Classification of Online Shoppers Purchasing Intention

Sahil Mathur   
Faculty of Engineering and Computing *Coventry University*Coventry, UK  
[mathurs2@uni.coventry.ac.uk](mailto:mathurs2@uni.coventry.ac.uk)

Sukhmani Saini  
Faculty of Engineering and Computing *Coventry University*Coventry, UK  
[sainis9@uni.coventry.ac.uk](mailto:sainis9@uni.coventry.ac.uk)

**Abstract— The primary objective of this paper is to analyse the intention/behaviour of online shoppers using machine learning techniques on the given dataset. Utilising Classification methods including K-nearest Neighbour, Naïve Bayes, Decision Trees and Random Forest , a comparison is drawn from the results and a conclusion is formulated on the most suitable method. Python programming language was used to implement techniques and algorithms utilizing machine learning repositories, prediction results and algorithm performance measures were obtained, and visualized for comparison and discussion.**

Keywords— machine learning; Classification; python; categorical data; decision tree; k-nearest neighbour; random forest, k- nearest.

Key: ML = Machine Learning, KNN = k-Nearest Neighbour, Random Forest = RF, PLT = Matplotlib

# Introduction

With the boom in internet access across the world, the normal retail shopping has been rapidly shifting to E-commerce or online shopping and as a result, the shopping dynamics have been continuously changing and has already become a major part of the retail market. A typical customer prefers shopping online in their own comfort as compared to travelling to a physical shop and buying the products. Resulting in Businesses to start implementing various online marketing techniques to lure customers to their shopping portal. Customers who visit these web portals might not make any purchases at all. This could be for a variety of reasons, like high product pricing or window shopping. There have been multiple studies (Santini 2018) that imply that retention strategies such as an appropriate recommendation system play a critical role in converting sales. For example, if the ML solution predicts a strong customer purchase intention, the recommend system may suggest a higher quality or more expensive product because it can be inferred that the user is ready to consider a better or more expensive product if their intent to buy a particular item is very strong. If the solution forecasts a lower intention to buy, the recommendation system may suggest discounted products or products with special deals. Cambridge Analytica used similar strategies, although on a larger scale, to influence voter decisions in American elections (The Guardian, 2018). This demonstrates the strength of a proper machine learning technique if applied effectively could vary a majority population’s intent.

# Objective

This project intends to make advantage of information that customers may leave behind in the form of browsing history data or user information when they browse an ecommerce platform. Utilizing clickstream and activity data information, the study seeks to forecast online buyers' purchasing intentions using this information. The research intends to build a machine learning model based on this data to forecast client purchasing intentions. The project's goal is to create a Machine Learning model that can anticipate customer purchase intent as accurate as possible.

# Dataset

The dataset used in the project was obtained from UCI machine learning. The dataset has been provided by Authors Sakar and Y. Kastro (Sakar et al., 2018).

The data has 12330 instances, each with 18 characteristics and no missing data. Each instance represents a single person's visit to the site. When a repeat consumer visits the site several times, only the initial visit from the time frame examined is taken into account, with subsequent visits being excluded. The first 17 attributes are visitor-related features that include 10 numerical values and 7 categorical elements. The last 18th attribute that is ‘Revenue’ reflects the category if the visit resulted in revenue or not, which means that if the visitor made a purchase, it is labelled as True or False.

The categorical and numerical features used in the prediction model are shown in Tables 1 and 2, respectively.

Table 1. Categorical Features

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Description** | **Number of values** |
| OperatingSystems | Visitor’s Operating System | 8 |
| Browser | Visitor’s Browser type | 19 |
| Region | Geographical Region of the visitor | 9 |
| TrafficType | Source of redirect by visitor | 20 |
| VisitorType | Type of Visitor (“New”, “Returning”, “Other”) | 3 |
| Weekend | Indicating if the day of visit is a weekend or not by True/False | 2 |
| Month | Month of visit | 10 |
| Revenue | Boolean to indicate if revenue was generated | 2 |

Table 2. Numerical Values

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Description** | **Number of values** |
| Administrative | Number of pages visited by the visitor | 12330 |
| Administrative\_Duration | Total amount of time (in seconds) spent by the visitor | 12330 |
| Informational | Number of pages visited by the visitor about Web site, communication and address information of the shopping site | 12330 |
| Informational\_Duration | Total amount of time (in seconds) spent by the visitor on informational pages | 12330 |
| ProductRelated | Number of pages visited by visitor about product related pages | 12330 |
| ProductRelated\_Duration | Total amount of time (in seconds) spent by the visitor on product related pages | 12330 |
| BounceRates | Average bounce rate value of the pages visited by the visitor | 12330 |
| ExitRates | Average exit rate value of the pages visited by the visitor | 12330 |
| PageValues | Average page value of the pages visited by the visitor | 12330 |
| SpecialDay | - | 12330 |

# Data Cleaning

The dataset is available in a good condition with no null values and some minor issues which we decided not to handle beforehand.

# Exploratory Data Analysis (EDA)

The dataset has both numerical and categorical variables. The "BounceRates," "ExitRates," and "PageValues" features, among others, depict the metrics for every session. Our data analysis yielded the following major findings:

The sample is imbalanced, with only 15% of visits resulting in a purchase.

PageValues may be the most critical feature in predicting purchase conversion.

Many numerical features are skewed to the right.

Data Imbalance –

Table

Description automatically generatedThis dataset is skewed, with only 15% of visits resulting in a purchase. While 15% is not a severely imbalanced dataset, we can still investigate strategies or algorithms which are more effective at dealing with it.

PageValue Importance –

PageValues is described as the average page value of the user's pages visited. Values are typically allocated to essential pages in an e-commerce setting, such as checkout pages or pages following the checkouts.

As illustrated in Fig 2, a PageValues greater than 5 enhances the likelihood of sales conversion. As a result, the PageValues feature gives an excellent indication of whether or not the visitor will buy something.

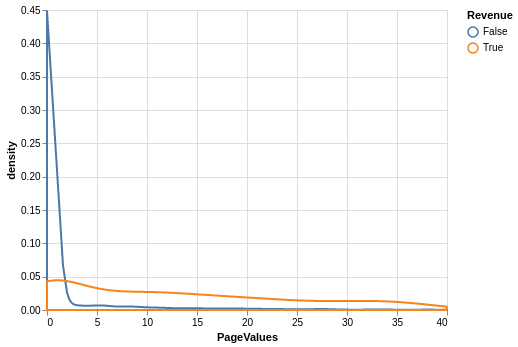


Figure 2. PageValues Density Plot

A picture containing chart

Description automatically generated

Figure 3. Heatmap of the dataset

From Fig 3, We can depict that Administrative and Administrative\_Duration are correlated. ExitRates, Information , ProductRelated and BounceRates show similar properties. Page Value appears to have a stronger relationship with Revenue.

Right-Skewness of the data –

Finally, we discovered that most of numerical features have right-skewed ends. This is typical in e-commerce environments, where some individuals have a disproportionately high utilisation rate. We can test whether removing outliers or using feature transformations like Box-Cox can improve model performance.

Chart, histogram

Description automatically generated

Figure 4. Right Skewness of the data

# Model Selection

References

The Guardian. (2018, May 7). Cambridge Analytica: How did it turn clicks into votes? the Guardian. <https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie>

Santini, R. (2018). Recommender systems as “tastemakers”: Collaborative filtering as a market strategy for online cultural products. Observatorio (OBS\*). <https://doi.org/10.15847/obsobs1222018847>

Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2018). Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Computing and Applications, 31(10), 6893–6908. https://doi.org/10.1007/s00521-018-3523-0

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**