Scoring 3 Project : Ad Click Prediction

Maria Bracque Vendrell

November 23, 2023

Introduction

- Billions of dollars are spent every year on online advertisements
- Important to help firms to target consumers more efficiently
- Goal: To create different prediction models to predict whether a user will click on an ad based on the characteristics of the user

Outline

1. Dataset
2. Cleaning
3. Data Preparation for Exploratory Analysis
4. Exploratory Data Analysis
5. Data Preparation for Prediction Models
6. Logistic Regression
7. Random Forest
8. XGBoost

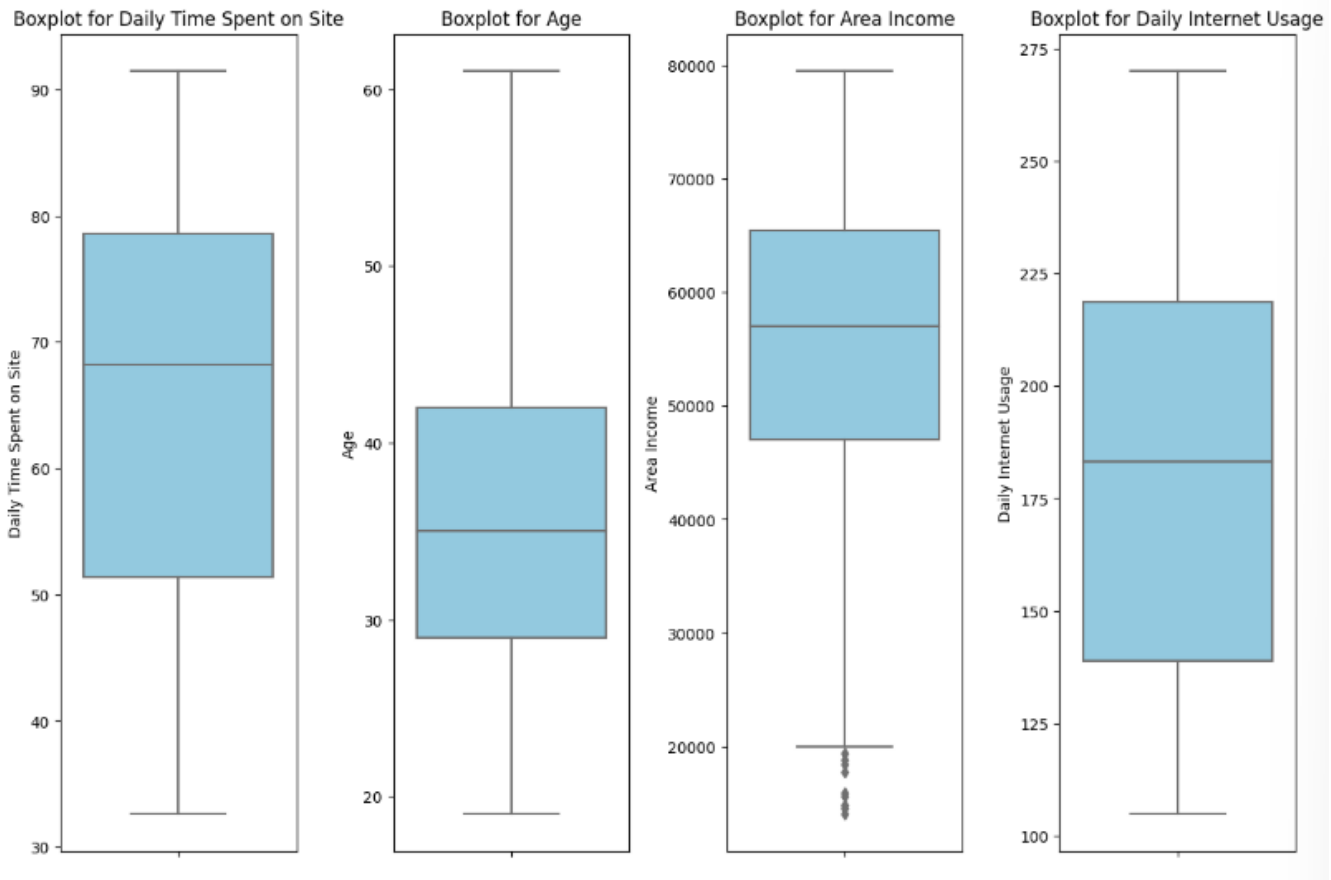
Conclusion

1. Dataset

- Source: Anonym marketing agency
- Population: 1000 internet users
- Data collected between 01/01/2016 and 24/07/2016
- 10 variables:
  - Daily Time Spent on Site
  - Age
  - Area Income
  - Daily Internet Usage
  - Ad Topic Line
  - City
  - Gender
  - Country
  - Timestamp
  - Clicked on Ad

2. Cleaning

- No missing data
- No duplicated data
- We don't delete any outlier

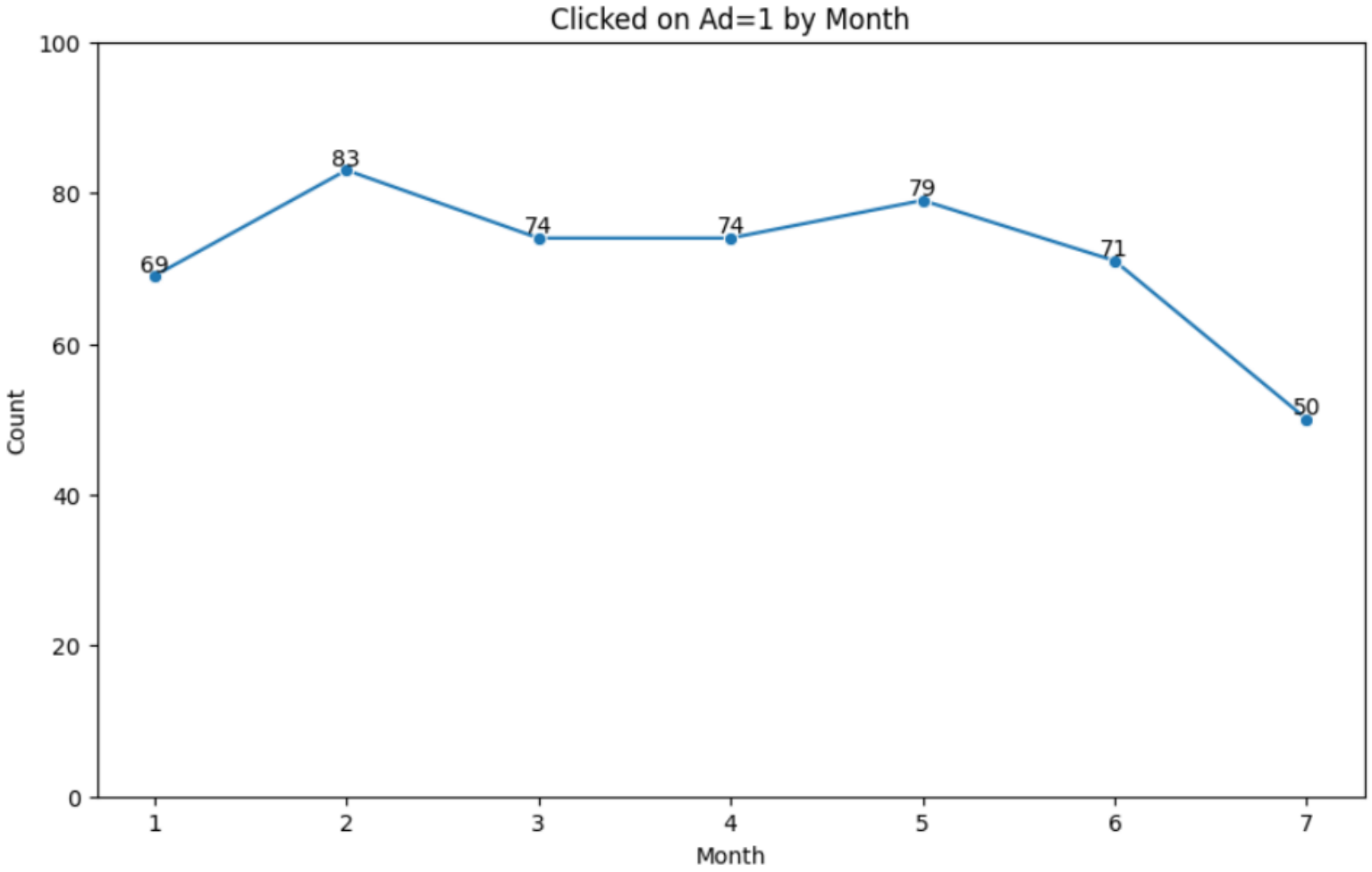


3. Data Preparation for Exploratory Analysis

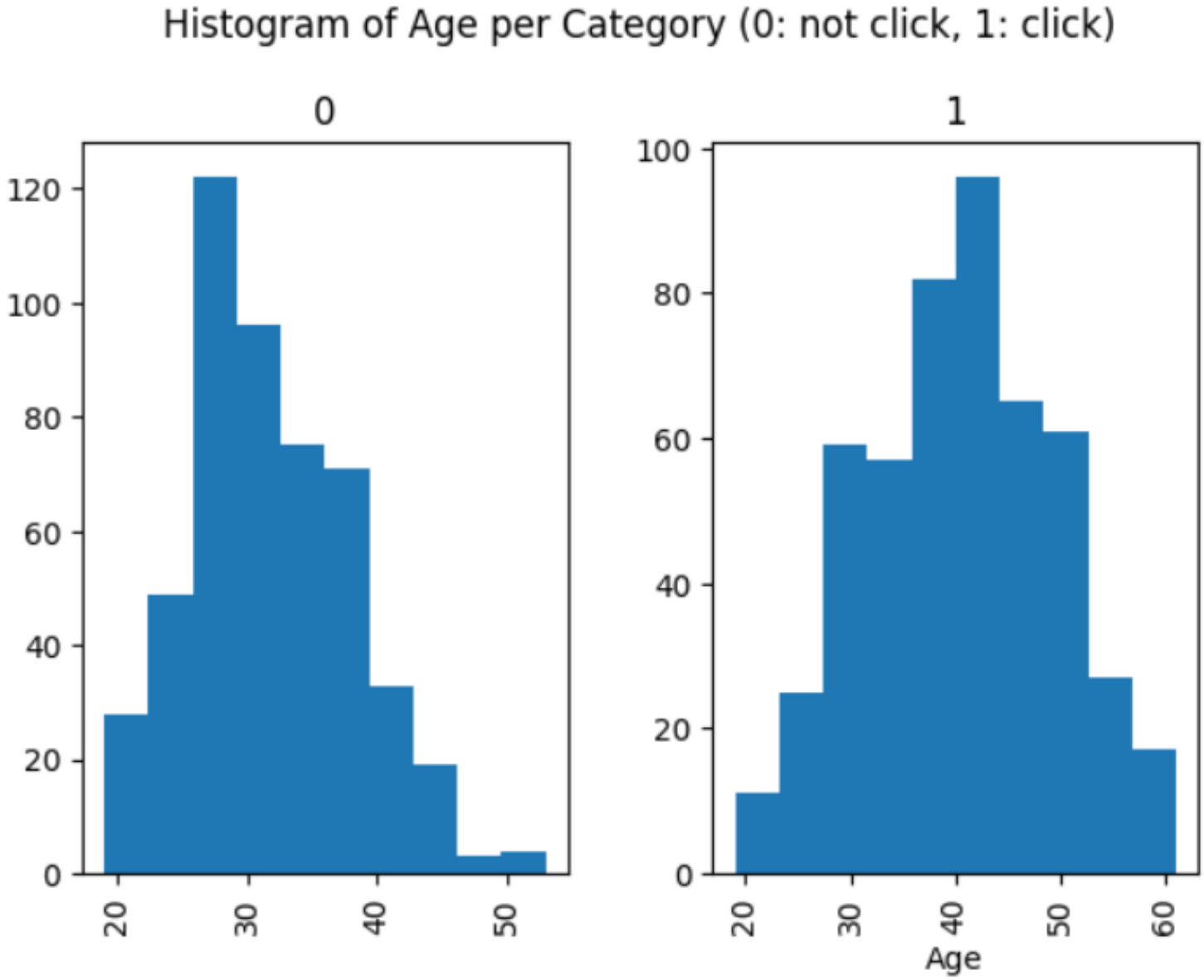
- Change the Timestamp variable to a datetime object
- Creation of new variables: Hour, DayOfWeek, Month, Date

Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Timestamp	Clicked on Ad	Hour	DayOfWeek	Month	Date
35	61833.90	256.09	Cloned 5thgeneration orchestration	Wrightburgh	0	Tunisia	2016-03-27 00:53:11	0	0	6	3	2016-03-27
31	68441.85	193.77	Monitored national standardization	West Jodi	1	Nauru	2016-04-04 01:39:02	0	1	0	4	2016-04-04

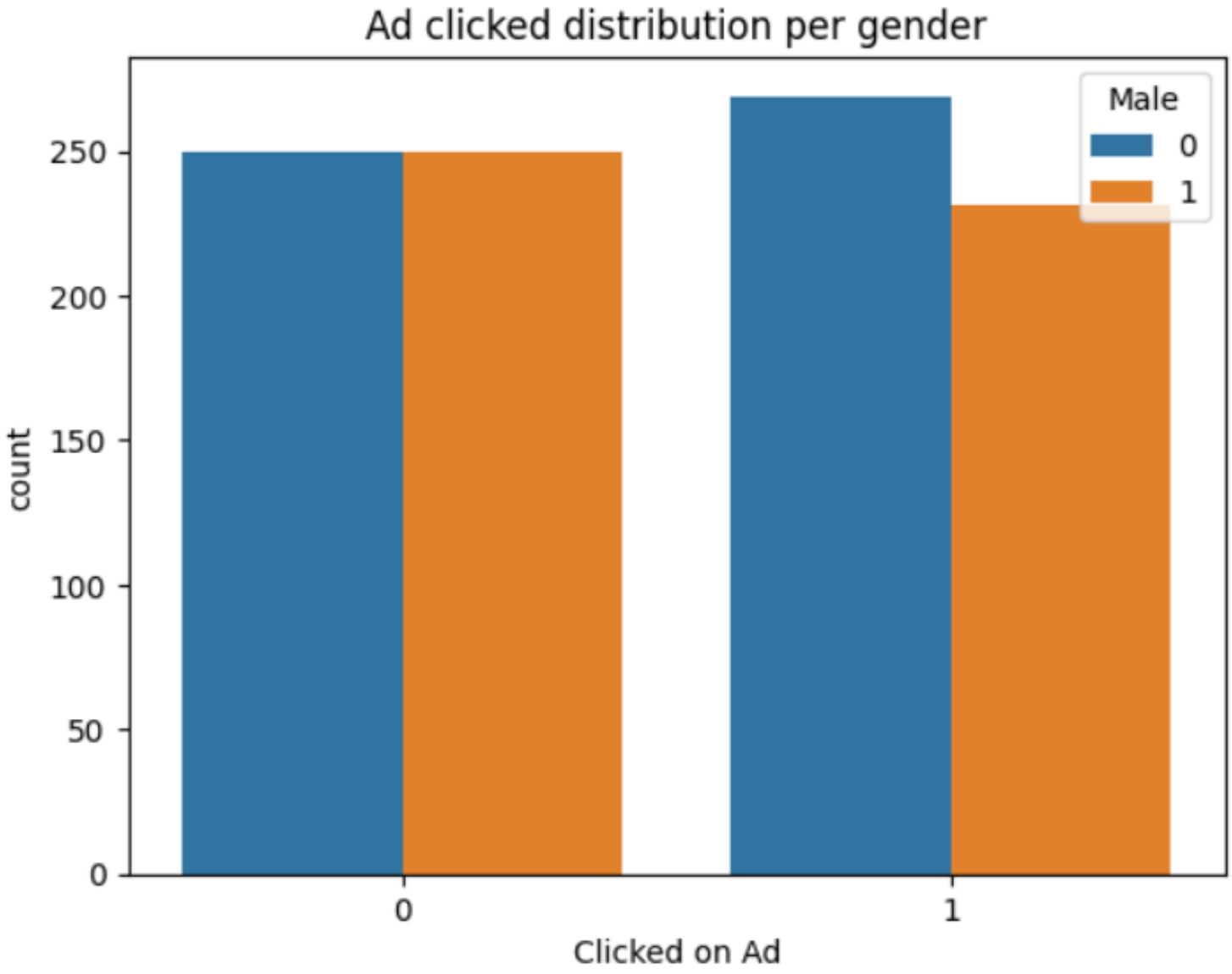
4. Exploratory Data Analysis



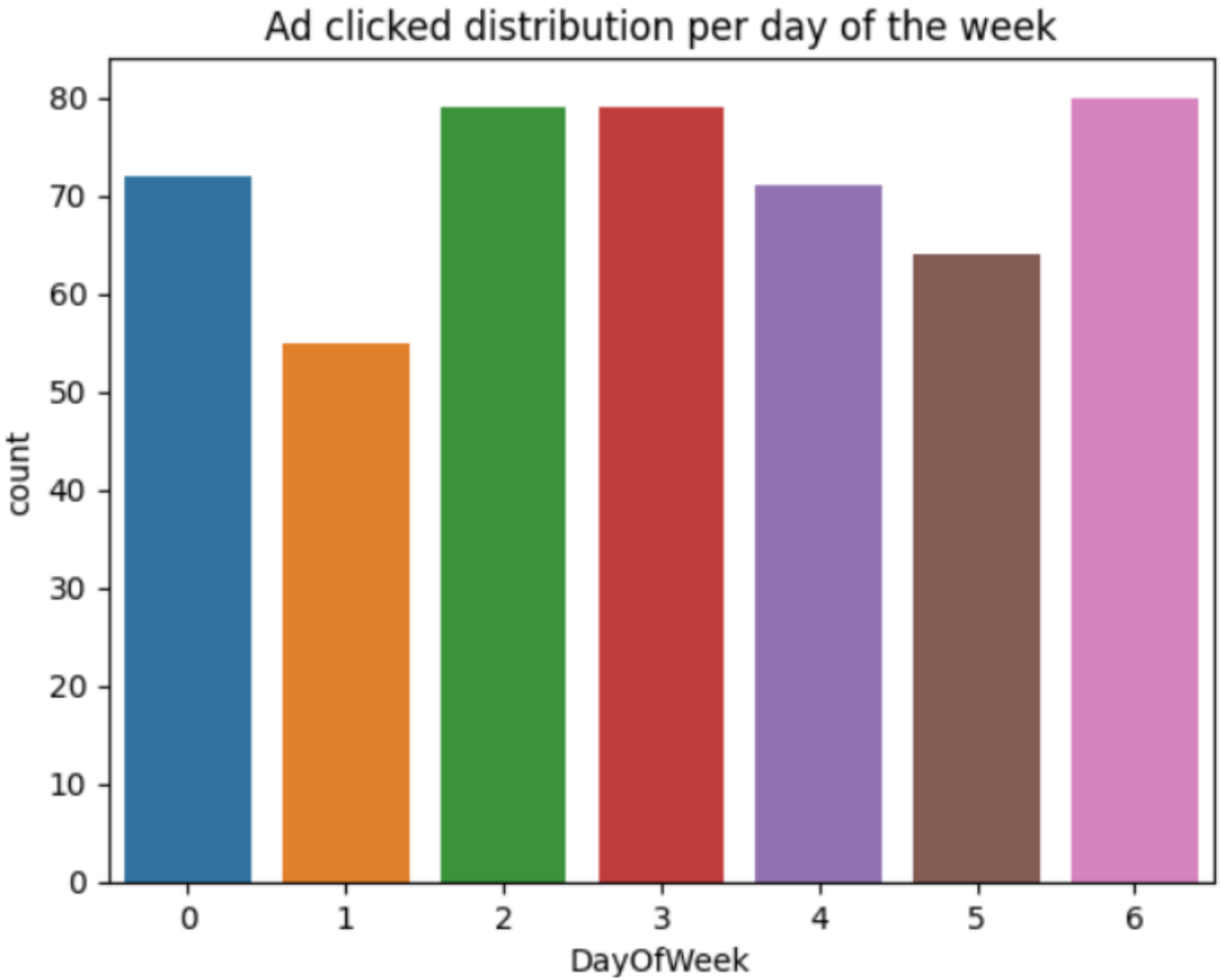
4. Exploratory Data Analysis



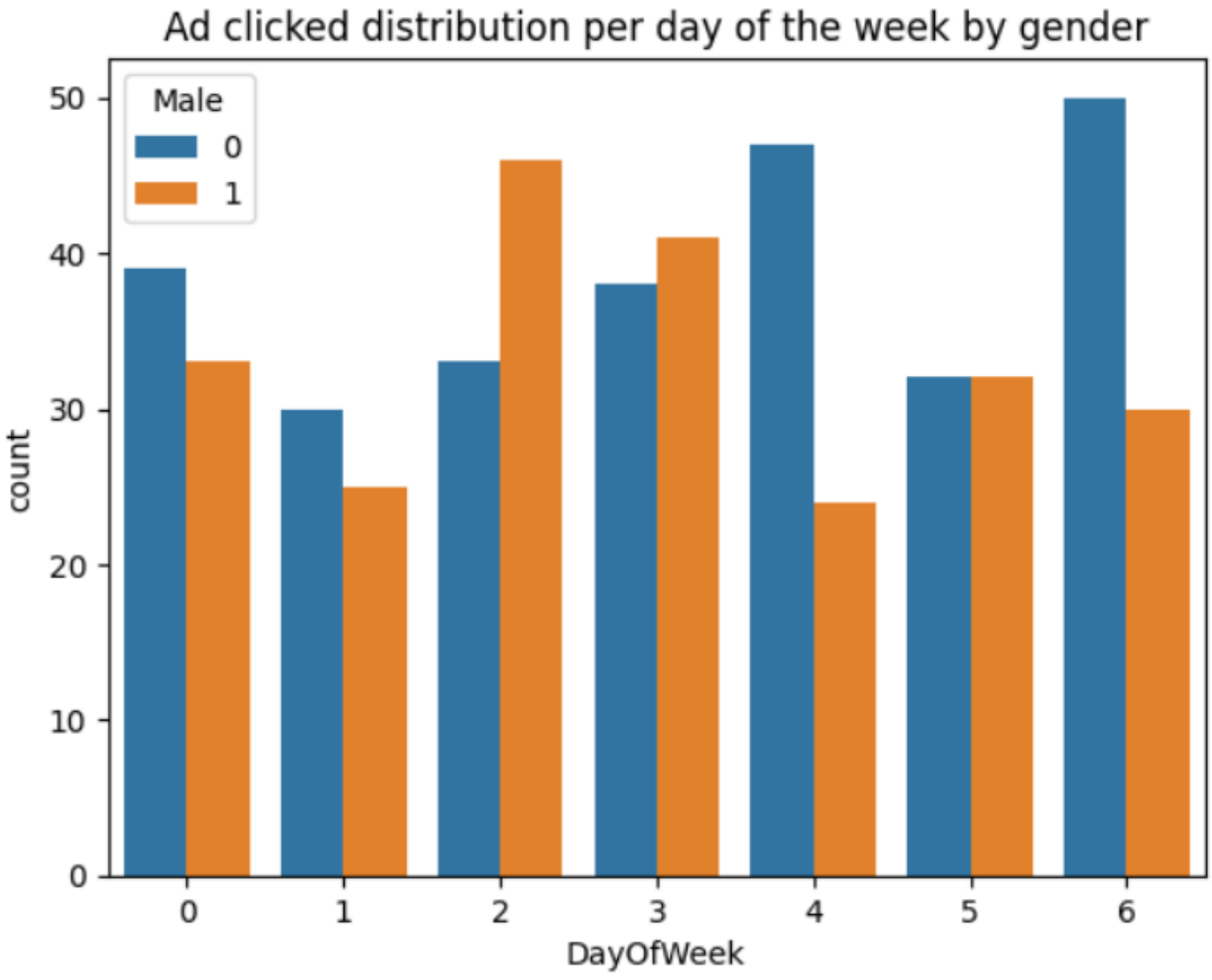
4. Exploratory Data Analysis



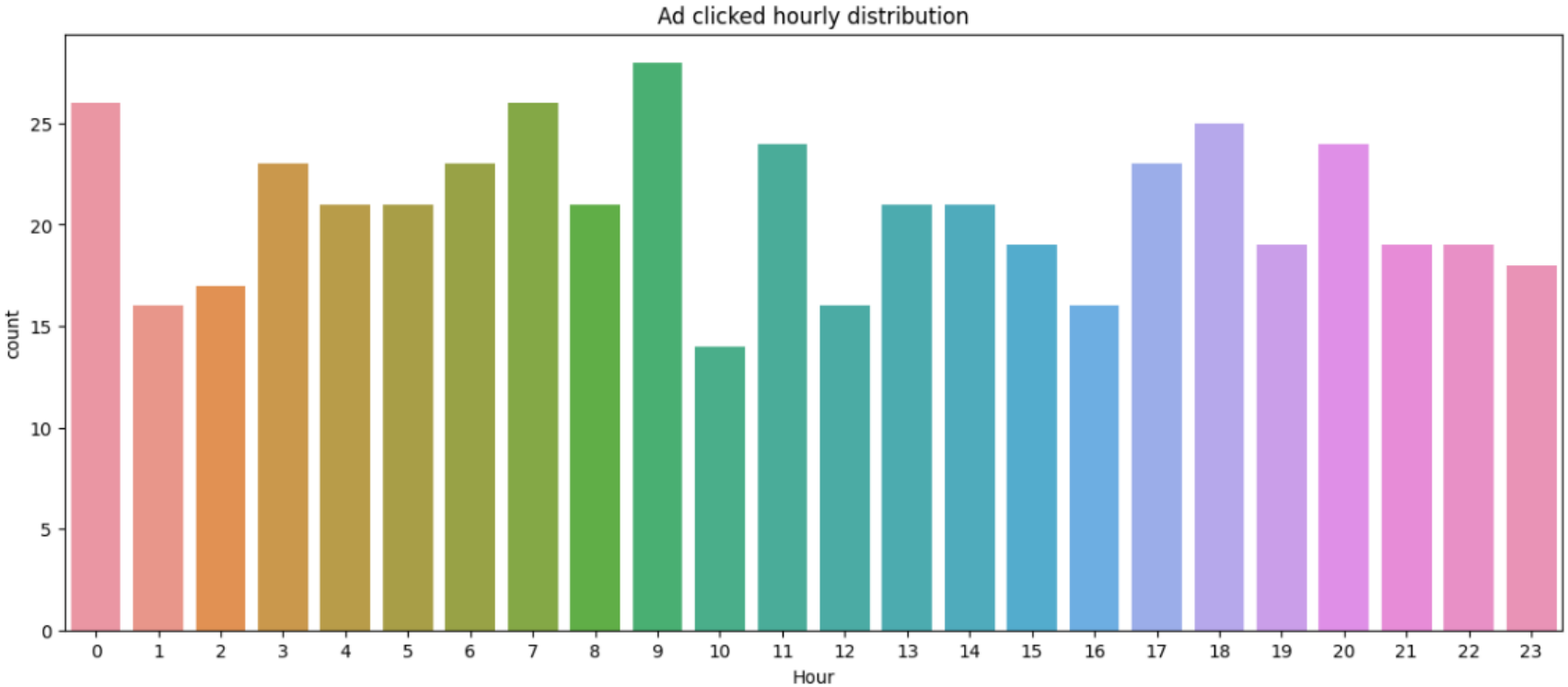
4. Exploratory Data Analysis



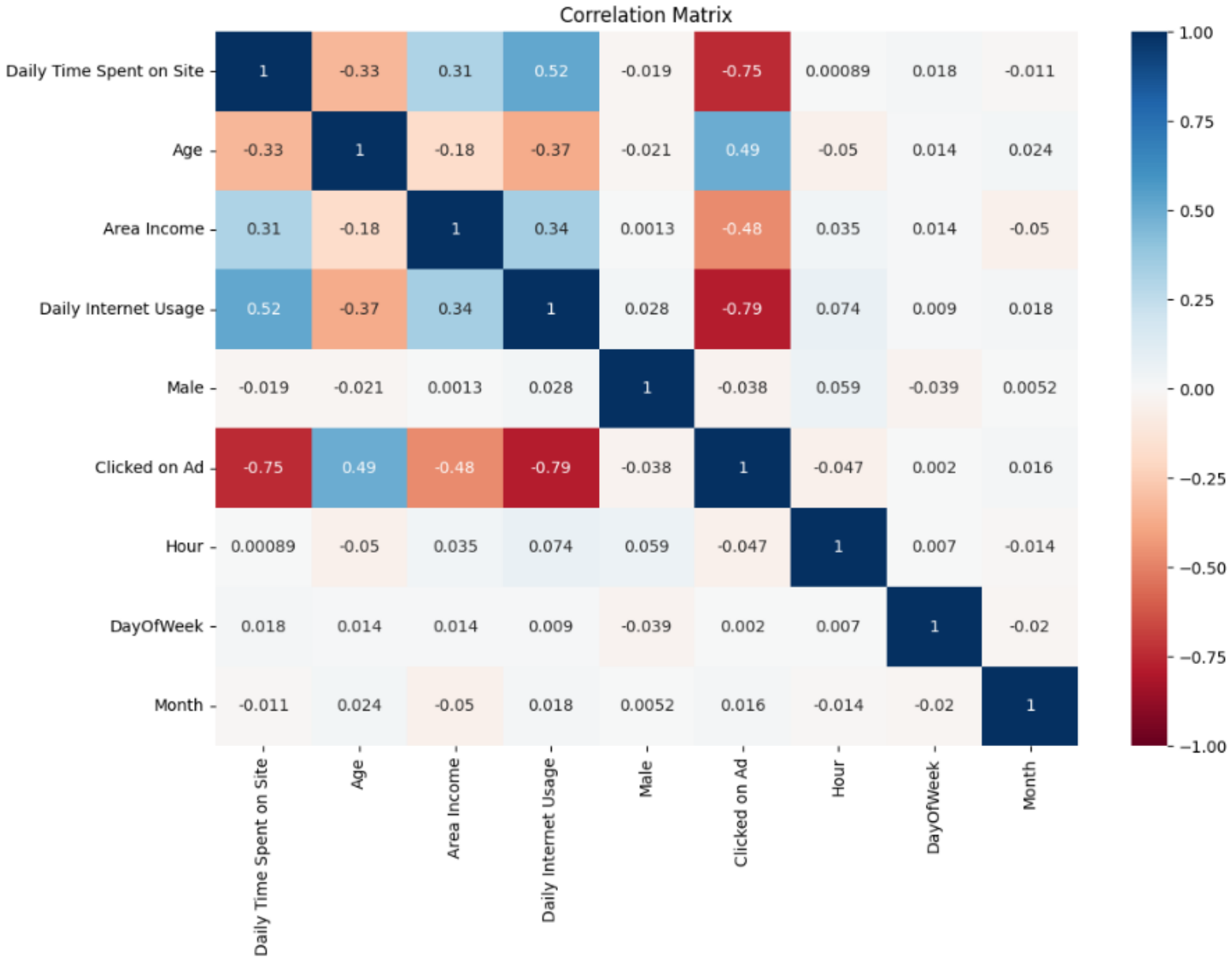
4. Exploratory Data Analysis



4. Exploratory Data Analysis



4. Exploratory Data Analysis



5. Data Preparation for Exploratory Analysis

- There are 1000 different categories (same as individuals), we can delete this variable
- Variables deleted: Timestamp, City, Country, Ad Topic Line, Hour, and Date
- Set variable y equal to 'Clicked on Ad'
- Split the dataset into training set (80%) and test set (20%)

6. Logistic Regression

First, let's do a Logistic Regression with the default parameters:

- Accuracy: 0.915
- Confusion matrix:

True/Predicted	True	False
True	99 (TN)	6 (FP)
False	11 (FN)	84 (FP)

- Classification report:

	Precision	Recall	f1-score	Support
0	0.90	0.94	0.92	105
1	0.93	0.88	0.91	95
Accuracy			0.92	200

6. Logistic Regression

Let's see if we can make it better by changing some parameters:

- Accuracy: 0.975
- Confusion matrix:

True/Predicted	True	False
True	104 (TN)	1 (FP)
False	4 (FN)	91 (FP)

- Classification report:

	Precision	Recall	f1-score	Support
0	0.96	0.99	0.98	105
1	0.99	0.96	0.97	95
Accuracy			0.97	200

6. Logistic Regression

Let's try to use GridSearch to find the best parameters for the solver liblinear:

- Accuracy: 0.975
- Confusion matrix:

True/Predicted	True	False
True	104 (TN)	1 (FP)
False	4 (FN)	91 (FP)

- Classification report:

	Precision	Recall	f1-score	Support
0	0.96	0.99	0.98	105
1	0.99	0.96	0.97	95
Accuracy			0.97	200

6. Logistic Regression

Remark: No need to do features selection with a penalty l1

- L1 penalty (Lasso regularization) performs a form of feature selection by encouraging sparsity in the coefficients
- Adds the sum of the absolute values of the coefficients as the penalty term
- Some coefficients may become exactly zero (excluding them from the model)

6. Logistic Regression

Let's look at the odds ratios:

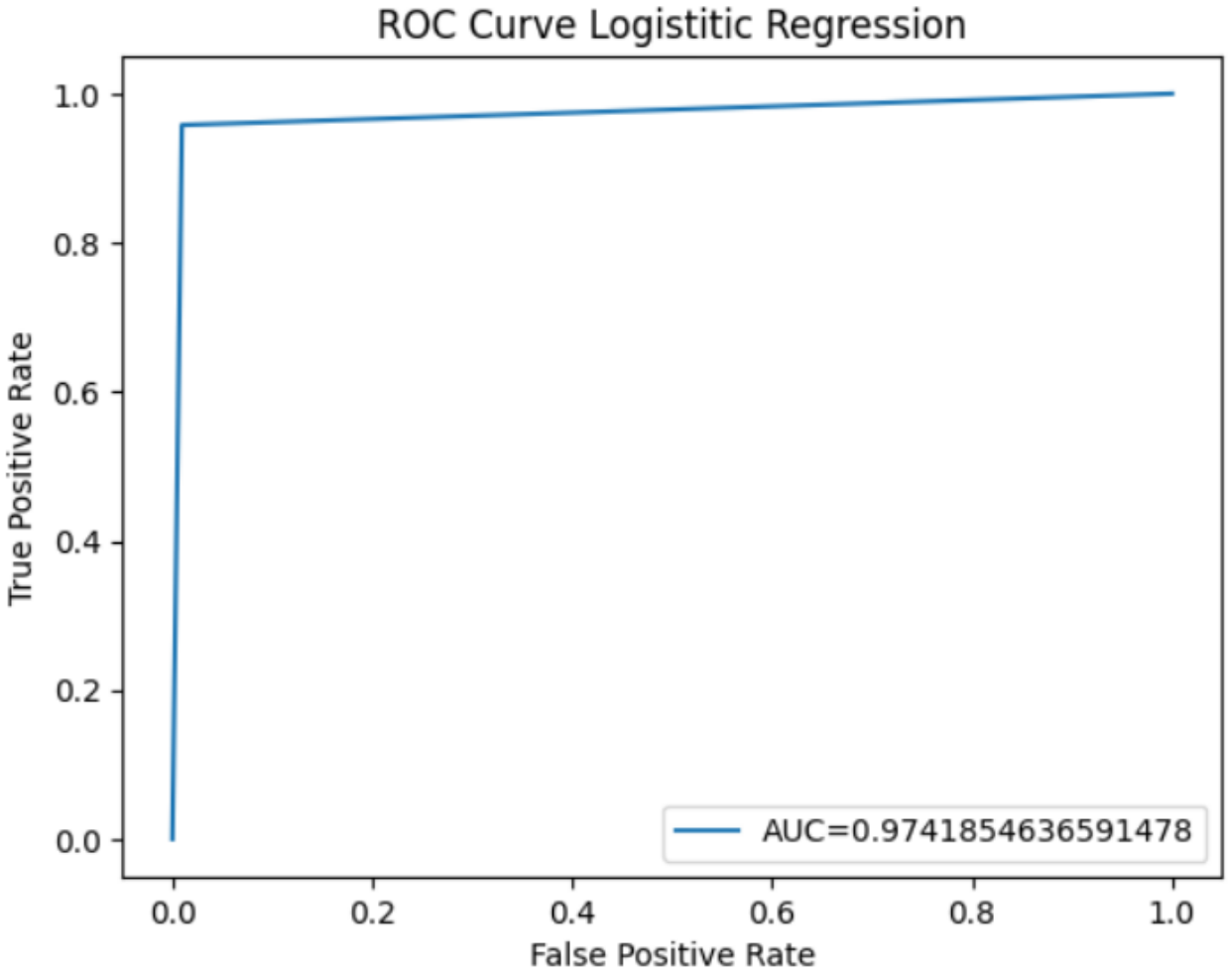
- Odds ratios:

	Coef
Day of the week	0.143817
Age	0.141713
Month	0.109447
Area Income	-0.000127
Daily Internet Usage	-0.061867
Daily Time Spent on Site	-0.180658
Male	-0.442551

- For female users, the odds of clicking are 0.44 times as larges as the odds for not clicking (when all other variables are held constant).
- As the user age increases by one, the odds for clicking are 0.14 as large as the odds for not clicking (when all other variables are held constant).

6. Logistic Regression

Let's plot the ROC Curve and AUC:



7. Random Forest

First, let's do a Random Forest with the default parameters:

- Accuracy: 0.96
- Confusion matrix:

True/Predicted	True	False
True	102 (TN)	3 (FP)
False	5 (FN)	90 (FP)

- Classification report:

	Precision	Recall	f1-score	Support
0	0.95	0.97	0.96	105
1	0.97	0.95	0.96	95
Accuracy			0.96	200

7. Random Forest

Let's try to use GridSearch to find the best parameters:

- Accuracy: 0.96
- Confusion matrix:

True/Predicted	True	False
True	104 (TN)	1 (FP)
False	4 (FN)	91 (FP)

- Classification report:

	Precision	Recall	f1-score	Support
0	0.96	0.96	0.96	105
1	0.96	0.96	0.96	95
Accuracy			0.96	200

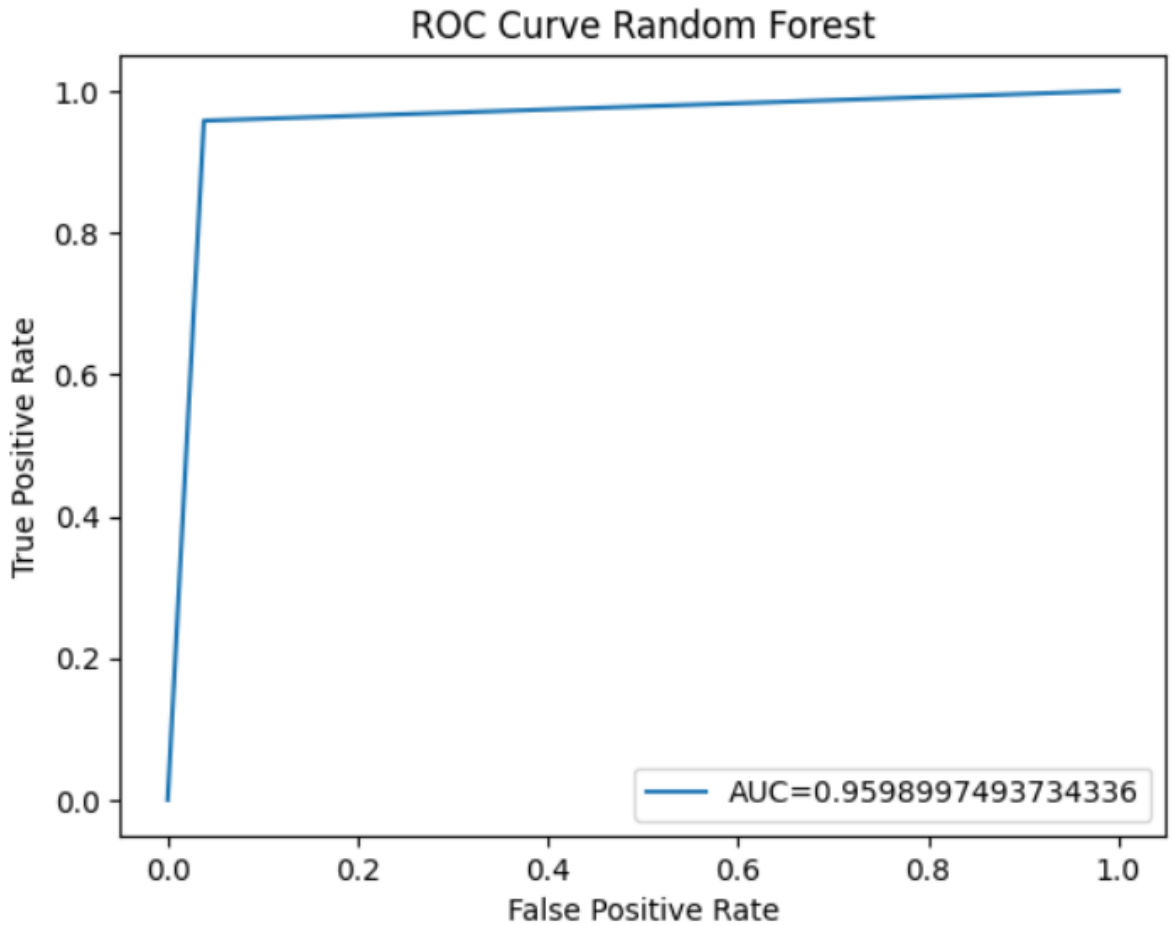
7. Random Forest

Remark: No need to do feature selection with Random Forest

- Built-in feature selection mechanisms due to the nature of the algorithm
- Calculate feature importance based on how much each feature contributes to reducing impurity (e.g., Gini impurity) across all decision trees in the forest
- Features with low importance can be considered less influential and have less impact on the final predictions

7. Random Forest

Let's plot the ROC Curve and AUC:



8. XGBoost

First, let's do a XGBoost with the default parameters:

- Accuracy: 0.95
- Confusion matrix:

True/Predicted	True	False
True	101 (TN)	4 (FP)
False	6 (FN)	89 (FP)

- Classification report:

	Precision	Recall	f1-score	Support
0	0.94	0.96	0.95	105
1	0.96	0.94	0.95	95
Accuracy			0.95	200

8. XGBoost

Let's try to use GridSearch to find the best parameters:

- Accuracy: 0.955
- Confusion matrix:

True/Predicted	True	False
True	102 (TN)	3 (FP)
False	6 (FN)	89 (FP)

- Classification report:

	Precision	Recall	f1-score	Support
0	0.94	0.97	0.96	105
1	0.97	0.94	0.95	95
Accuracy			0.95	200

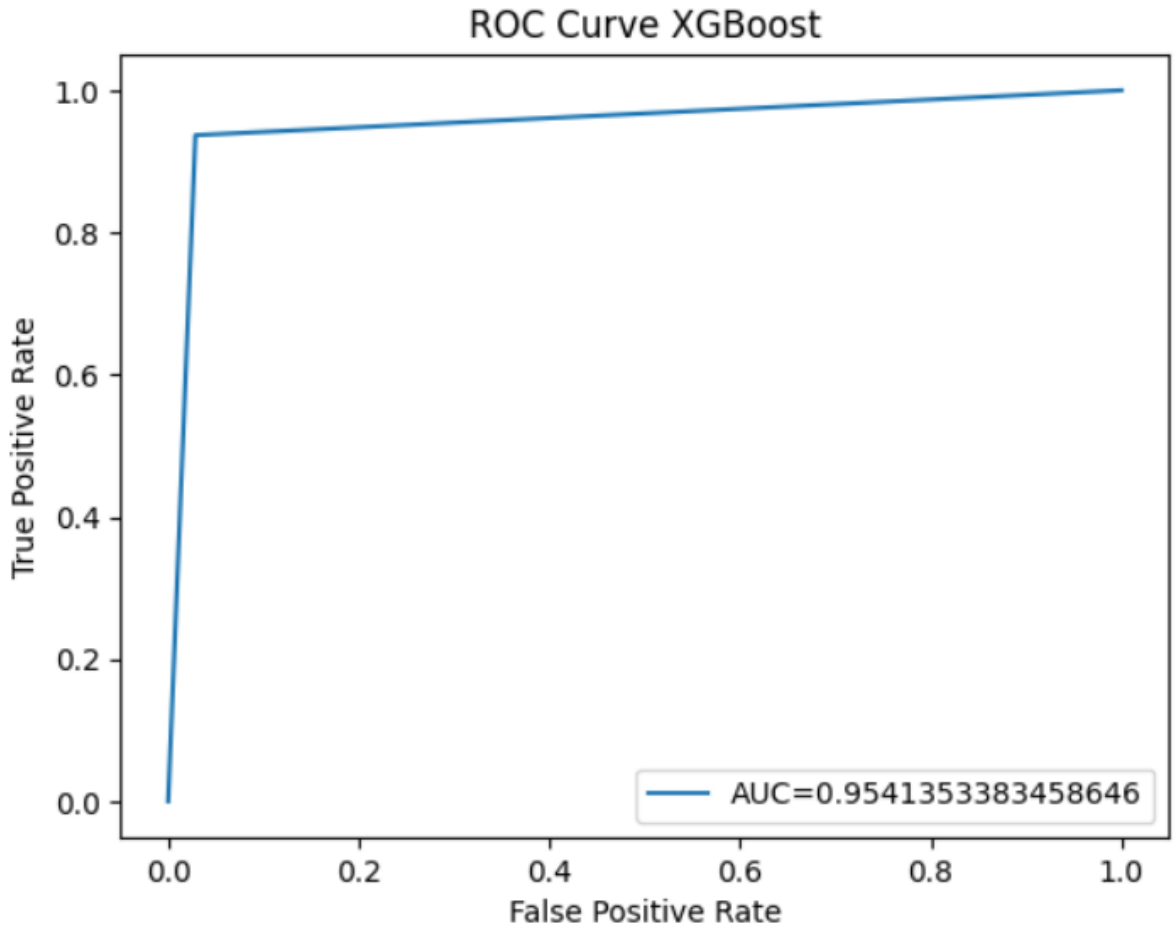
8. XGBoost

Remark: No need to do feature selection with XGBoost

- Built-in feature selection mechanism through the importance scores assigned to each feature during training
- Tree-based ensemble algorithm

8. XGBoost

Let's plot the ROC Curve and AUC:

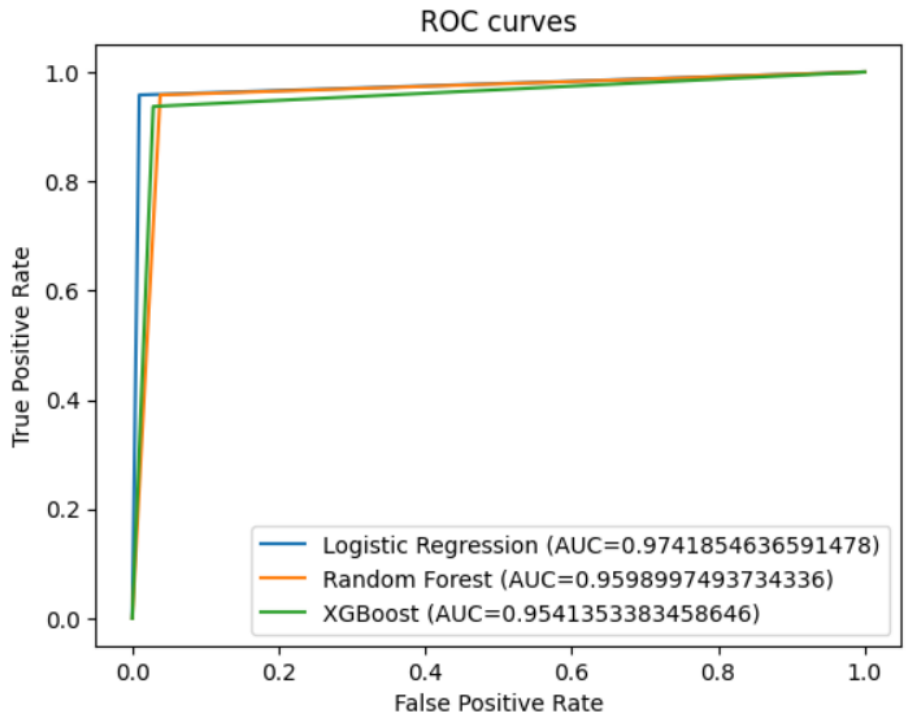


Conclusion

- Let's compare the three models using the accuracy score:

Model	Accuracy Score
Logistic Regression	0.975
Random Forest	0.96
XGBoost	0.955

- Let's compare the ROC Curves and AUC:



Bibliographie

- https://www.kaggle.com/datasets/tbyrnes/advertising
- https://scikit-learn.org/stable/index.html
- https://www.collimator.ai/reference-guides/what-is-l1-regularization