Click on Ads Prediction Analysis for Making Business Decisions

DECISION 520Q Section B Team 18

Yinan Chen, Tirth Pravin Gala, Yuhe (Tiffany) Jin, Chenjie (Angelina) Sun

1. Business Understanding

As the data analyst team in the marketing department, we specialize in analyzing the challenge of customer churn within the telecommunication industry. Churn, or the loss of customers to competitors, directly impacts revenue and raises acquisition costs. Each lost customer poses a potential threat to others through negative feedback. Our strategy prioritizes early identification of potential churners and improving our advertising for maximum retention rate.

We use data mining to draw connections between click rates and customer profiles so as to evaluate the advertisement performance. A disparity between high click rates and rising churn might indicate a misalignment between our advertising promises and the actual service experience. We employ logistic regression, LASSO, decision trees, and CART models to do the data mining process, directly addressing our telecom challenges. With tools like predictive modeling, we can forecast customer departures, and segmentation provides insights into varied customer behaviors, facilitating precise retention strategies. Correlation analysis further helps identify popular product or service combinations to inform our offerings. Decision Trees and CART will offer interpretable insights for tailored retention. LASSO helps us to select crucial variables, focusing resources where they matter most. Our goal goes beyond simply problem identification; we strive to develop well-informed solutions supported by data. Through a detailed data mining process and emphasizing data in our decision-making, we aim for optimized marketing campaigns and increased ad effectiveness, ultimately raising the bar for our marketing standards.

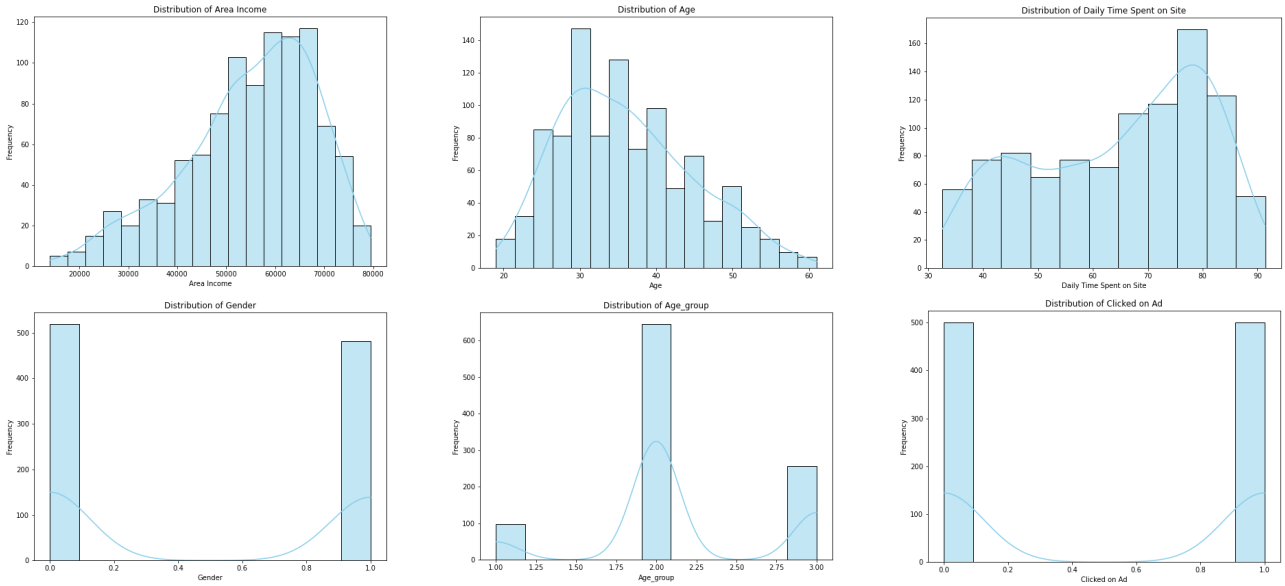
2. Data Understanding

2.1 Data Source

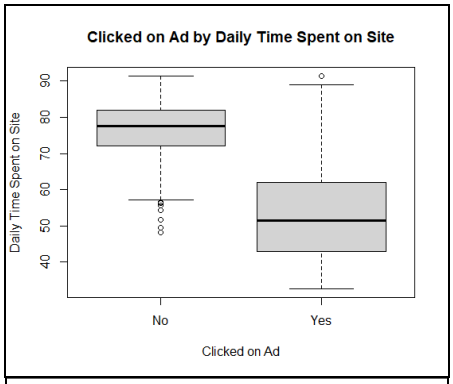
The Advertisement - Click on Ad dataset data is retrieved from *Kaggle*. The dataset has 10 variables describing the demographic features and user habits of the target audience for digital ads. Following is the list of variables along with its data type: *Daily.Time.Spent.on.Site* (numeric), *Age* (integer), *Area.Income* (numeric), *Daily.Internet.Usage* (numeric), *Ad.Topic.Line* (character), *Male* (integer), *City* (character), *State* (character), *Country* (character), *Timestamp* (character), *Clicked.on.Ad* (integer).

2.2 Exploratory Data Analysis (EDA)

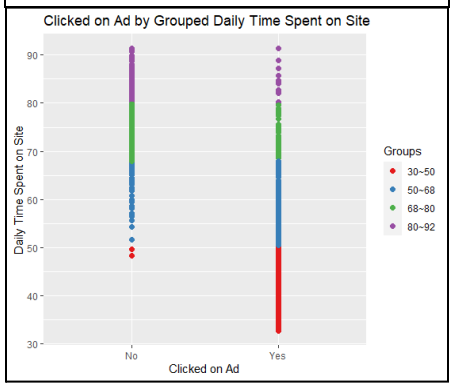
To construct a model for prediction the Click-on Rate of a certain ad, we look closely at the dataset. For the variable, Clicked on Ad, there are 500 records of clicking-on and 500 records of not-clicking-on. The three variables, Area Income, Daily Time Spent on Site, and Age, are continuous variables, and are distributed reasonably. For the dummy variable, Gender, there are slightly more females than male, which will not cause significant bias to our analysis. Generally speaking, our dataset is a representative display of the situation in real life.



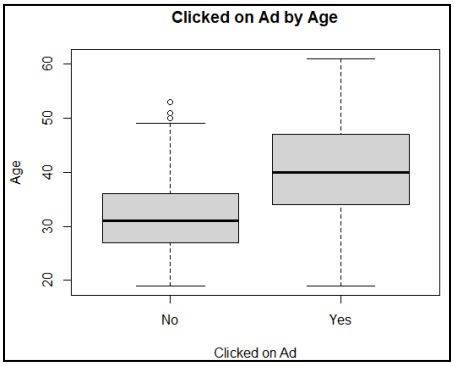
Daily Time Spent on Site



The variable, Daily Time Spent on Site, can be a good indicator. According to the boxplot, people who clicked on the ad and those who did not are quite different. The boxes for 'No' and 'Yes' are far from each other. Moreover, there is a significant difference between clicking on an ad across different groups of Daily Time Spent on the Site. In the scatterplot, the density of a certain group under 2 different situations, whether people click on the ad or not, is distinct. For instance, for people falling into the group 30-50 minutes daily time on site, apparently more people clicked on the ad than those who did not.

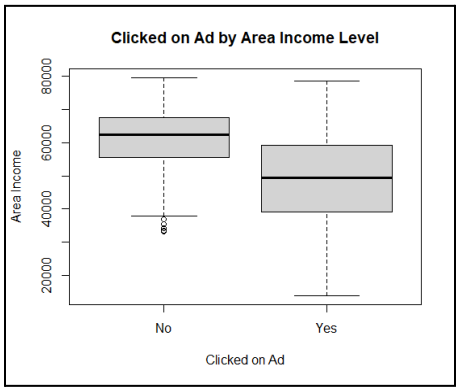


Age



The Age variable against Clicked on Ad also indicates that there is a significant difference between those who clicked on an ad or not regarding their age. Referring to the boxplot, we can infer that those who did not click on an ad are generally younger than the group who clicked. And thus, the variable, Age, can be an informative factor variable.

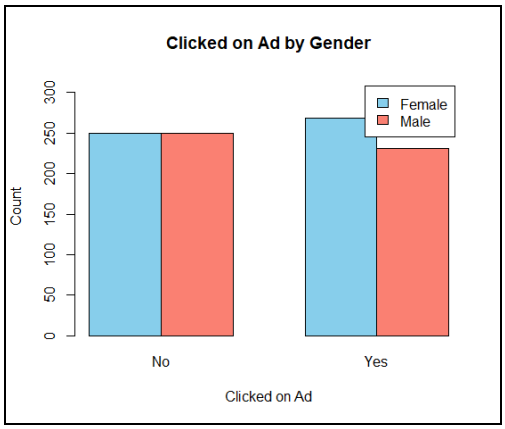
Area Income



A significant difference in area income exists between people who clicked on an ad or not. People in areas with a lower income level seem more likely to click on an ad displayed to them. This variable is also an informative one, taking into consideration some geographic and economic concerns.

Gender

Gender, as a variable, has an almost equal number of females and male, showing that the sample is representative to some extent. For those who clicked on an ad, females tend to be more likely to click than males, while females and males are performing similarly in not clicking on an ad. In a word, gender has some impact on people's choice of clicking on an ad.



3. Data Preparation

To prepare the data for the next-step data mining process, we first drop the column of *Daily.Internet.Usage*. The first column shows the time spent by each viewer on site. Thus, each viewer's Internet usage becomes an incompetent column. It correlates and rather gives less information about the viewer than the first column. Thus, as this column has no real informative value, it is dropped. We also rename the column of *Daily.Time.Spent.on.Site* to be *Time_Spend* with continuous values and change the column of *Male* to *Gender* with binary values with 1 being males and 0 being females.

4. Modeling

4.1 Core Task

Our core task is to build a model to predict the click rate of digital ads *(Clicked on Ad)* given the observed features of the target audience.

4.2 Data Splitting

In the data preparation process, we split the data into two sets and randomly assign observations

to the training set with a 70% probability and to the testing set with a 30% probability. The variables *Country, City, Daily Internet Usage, and Ad Topic Line* are excluded in both data sets.

4.3 Models

Since our target variable *(Clicked on Ad)* is a binary variable with its value to be 0 and 1, we employ the following 4 models to find the best model for our predictive task:

1) Logistic Regression model
2) Decision Tree model
3) CART model
4) LASSO model

For the logistic regression model, it produces probability estimates with the binary target variable with high interpretability and efficiency, but it can't explain the non-linear relationships.

For the decision tree model, while it better captures complex, non-linear relationships and handles both categorical and numerical data, the model has the risk of overfitting and high variance.

For the CART model, it's highly interpretable and can be easily visualized but is less stable and harder to reproduce.

For the LASSO model, it better prevents overfitting problems but may not capture complex, non-linear relationships in the data as effectively as other non-linear models.

5. Evaluation

5.1 Evaluation Metric

To choose the best model that fits this data set for prediction, we apply a 5-fold cross validation for each model and compare both the in-sample and out-of-sample accuracy as evaluation metric.

	In-Sample Accuracy	Out-of-Sample Accuracy
Logistic Regression	0.9206	0.9255
Decision Tree	0.7998	0.8315
CART	0.8374	0.8122
LASSO	0.9168	0.9205

5.2 Final Model

By comparing both the in-sample and out-of-sample accuracy across all models, we select the Logistic Regression Model as our final choice since it has the highest accuracy that ensures its high-quality predictive performance. Based on the logistic regression model, we predict the probability of clicking on the ads that fall within the range from 0 to 1, where 0 represents the probability of not clicking on the ad, and 1 represents the probability of clicking on the ad.

Mathematical representation of Logistic Regression can as follows:

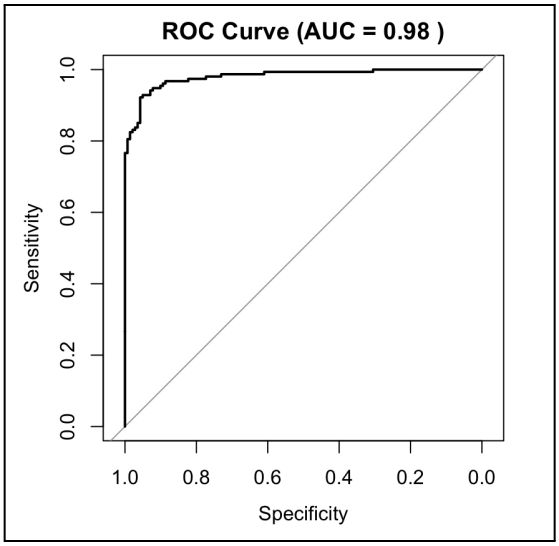
$$log(p(Y = 1) / (1 - p(Y = 1))) = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \beta 5 * X5$$

Here, P(Y=1) is the probability that Y equals 1 (i.e., the probability of clicking on the ad).
β0 is the intercept of the model, representing the log-odds of Y=1 when all predictor variables are zero. β1, β2, β3, β4, and β5 are the coefficients associated with the respective predictors X1, X2, X3, X4, and X5. These coefficients represent the change in the log-odds of Y=1 for a one-unit change in the corresponding predictor while holding all other predictors constant.

5.3 ROC Curve & AUC

To validate the predictive performance , we create an ROC curve and calculate the AUC of the logistic model to evaluate the model's discrimination ability.

From the visualization below, the classifiers that give curves closer to the top-left corner indicate a better performance. An AUC of 0.98 also indicates perfect discrimination between positive and negative instances and is considered very indicative in most binary classification scenarios.



6. Deployment

6.1 Deployment of the Data Mining Result

We have an advertisement to be placed online to improve our brand awareness and people's understanding of a certain product or service. With a limited budget, it is important to choose the websites where the ad can bring in the highest Clicks, and place our ad on those potentially rewarding websites. To determine which websites to display our ad, we feed the data featuring characters of visitors of a certain website into our model, and predict a *Clicked on Ad* Rate for the exact ad. Among all the websites tested, we will finally choose the websites with the highest *Clicked on Ad* Rate to ensure a more profitable investment.

6.2 Issues to be Aware of

Biases and Inaccurate Results: The model we constructed through data analytics might contain biases due to the inherent biases in the dataset used for training.

Model Robustness and Generalization: Our model might not perform optimally in real-life scenarios because the dataset used for its training is somewhat limited. To enhance the model's robustness and its performance in real-world situations, additional datasets are needed to account for variations and anomalies.

Dramatic Changes in the Market: The model is vulnerable to specific incidents, such as significant shifts in technology and market trends. These can introduce additional biases into the model, leading to potentially misleading predictions.

6.3 Ethical Consideration

The models we constructed are based on specific datasets. The process of obtaining this data can raise certain ethical concerns. At times, the dataset used for training may not be representative of the entire population, resulting in a biased model.

Furthermore, determining when and where to display our ads can lead to unforeseen ethical dilemmas. The audiences of various websites differ greatly. Simply choosing to display our ads on certain sites can pose ethical challenges. Some audience groups might not access the content, while others might be overly exposed to the ads. Overall, ads placement requires thoughtful consideration to minimize potential ethical concerns.

6.4 Other Risk and Proposed Solution

Our model doesn’t account for the unique attributes of a specific ad. Elements such as the ad’s design, its slogan, the script, and the timing of its display can significantly influence people's responses to an online ad, beyond the general features our model considers. To address these

shortcomings, we recommend conducting A/B testing on ad placement platforms. Before rolling out our ad widely online, we can select the ad version with the highest click-through rate as the final version for broad display. This approach aims to minimize risks and maximize the utility of our model.

7. Conclusion with Insights

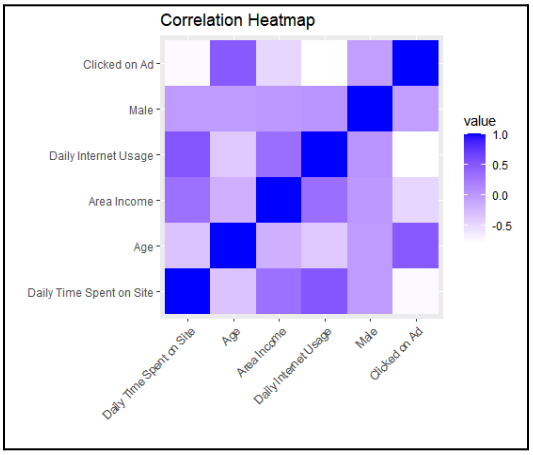
In this report, we aim to comprehend viewing behaviors related to telecom advertisements on third-party websites. We sourced our data from the esteemed *Kaggle* platform, then meticulously cleaned and restructured it to focus on the most relevant variables for our analytical endeavors. We evaluated various models, including logistic regression, decision tree, CART, and LASSO, to measure the likelihood of ad clicks.

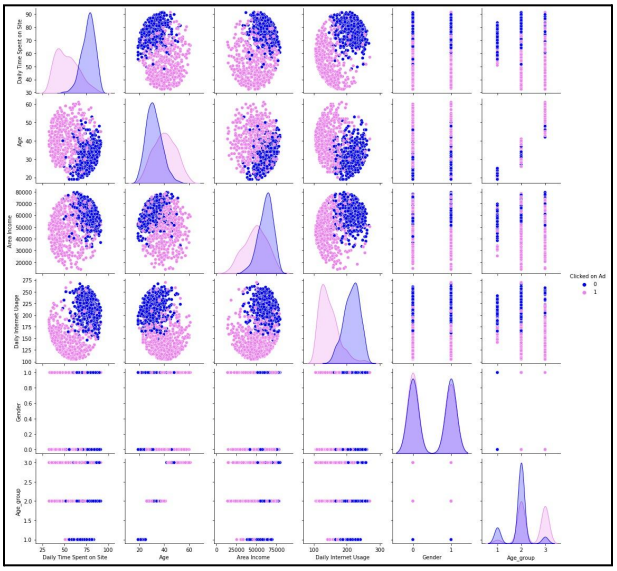
A significant finding from our study highlights the importance of the total click count over the click-through rate. Due to the absence of impression data in our dataset, we emphasized raw click counts as they offer a more direct measure of user engagement and potential sales conversions.

After rigorous cross-validation, the logistic regression model emerged as the best performer, boasting remarkable accuracy metrics. This was further confirmed by an exemplary AUC of 0.98 on the ROC curve, emphasizing the model's exceptional ability to differentiate between positive and negative outcomes. Beyond providing the telecom company with valuable insights into potential customer interactions, this analysis accentuates the crucial role that advanced predictive modeling plays in refining digital advertising strategies.

Appendix

I. Correlation Heatmap & Pairplot





II. Contribution:

Yuhe (Tiffany) - Data Modeling & Evaluation, Report write-up (Modeling, Evaluation)

Tirth - Data Cleaning, Report write-up (Data understanding, Evaluation), Presentation

Yinan- Report write-up (Business Understanding, Conclusion), Presentation, Presentation slides

Chenjie (Angelina) - EDA, Report write-up (Deployment), Presentation

Caryl - N/A……