Online Shopper's intention

Filip Mordarski

10/1/2019

Contents

Introduction

This report is part of the capstone project of the EdX course 'HarvardX: PH125.9x Data Science: Capstone'. Participants had to create their own project. Firstly, they need to seek out a dataset. I chose Online Shopper's Intention database from Kaggle. URL: Link

Executive Summary

The main goal of this project is building algorithms to predict if given online shopper will be generated revenue or not. It can be really useful information for entrepreneur. Online Shopper's Intention dataset contains around 12 thousands sessions of online shoppers. The information from the analysis of this database is used to generate revenue predictions that are compared with actual zero one variable revenue to check the quality of the forecasting algorithm.

Overview

This report is split in four sections. First, **Introduction** describes content of data set, summarizes the goal of the project and key steps that were performed. **Analysis** section explains the process and techniques used, such as data cleaning, data exploration and visualization and my modeling approach. In a **Results** section I present the modeling results and discuss the model performance. **Conclusion** section gives us a brief summary of the report, its limitations and future work.

Data set description

The dataset consists of feature vectors belonging to 12,330 sessions. The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. The dataset consists of 10 numerical and 8 categorical attributes.

- Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration

These variables represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories.

- Bounce Rate

Feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session.

- Exit Rate

Feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session.

- Page Value

Feature represents the average value for a web page that a user visited before completing an e-commerce transaction.

- Special Day

Feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother's Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina's day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.

- Month

Attribute represents which month was during the visit.

- Operating System, Browser, Region, Traffic type

Attributes represent what operating system (8 different operating systems) and browser (13 different browsers) online shoppers were using. Region (9 different regions) and traffic type (20 different types) indicate region of the user and traffic type during shopping. We can not say anything more about these attributes, because they are numeric values. We can only determine their relationships between them and the impact on the dependent variable.

- Visitor type

Feature indicates visitors as returning or new visitor.

- Weekend

Boolean value indicating whether the date of the visit is weekend, and month of the year.

- Revenue

Attribute indicates if visitor will be generated revenue or not.

Analysis

This section expains the process and techniques used. It shows us some data visualization or data exploration. This section explains also modelling approach.

Data cleaning

Firstly I would like to remove all rows with any NA values. The code below.

```r
# removing any rows with NA value
data <- data[complete.cases(data), ]
```

Now I am prepared to create train and testset for my algorithm.

```r
y <- data$Revenue

# Create trainset and test set
set.seed(1)
test_index <- createDataPartition(y, times = 1, p = 0.1, list = FALSE)
test_set <- data[test_index, ]
train_set <- data[-test_index, ]
```

Data exploration and visualization

Now we can begin the eye-pleasing part of the report. The below section will be full of plots and tables.

Table 1: What percentage of visitors generate revenue?

FreqPercTrue	FreqPercFalse
0.15	0.85

Above table shows us what percentage of visitors generate revenue or not.