Predicting Ad Click-Through Rates Using Logistic Regression

Abstract

In the digital age, understanding customer behavior is crucial for optimizing online advertising strategies. This project leverages logistic regression to predict whether users will click on advertisements based on demographic and behavioral data. The dataset used includes features such as age, gender, daily time spent on a website, area income, and daily internet usage. Using Python libraries like Pandas, Seaborn, and Scikit-learn, exploratory data analysis (EDA) and predictive modeling were conducted. The results include insights from visualization and a classification report evaluating the model's performance, demonstrating the utility of machine learning in targeted advertising.

Contents

1 Introduction

Digital advertisements are a key driver of revenue for online businesses. Predicting whether a user clicks on an ad allows marketers to design more effective campaigns. This project aims to analyze a dataset of user behavior and demographic characteristics to build a logistic regression model that predicts ad clicks.

2 Dataset Description

The dataset comprises the following features:

- **Daily Time Spent on Site**: Average time users spend on the website daily (in minutes).
- **Age**: User's age (in years).
- **Area Income**: Average income of the user's geographical area.
- **Daily Internet Usage**: Average daily internet usage (in minutes).
- **Male**: Gender indicator (1 for male, 0 for female).
- **Clicked on Ad**: Target variable (1 if the ad was clicked, 0 otherwise).

3 Methodology

3.1 Exploratory Data Analysis (EDA)

- Summary statistics and dataset information were reviewed.
- Visualizations were created using Matplotlib and Seaborn, including histograms, joint plots, and pair plots to understand feature distributions and relationships.

3.2 Feature Selection

Features deemed most relevant were selected for model building:

- Daily Time Spent on Site

- Age
- Area Income
- Daily Internet Usage
- Male

3.3 Model Building

- The dataset was split into training and testing sets (67
- A logistic regression model was built and trained using the training set.
- Predictions were made on the test set.

4 Results

4.1 Visualizations

- Age distribution showed a concentration in the younger demographic.
- Joint plots revealed correlations, such as higher daily site time corresponding to higher ad clicks.
- Pair plots indicated patterns across multiple features relative to the target variable.

4.2 Model Performance

- **Classification Report:**
  - Precision, recall, F1-score, and support metrics were calculated for each class.
  - The logistic regression model performed effectively, achieving a balance between precision and recall.

5 Conclusion

The project demonstrates the effectiveness of logistic regression in predicting ad click-through rates. The model's performance highlights its potential for deployment in real-world advertising platforms, enabling data-driven marketing strategies. Future improvements could include testing additional algorithms or incorporating more nuanced behavioral data.

6 References

- Dataset source (if applicable).
- Scikit-learn documentation.
- Python Data Science Handbook.