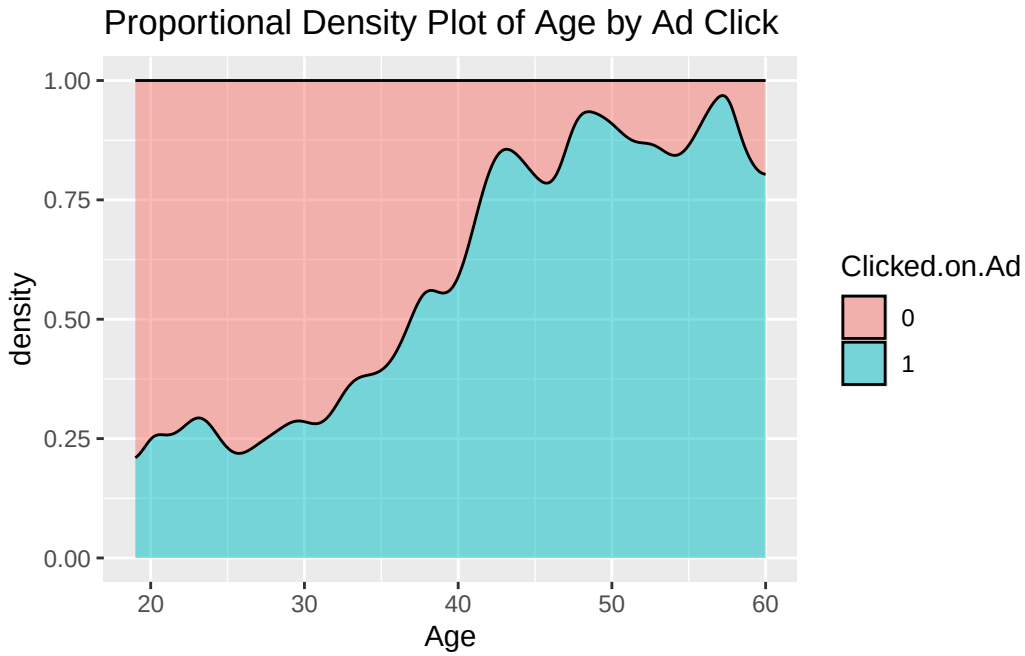
Evening shows a slightly higher CTR, suggesting users may be more engaged with ads after work hours.



shows how much time users spend on the website daily.



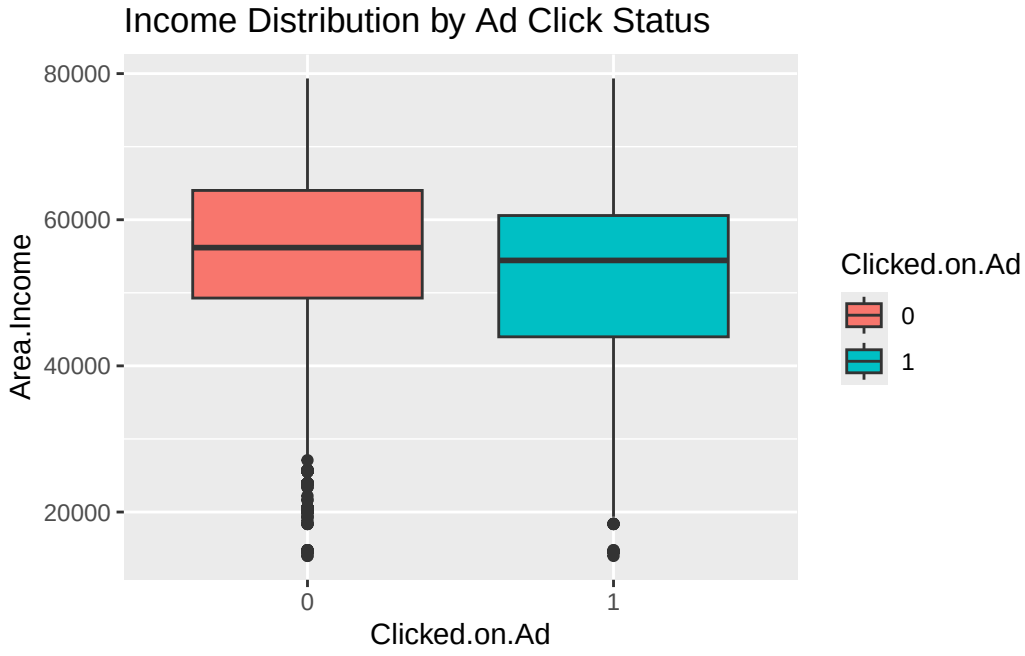
Helps understand how long users stay online daily.



stacked proportional density plot of Age by Ad Click. Younger individuals (age < 30): The proportion of not clicking on ads (red) is higher. Middle-aged individuals (age ~30-50): The

likelihood of clicking on an ad (blue) increases significantly. Older individuals (age >50): More people tend to click on ads, but there's still a portion who do not.

```
# A tibble: 2 x 5
  Clicked.on.Ad Mean_Internet_Usage Median_Internet_Usage SD_Internet_Usage
  <fct>                       <dbl>                 <dbl>             <dbl>
1 0                            183.                  182.              39.6
2 1                            172.                  168.              41.3
# i 1 more variable: Count <int>
```



7. Machine Learning Models and Comparison

To predict CTR, three machine learning models were tested: Random Forest, Lasso Regression, and Ordinary Least Squares (OLS) Regression. Random Forest was chosen as the primary model due to its ability to capture nonlinear relationships and rank feature importance. A hyperparameter tuning process was conducted, identifying the optimal mtry = 3, which determines the number of features considered at each tree split. Lasso Regression was applied to select the most important predictors by penalizing less relevant variables. The results indicated that the interaction term (Age_InternetUsage) did not significantly improve prediction, leading to its exclusion in the final Lasso model. OLS Regression was used as an interpretable baseline model, confirming key findings such as the positive effect of age on CTR and the negative effects of internet usage and income on engagement.

Random Forest

Random Forest was chosen to model the CTR prediction due to its ability to capture nonlinear relationships and feature importance rankings. Hyperparameter tuning identified mtry = 3 as the optimal feature subset per split.The results show that the model achieved an accuracy of 80.3%, with Age being the most important feature, followed by Daily Internet Usage and Area Income.

```
# Train Random Forest Model with Hyperparameter Tuning
# Create interaction term
df$Age_InternetUsage <- df$Age * df$Daily.Internet.Usage
```

```
# Split data into training (80%) and testing (20%)
set.seed(42)
trainIndex <- createDataPartition(df$Clicked.on.Ad, p = 0.8, list = FALSE)
train_data <- df[trainIndex, ]
test_data <- df[-trainIndex, ]
```

```
# Define hyperparameter grid
rf_grid <- expand.grid(
  mtry = c(2, 3, 4)
)
```

```
# Train Random Forest Model using 5-fold Cross-Validation
set.seed(42)
rf_model <- train(
  Clicked.on.Ad ~ Daily.Internet.Usage + Age + Age_InternetUsage + Area.Income + Time_of_Day
  data = train_data,
  method = "rf",
  trControl = trainControl(method = "cv", number = 5),
  tuneGrid = rf_grid,
  importance = TRUE
)
```

```
# Print best hyperparameters
print(rf_model$bestTune)
```

```
  mtry
2    3
```

The tuning process found mtry = 3 to be optimal for maximizing predictive performance.

```r
# Predict on test data
rf_predictions <- predict(rf_model, test_data)
rf_probabilities <- predict(rf_model, test_data, type = "prob")[, 2]

# Model Performance
confusionMatrix(rf_predictions, test_data$Clicked.on.Ad)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 832 209
         1 184 774

               Accuracy : 0.8034
                 95% CI : (0.7853, 0.8206)
    No Information Rate : 0.5083
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6065

 Mcnemar's Test P-Value : 0.226

            Sensitivity : 0.8189
            Specificity : 0.7874
         Pos Pred Value : 0.7992
         Neg Pred Value : 0.8079
             Prevalence : 0.5083
         Detection Rate : 0.4162
   Detection Prevalence : 0.5208
      Balanced Accuracy : 0.8031

       'Positive' Class : 0
```

```r
# AUC-ROC Curve
library(pROC)
```

```
Type 'citation("pROC")' for a citation.

Attaching package: 'pROC'
```

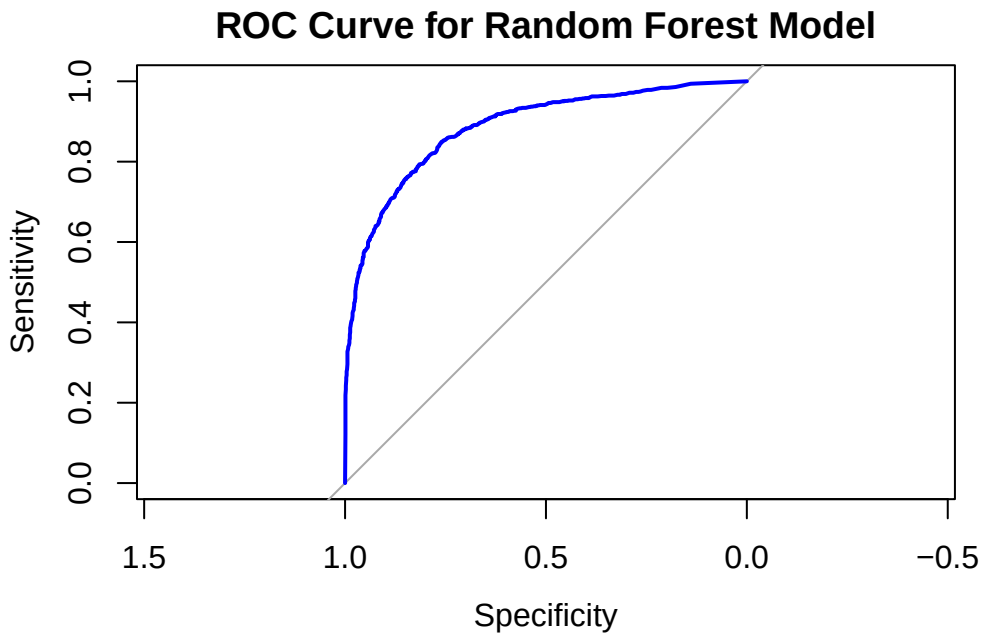
```
The following objects are masked from 'package:stats':

    cov, smooth, var
```

```
roc_curve <- roc(test_data$Clicked.on.Ad, rf_probabilities)
```

```
Setting levels: control = 0, case = 1
```

```
Setting direction: controls < cases
```

```
plot(roc_curve, col = "blue", main = "ROC Curve for Random Forest Model")
```



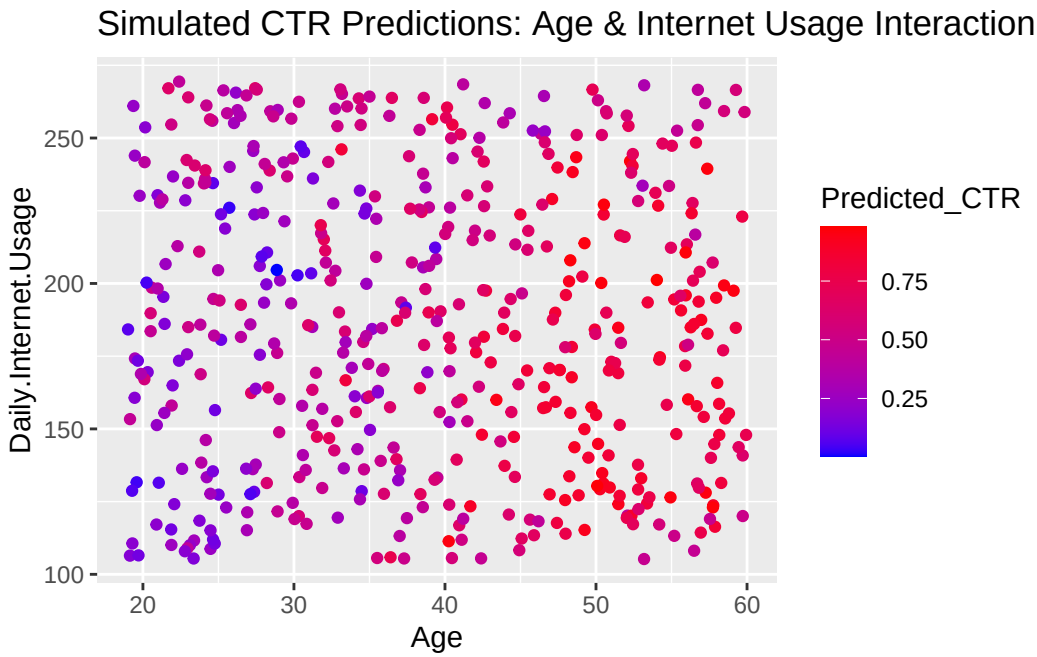
Accuracy = 0.8034 (80.34%). Since the blue ROC curve is significantly above this diagonal, it indicates that the Random Forest model performs well at distinguishing between users who click and those who do not.

Simulated Data Analysis

To further analyze CTR trends, a simulated dataset was created, generating 500 synthetic user profiles with random values drawn from real data distributions. The Random Forest model was applied to predict CTR probabilities, and the results show that CTR varied across age

groups, confirming that younger users are less likely to engage with ads.The plot show that there is no strong interaction effect was detected between Age and Internet Usage, indicating that these variables exert independent effects on CTR.

```
# Applying the Random Forest Model to Simulated Data for Further Analysis
set.seed(42)

# Generate simulated data
simulated_data <- data.frame(
  Daily.Internet.Usage = runif(500, min(df$Daily.Internet.Usage), max(df$Daily.Internet.Usage
  Age = runif(500, min(df$Age), max(df$Age)),
  Area.Income = runif(500, min(df$Area.Income), max(df$Area.Income))
)

# Create interaction term
simulated_data$Age_InternetUsage <- simulated_data$Age * simulated_data$Daily.Internet.Usage

# Generate random time-of-day categories
simulated_data$Hour <- sample(0:23, 500, replace = TRUE)
simulated_data$Time_of_Day <- case_when(
  simulated_data$Hour >= 6 & simulated_data$Hour < 12 ~ "Morning",
  simulated_data$Hour >= 12 & simulated_data$Hour < 18 ~ "Afternoon",
  simulated_data$Hour >= 18 & simulated_data$Hour < 24 ~ "Evening",
  TRUE ~ "Night"
)

# Generate random weekend indicator
simulated_data$Weekend <- sample(0:1, 500, replace = TRUE)

# Convert categorical variables to factors
simulated_data$Time_of_Day <- as.factor(simulated_data$Time_of_Day)
simulated_data$Weekend <- as.numeric(simulated_data$Weekend)

# Predict CTR on simulated data
simulated_data$Predicted_CTR <- predict(rf_model, simulated_data, type = "prob")[, 2]

# Visualization: Age & Internet Usage Interaction
ggplot(simulated_data, aes(x = Age, y = Daily.Internet.Usage, color = Predicted_CTR)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red") +
  labs(title = "Simulated CTR Predictions: Age & Internet Usage Interaction")
```

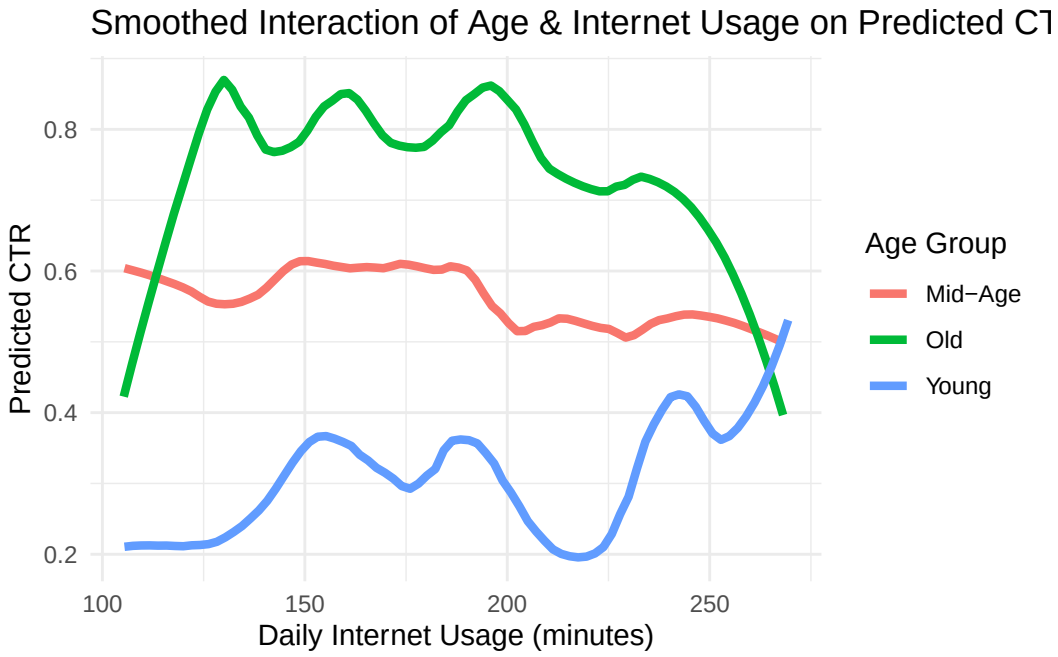



```
#segment users into different age categories
# Load necessary libraries
library(ggplot2)
library(dplyr)

# Create Age bins (Young, Mid-Age, Old)
simulated_data <- simulated_data %>%
  mutate(Age_Group = case_when(
    Age <= 30 ~ "Young",
    Age > 30 & Age <= 50 ~ "Mid-Age",
    Age > 50 ~ "Old"
  ))

# Aggregate CTR predictions by Age Group & Daily Internet Usage
simulated_summary <- simulated_data %>%
  group_by(Age_Group, Daily.Internet.Usage) %>%
  summarise(Avg_CTR = mean(Predicted_CTR), .groups = "drop")

# Smooth line plot using LOESS regression
ggplot(simulated_summary, aes(x = Daily.Internet.Usage, y = Avg_CTR, color = Age_Group)) +
  geom_smooth(method = "loess", span = 0.3, se = FALSE, size = 1.5) +  # Smooth trend line
  labs(title = "Smoothed Interaction of Age & Internet Usage on Predicted CTR",
       x = "Daily Internet Usage (minutes)",
```

```
       y = "Predicted CTR",
       color = "Age Group") +
  theme_minimal()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

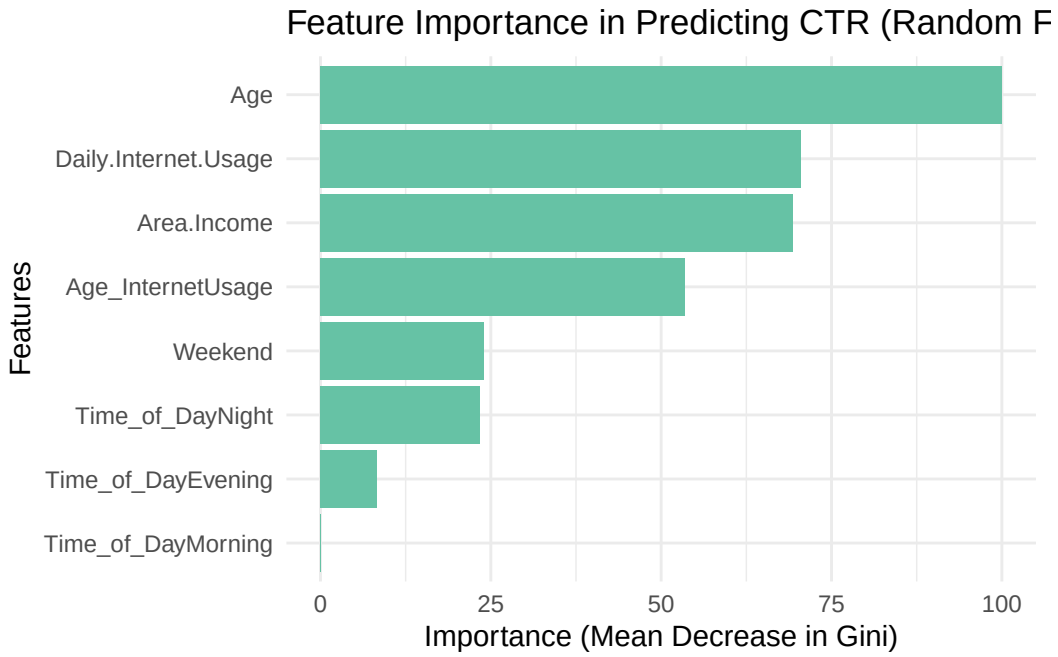
```
`geom_smooth()` using formula = 'y ~ x'
```



segment users into different age categories, aggregate predicted CTR values, and visualize the relationship between Daily Internet Usage and CTR across different age groups using LOESS regression smoothing.

```
# Extract feature importance
importance_df <- as.data.frame(varImp(rf_model)$importance)  # Extract importance values
importance_df$Feature <- rownames(importance_df)  # Add feature names

# Rename column for clarity
colnames(importance_df)[1] <- "Importance"

# Plot Feature Importance
ggplot(importance_df, aes(x = reorder(Feature, Importance), y = Importance)) +
```

```r
  geom_bar(stat = "identity", fill = "#66C2A5") +  # Set all bars to the same color (Green)
  coord_flip() +  # Flip to make it horizontal
  labs(title = "Feature Importance in Predicting CTR (Random Forest)",
       x = "Features",
       y = "Importance (Mean Decrease in Gini)") +
  theme_minimal()
```



Area Income (Most Important):Highest Mean Decrease in Gini → Strong influence on CTR. Age (Second Most Important): Significant impact on CTR. Daily Internet Usage (Moderate Importance): Less influence compared to Age & Area Income.

Discussion on the results of random forest

From the results, Age emerges as the most significant predictor, indicating that user age plays a critical role in determining ad engagement. This suggests that certain age groups are more likely to click on ads, possibly due to differing levels of digital engagement, trust in advertisements, or purchasing behavior. Daily Internet Usage and Area Income are also highly important. The importance of Daily Internet Usage suggests that online behavior patterns influence CTR in a way that users who spend more time online may either develop ad fatigue or be more exposed to tailored advertisements. Similarly, Area Income being a key predictor showing that socioeconomic factors affect how users respond to digital ads, with higher-income users potentially engaging less with online advertisements. The interaction term

is also moderate important, indicating that the effect of age on CTR is modified by how much time a user spends online. This supports the idea that digital habits interact with demographic factors in shaping ad engagement. Interestingly, time-of-day effects (Morning, Evening, Night) and Weekend status show relatively lower importance but are still relevant. The lower ranking suggests that while time-based behaviors influence ad clicks, they are not as crucial as user demographics and online behavior.

lasso regression

To better understand the relationship between users' demographics and their online behaviors with the dependent variable CTR, Lasso regression was performed to select important variables by shrinking the coefficient of less important to zero, removing unimportant variables from the model. Also, Lasso select fewer variables for better interpretability. The results show that Age, Daily internet use, and income are important variables in modeling CTR. the interaction term (Age $\times$ Internet Usage) was removed, suggesting it does not add predictive value beyond individual effects.

```
# 3. Lasso Regression for Feature Selection

# Reload dataset to avoid transformations from previous steps
df <- read.csv("ad_10000records.csv")
```

```
# Convert categorical variables
df$Clicked.on.Ad <- as.factor(df$Clicked.on.Ad)
df$Gender <- as.factor(df$Gender)

# Create interaction term
df$Age_InternetUsage <- df$Age * df$Daily.Internet.Usage

# Prepare data for Lasso Regression
X <- model.matrix(Clicked.on.Ad ~ Daily.Internet.Usage + Age + Area.Income + Age_InternetUsa
y <- as.numeric(as.character(df$Clicked.on.Ad))  # Convert factor to numeric
```

```
# Perform Lasso Regression
set.seed(42)
lasso_model <- cv.glmnet(X, y, alpha = 1, family = "binomial")
```

```
# Best lambda
best_lambda <- lasso_model$lambda.min
print(paste("Best Lambda for Lasso:", best_lambda))
```

```
[1] "Best Lambda for Lasso: 0.00200181033208852"
```

```r
# Extract important features
lasso_coefs <- coef(lasso_model, s = best_lambda)
selected_features <- rownames(lasso_coefs)[which(lasso_coefs != 0)]
print("Selected Features:")
```

```
[1] "Selected Features:"
```

```r
print(selected_features)
```

```
[1] "(Intercept)"          "Daily.Internet.Usage" "Age"
[4] "Area.Income"
```

Features not listed here were removed by Lasso, meaning they did not contribute significantly to predicting whether a user clicked on an ad. the interaction between age and internet usage does not add significant predictive value beyond the individual features.

OLS regression

To further understand its statistical significance and inference, OLS model was preformed after Lasso. Since lasso doesn't naturally provide p-values or confidence intervals. By following with OLS on the selected variables, we can gain access to these standard statistical inference tools. Also, we get a simpler model from Lasso variable selection while still having statistically valid coefficients and inference measures. The result show that Age positively correlated with CTR, reinforcing that older users engage more with ads. Daily Internet Usage negatively correlated with CTR, which might be attributed to overexposure leading to fatigue. Higher Income was associated with lower CTR, indicating that wealthier individuals are either less interested in online ads or engage with them differently.

```r
# Prepare Data for OLS Regression
selected_features <- selected_features[-1]  # Remove intercept
formula <- as.formula(paste("Clicked.on.Ad ~", paste(selected_features, collapse = " + ")))
```

```r
# Fit OLS Model
ols_model <- glm(formula, data = df, family = "binomial")
summary_ols <- summary(ols_model)
print(summary_ols)
```

```
Call:
glm(formula = formula, family = "binomial", data = df)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.424e+00  1.779e-01 -19.244  < 2e-16 ***
Daily.Internet.Usage -4.552e-03  5.540e-04  -8.217  < 2e-16 ***
Age                   1.287e-01  3.143e-03  40.939  < 2e-16 ***
Area.Income          -7.495e-06  1.675e-06  -4.474 7.68e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13860  on 9999  degrees of freedom
Residual deviance: 11437  on 9996  degrees of freedom
AIC: 11445

Number of Fisher Scoring iterations: 3
```

Intercept (-3.424): The baseline log-odds of clicking an ad when all predictor variables are zero. Since it's negative, the probability is quite low at the baseline.

Daily Internet Usage Negative coefficient suggests that as internet usage increases, the likelihood of clicking an ad decreases. This could imply ad fatigue—users spending more time online are less likely to engage with ads.

Age Positive coefficient means older individuals are more likely to click on ads. This could indicate that older users are more engaged with online ads or are part of a demographic that responds more to them.

Area Income Negative coefficient suggests that individuals from higher-income areas are less likely to click ads. This could indicate wealthier individuals are either less interested in online ads or engage with them differently.

All predictor variables have p-values < 0.001, which means they are highly significant in predicting ad clicks.

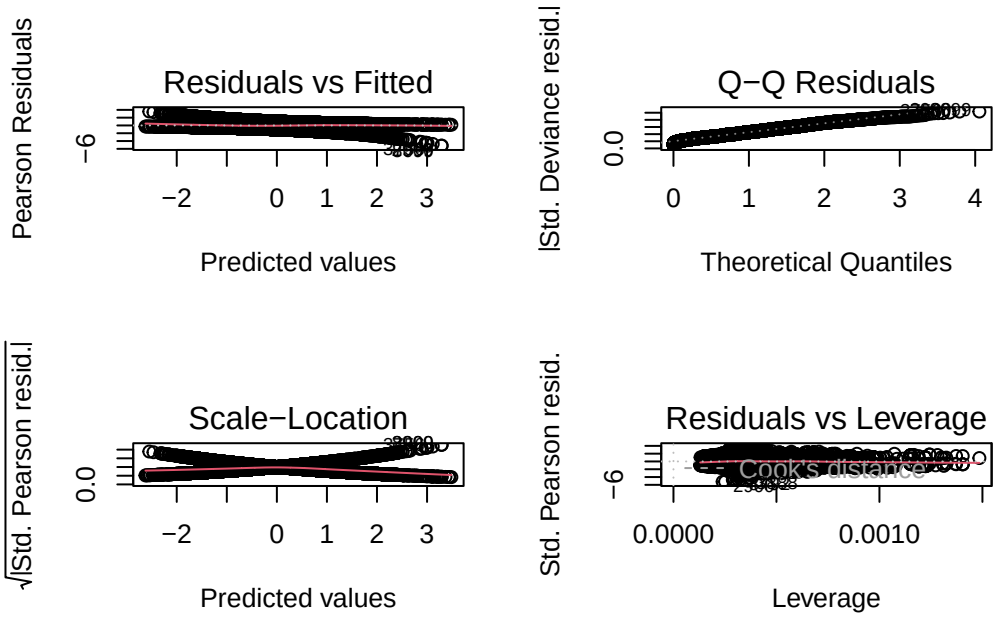
```r
# Extract Coefficients and P-values
coefficients_df <- tidy(ols_model)
print("OLS Coefficients and P-values:")
```

```
[1] "OLS Coefficients and P-values:"
```

```
print(coefficients_df)
```

```
# A tibble: 4 x 5
  term                    estimate  std.error statistic  p.value
  <chr>                      <dbl>      <dbl>     <dbl>    <dbl>
1 (Intercept)          -3.42       0.178         -19.2  1.57e-82
2 Daily.Internet.Usage -0.00455    0.000554       -8.22 2.08e-16
3 Age                   0.129      0.00314        40.9  0
4 Area.Income          -0.00000749 0.00000168     -4.47 7.68e- 6
```

```
# Model Diagnostics
par(mfrow = c(2, 2))
plot(ols_model)
```



8. Results and Findings

The feature importance analysis from the Random Forest model identified Age as the most influential predictor, followed by Daily Internet Usage and Area Income. This suggests that older users tend to engage more with ads, while users with higher internet usage and income levels engage less. The interaction term (Age_InternetUsage) was moderately important, indicating that age influences CTR differently depending on online behavior. Time-of-Day and

Weekend effects were found to have lower predictive power, suggesting that while temporal patterns exist, they are not the dominant drivers of ad engagement. Lasso regression and OLS models helped refine feature selection, confirming that Age positively correlates with CTR, while higher-income users and those with higher daily internet usage are less likely to click ads. The negative effect of Daily Internet Usage suggests a possible ad fatigue phenomenon, where users who spend more time online become desensitized to advertisements. Interestingly, the interaction term between Age and Internet Usage was removed in Lasso regression, reinforcing the idea that these variables exert independent effects rather than working together to influence CTR. Together, these results indicate that CTR are driven by demographic and behavioral tendencies rather than simple internet usage patterns. Younger users may be less responsive to ads due to digital saturation, while older users exhibit higher engagement, potentially due to increased curiosity or product relevance. The role of income as a negative predictor suggests that higher-income individuals may be more selective in their engagement with online ads. Future research could explore additional psychological and contextual factors, such as ad content personalization or cognitive engagement levels across age groups, to refine our understanding of user behavior in digital advertising.

9. Final Model Selection and Discussion

The Random Forest model emerged as the best-performing model, providing both high accuracy (80.3%) and implications through feature importance rankings. This model effectively captured nonlinear relationships and feature interactions, making it the most suitable choice for predicting CTR. However, Lasso Regression and OLS Regression also contributed valuable insights by refining feature selection and explaining the specific relationship between key variables and the CTR. While Lasso helped identify the most predictive variables, OLS regression provided statistical validation, reinforcing the relationship between age, internet usage, and ad engagement. Several important insights emerged from random forest analysis. Age was found to be the strongest predictor of CTR, with middle-aged users (30–50) engaging with ads the most. This suggests that advertisers should tailor their campaigns toward this demographic, as they are the most responsive to digital ads. Additionally, the results might be attributed to the presence of ad fatigue, where higher daily internet usage correlates with lower CTR. Users who spend extended time online may become desensitized to advertisements, reducing engagement. This insight has particular implication for advertisers, as it highlights the importance of ad placement, frequency, and content personalization to mitigate the effects of ad fatigue. Economic factors such as cincome level also play a crucial role in ad engagement. Higher-income users were found to be less likely to click on ads, suggesting that wealthier individuals engage with advertisements differently. This may indicate that higher-income users prefer non-intrusive or highly personalized ad experiences. Finally, while temporal effects (time-of-day and weekends) had a minor but notable impact, they were not as influential as age and behavioral patterns. This finding suggests that while ad engagement varies slightly across different times of the day, it is not the primary driver of CTR. These findings provide empirical evidence that user demographics and online behavior significantly impact CTR. The

study confirms that age, daily internet usage, and income levels drive engagement patterns, making them critical factors in ad targeting strategies. The interaction term between Age and Daily Internet Usage was found to be insignificant. This research aligns with previous studies on CTR prediction, particularly in highlighting that middle-aged users are the most engaged and that high internet usage can lead to lower ad interaction. Future research should explore advanced deep learning models, such as the Deep Interest Evolution Network (DIEN), which can capture long-term behavioral trends and evolving user preferences. Incorporating personalized ad-serving mechanisms based on individual engagement histories could further enhance ad targeting effectiveness. CTR prediction is a long-studying problem that requires an in-depth understanding of demographics, online behavior, and ad interaction patterns. By leveraging Random Forest with refined feature selection, this study provides valuable insights into how different user groups engage with online advertisements. These findings have practical implications for digital marketing strategies, which can help advertisers optimize ad targeting and content optimization to maximize user engagement. Moreover, integrating deep learning approaches might further refine CTR prediction models.Incorporating ad-specific features, such as ad content, format, and placement, could provide a more comprehensive understanding of CTR prediction as well as what drives ads engagement.

References

Chen, Leqi. 2023. "Research on Advertising Click-Through Rate Prediction Model Based on Taobao Big Data." *Highlights in Science, Engineering and Technology* 56 (July): 179–87. https://doi.org/10.54097/hset.v56i.10102.

Qin, Jiarui, Weinan Zhang, Xin Wu, Jiarui Jin, Yuchen Fang, and Yong Yu. 2020. "User Behavior Retrieval for Click-Through Rate Prediction." In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2347–56. Virtual Event China: ACM. https://doi.org/10.1145/3397271.3401440.

Zhou, Guorui, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. "Deep Interest Evolution Network for Click-Through Rate Prediction." *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (01): 5941–48. https://doi.org/10.1609/aaai.v33i01.33015941.