VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

Probability and Statistics

Final Project Report

Internet Advertisement Classification using Random Forest

Advisor(s):	[Instructor Name]	
Student(s):	NGUYỄN LÊ THANH HÀO	1952242
	PHẠM NHẬT KHÔI	2352623
	Nguyễn Song Đạt	1952646
	Nguyễn Phước Thịnh	1852765

HO CHI MINH CITY, AUGUST 2025

Contents

List of Figures

List of Tables

Listings

1 Introduction

The proliferation of the internet has led to a massive increase in online advertising. While advertisements can be useful, they can also be intrusive and detract from the user experience. Therefore, the ability to automatically distinguish between advertising and non-advertising content is a significant challenge in modern data science. This project addresses this challenge by developing a model to classify internet images as either advertisements ('ad') or non-advertisements ('nonad').

This report details the process of analyzing the "Internet Advertisements Data Set" sourced from the UCI Machine Learning Repository[4]. We will explore the data, clean it for analysis, and apply several machine learning models to build an effective classifier. The primary goal is to construct a robust model that can accurately predict whether an image is an advertisement based on its various features.

2 Data Description

2.1 Dataset Overview

The dataset used in this project is the "Internet Advertisements Data Set" from the UCI Machine Learning Repository. This dataset contains information about internet images and their classification as advertisements or non-advertisements.

The original dataset consists of 3,279 observations with 1,560 columns. The first column serves as an index and is removed during preprocessing, leaving 1,559 features for analysis. The target variable, located in the last column (X1558), indicates whether an image is an advertisement ('ad.') or not ('nonad.').

2.2 Data Loading

The dataset is loaded from a CSV file using the following R code:

```r
# Load the dataset from the CSV file
data <- read.csv("../add.csv", header = TRUE,
   stringsAsFactors = FALSE)

# Identify the target column name
target_col <- names(data)[ncol(data)]
```

2.3 Target Variable Distribution

The target variable shows a significant class imbalance:

- Advertisement images (ad): 459 observations (14%)
- Non-advertisement images (nonad): 2,820 observations (86%)

This imbalance is typical in advertisement detection problems and will need to be considered when building and evaluating classification models.

2.4 Missing Values

The dataset contains missing values represented as "?" strings. Initial analysis revealed:

- Total missing values: 15 occurrences

- All missing values are concentrated in column 5
- Missing values represent approximately 0.46% of column 5's data

2.5 Data Preprocessing

Several preprocessing steps were applied to prepare the data for analysis:

```r
# 1. Remove the first column which is an unnecessary index
data <- data[, -1]

# 2. Handle missing values represented by "?"
# Convert "?" to NA for all feature columns
for(i in 1:(ncol(data)-1)) {
  data[[i]] <- as.numeric(ifelse(data[[i]] == "?", NA, data[[
     i]]))
}

# 3. Impute missing values using the median of each column
for(i in 1:(ncol(data)-1)) {
  if(any(is.na(data[[i]]))) {
    median_val <- median(data[[i]], na.rm = TRUE)
    data[[i]][is.na(data[[i]])] <- median_val
  }
}

# 4. Clean and format the target variable
data[[target_col]] <- factor(data[[target_col]],
                              levels = c("ad.", "nonad."),
                              labels = c("ad", "nonad"))
```

The preprocessing pipeline includes:

1. **Index removal**: The first column containing row indices was removed
2. **Missing value handling**: All "?" strings were converted to NA values
3. **Median imputation**: Missing values were replaced with the median of their respective columns

4. **Target variable formatting**: The target variable was converted to a factor with clear labels ("ad" and "nonad")

2.6 Final Dataset Characteristics

After preprocessing, the cleaned dataset has the following properties:

- Dimensions: 3,279 rows $\times$ 1,559 columns
- All missing values have been imputed
- Target variable is properly formatted as a factor
- All feature columns are numeric
- Class distribution remains: 459 advertisements, 2,820 non-advertisements

2.7 Basic Statistics

Preliminary statistics for the first five feature columns show:

- Column 1: Mean = 1,639, Median = 1,639, SD = 946.71
- Column 2: Mean = 64.02, Median = 51, SD = 54.87
- Column 3: Mean = 155.34, Median = 110, SD = 130.03
- Column 4: Mean = 3.91, Median = 2.1, SD = 6.04
- Column 5: Mean = 0.77, Median = 1, SD = 0.42

These statistics indicate varying scales across features, suggesting that normalization or standardization may be beneficial for certain machine learning algorithms.

3 Descriptive Statistics

This section presents a comprehensive exploratory data analysis (EDA) of the Internet Advertisements dataset to understand the data distribution, relationships between variables, and key characteristics that will inform our modeling approach.

3.1 Target Variable Analysis

The target variable distribution reveals a significant class imbalance in the dataset:

- Advertisement images (ad): 459 observations (14%)
- Non-advertisement images (nonad): 2,820 observations (86%)

This 1:6 ratio between advertisement and non-advertisement classes is typical in real-world advertisement detection scenarios and must be considered when evaluating model performance.

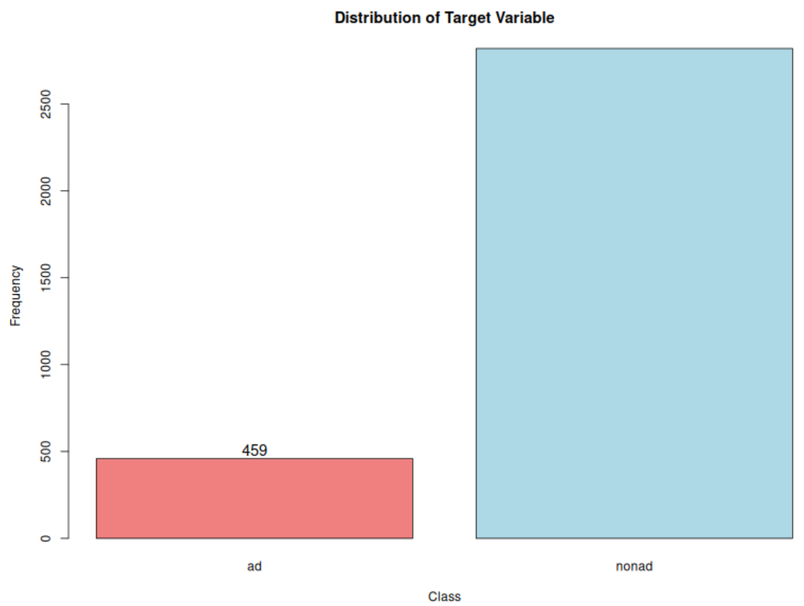

Figure 3.1: Distribution of target variable showing class imbalance

The target distribution visualization was generated using:

```
# Bar plot for target variable distribution
target_table <- table(data[[target_col]])
barplot(target_table,
        main = "Distribution of Target Variable",
        xlab = "Class", ylab = "Frequency",
        col = c("lightcoral", "lightblue"),
        border = "black")
```

3.2 Feature Distribution Analysis

To understand the characteristics of individual features, we analyzed the distribution of the first 10 features in the dataset. The summary statistics reveal varying scales and distributions across features:

Table 3.1: Summary statistics for the first 10 features

Feature	Min	Q1	Median	Mean	Q3	Max	SD
X0	1.00	32.50	51.00	60.44	61.00	640	47.06
X1	1.00	90.00	110.00	142.89	144.00	640	112.56
X2	0.00	1.28	2.10	3.41	3.90	60	5.20
X3	0.00	1.00	1.00	0.77	1.00	1	0.42
X4	0.00	0.00	0.00	0.004	0.00	1	0.065

Histograms for the first six features show diverse distribution patterns:

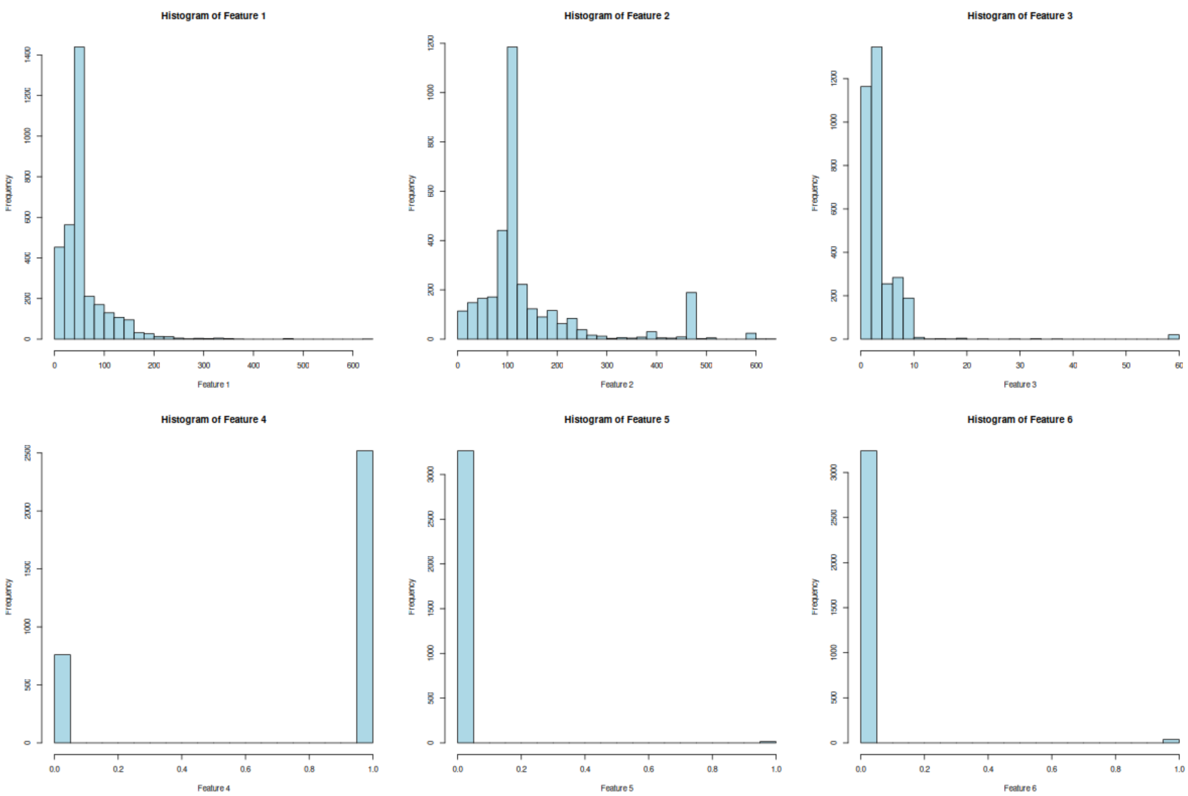

Figure 3.2: Histograms of the first six features showing distribution patterns

The histogram generation code:

```
# Histograms for the first 6 features
par(mfrow = c(2, 3))
for(i in 1:6) {
  hist(data[[i]],
       main = paste("Histogram of Feature", i),
       xlab = paste("Feature", i),
       ylab = "Frequency",
       col = "lightblue",
       border = "black",
       breaks = 30)
}
```

3.3 Class-wise Feature Analysis

To understand how features differ between advertisement and non-advertisement classes, we created boxplots comparing the distribution of key features across classes:

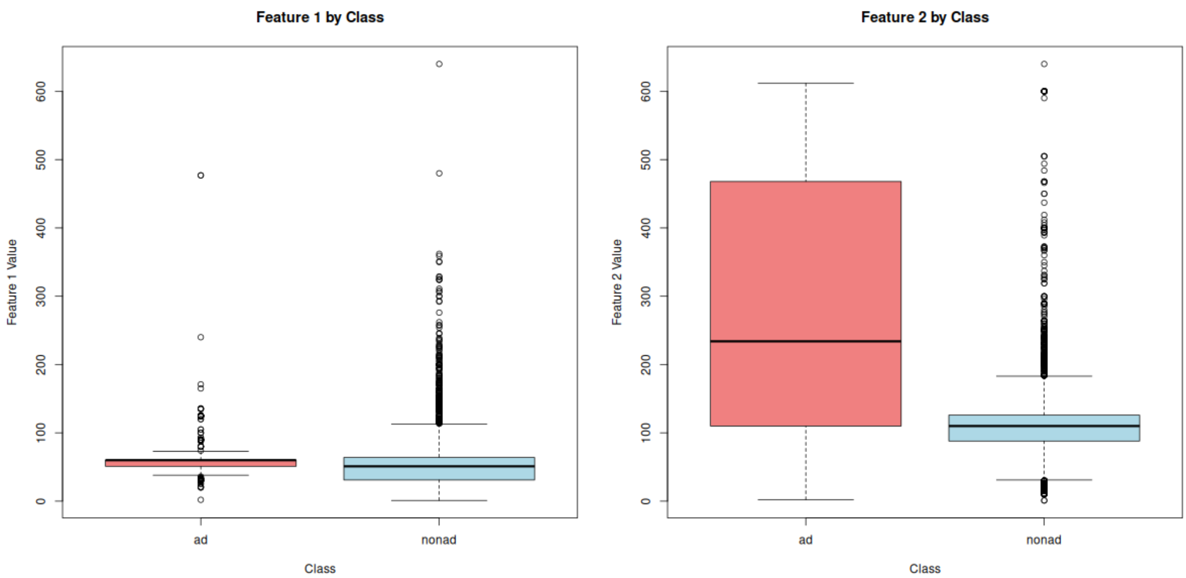

Figure 3.3: Boxplots of features 1 and 2 grouped by target class

The boxplot analysis was implemented as:

```
# Boxplots of features 1 and 2, grouped by target class
```

```
par(mfrow = c(1, 2))
boxplot(data[[1]] ~ data[[target_col]],
        main = "Feature 1 by Class",
        xlab = "Class", ylab = "Feature 1 Value",
        col = c("lightcoral", "lightblue"))
boxplot(data[[2]] ~ data[[target_col]],
        main = "Feature 2 by Class",
        xlab = "Class", ylab = "Feature 2 Value",
        col = c("lightcoral", "lightblue"))
```

3.4 Feature Relationships and Scatter Analysis

Scatter plots and density plots provide insights into feature relationships and class separability:

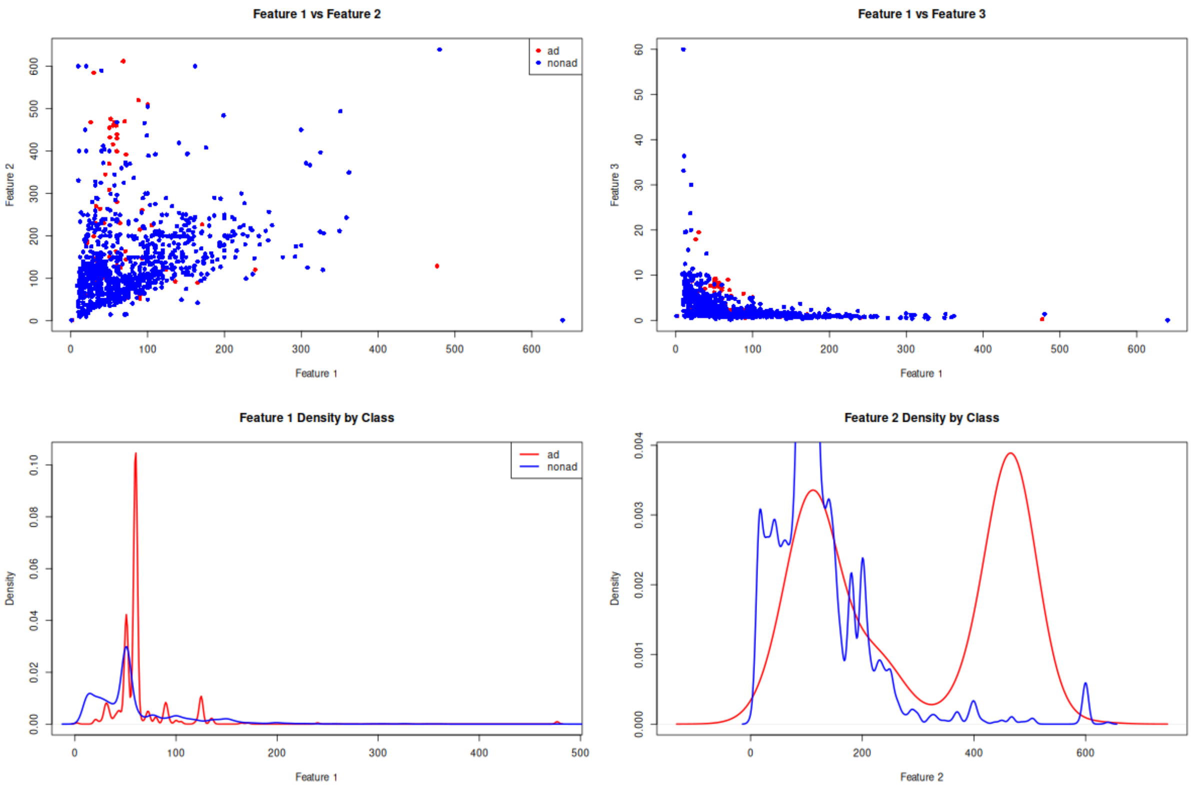

Figure 3.4: Scatter plots and density plots showing feature relationships and class distributions

The visualization combines scatter plots and density analysis:

```r
# Scatter plots and density plots
par(mfrow = c(2, 2))
colors <- c("red", "blue")
class_colors <- colors[as.numeric(data[[target_col]])]
plot(data[[1]], data[[2]],
     main = "Feature 1 vs Feature 2",
     xlab = "Feature 1", ylab = "Feature 2",
     col = class_colors, pch = 16)
legend("topright", legend = levels(data[[target_col]]),
       col = colors, pch = 16)
```

3.5 Correlation Analysis

Correlation analysis reveals important relationships between features. The correlation heatmap for the first 10 features shows the strength of linear relationships:

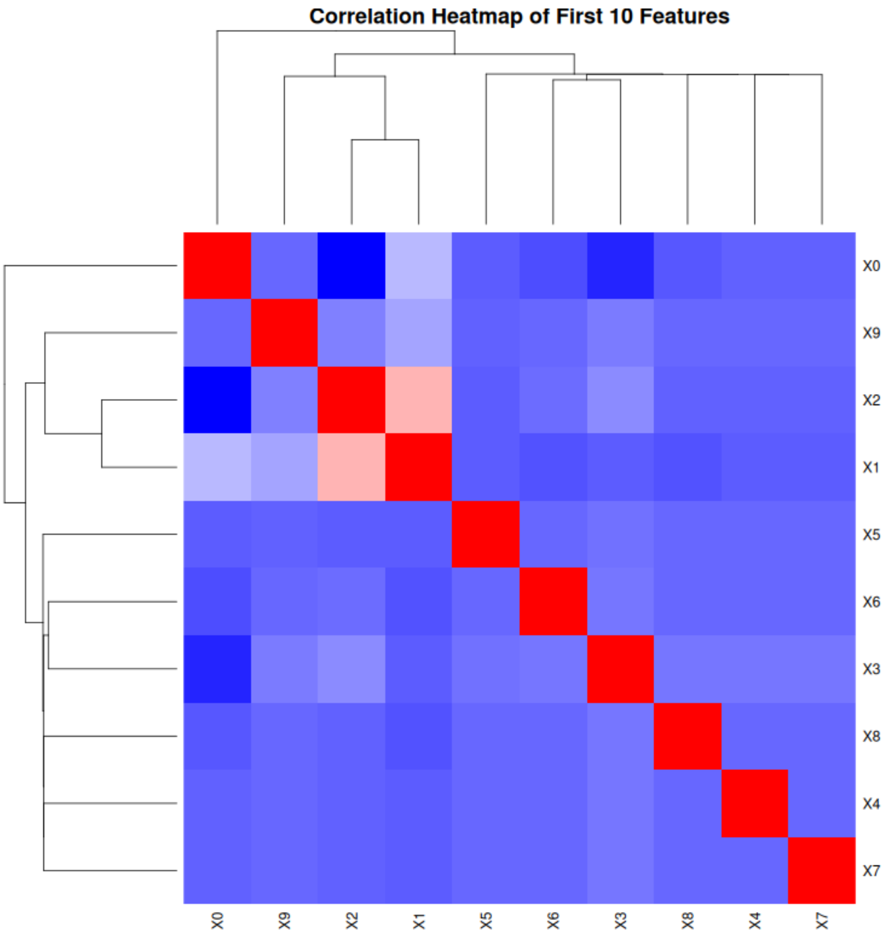

Figure 3.5: Correlation heatmap of the first 10 features

The correlation analysis identified several perfect correlations (correlation = 1.0) between different feature pairs, indicating potential redundancy in the dataset:

- Features X11 and X14: Perfect positive correlation
- Features X8 and X15: Perfect positive correlation
- Features X13 and X38: Perfect positive correlation
- Features X44 and X46: Perfect positive correlation

The correlation matrix was computed using:

```r
# Correlation heatmap for the first 10 numeric features
numeric_data <- data[, sapply(data, is.numeric)]
cor_matrix <- cor(numeric_data[, 1:10], use = "complete.obs")
```

```
heatmap(cor_matrix,
        symm = TRUE,
        main = "Correlation Heatmap of First 10 Features",
        col = colorRampPalette(c("blue", "white", "red"))
           (100))
```

3.6 Key Findings from Descriptive Analysis

The exploratory data analysis reveals several important characteristics:

1. **Class Imbalance**: The dataset has a significant imbalance with 86% non-advertisements
2. **Feature Diversity**: Features show diverse scales and distributions, suggesting the need for normalization
3. **Perfect Correlations**: Multiple feature pairs show perfect correlation, indicating potential redundancy
4. **Class Separability**: Some features show different distributions between advertisement and non-advertisement classes
5. **High Dimensionality**: With 1,559 features, dimensionality reduction techniques may be beneficial

These findings will inform our modeling approach, particularly regarding feature selection, data preprocessing, and evaluation metrics that account for class imbalance.

References

[1] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media, 2009.

[4] M. Lichman. UCI machine learning repository, 2013.

[5] J Ross Quinlan. *C4. 5: programs for machine learning.* Morgan kaufmann, 1993.