CMPE 251 Project Report

Online Shopping Website Analysis

Bryan Hoang

2021-11-14

Table Of Contents

List of Figures

List of Tables

1 Introduction

Online shopping is prominent form of e-commerce in the world today that has a large market for consumers, as evident by the popularity of online shopping platforms such as Amazon. As such, there have been efforts to identify different types of online shoppers to improve an online shopping website's sales, or revenue [1].

A similar effort will occur given a dataset describing the properties and behaviours of online shopper visiting an unnamed online shopping site, that will be referred to as Nozama throughout the rest of the report. The goals of this report are to create an accurate predictor of which customer will buy something on Nozama and to determine which properties of users are associated with eventually buying items.

The analysis will first involve analyzing the datasets' properties, such as distribution and correlation, to help remove possibly redundant attributes and provide metrics for evaluating prediction models. Then the attributes of the dataset will be ranked and filtered based on importance and redundancy to improve clustering analysis and prediction model building. Once the dataset has been assessed wholistically, an analysis of clusters in the data to find insights regarding properties of eventual buyers and to help with prediction model building will occur. After that, a predictor will be constructed based on the findings in the data and assessed before finally making recommendations for Nozama's website.

2 Properties of the Dataset

The dataset has no missing attribute values, which avoids the challenge of having to fill in or compensate for any missing values. In lieu of this, practically all numerical attributes in the dataset are unimodally distributed with a positive skew in the range of 1 to 7.5. For instance, a histogram of the Product Related attribute values seen in Figure 2.1 shows the positively skewed distribution of the attribute.

Figure 2.1: Histogram of the Product Related attribute values

The significance of the distribution increases for the target attribute we want to predict. The Revenue attribute has two classes, TRUE and FALSE, representing whether a visitor eventually bought an item from Nozama's website. Of the 12 330 records in the dataset, 84.5 % of the values for Revenue are FALSE, which implies that any predictor could easily attain a prediction accuracy of 84.5 % if it always predicted Revenue to be false. Therefore, the evaluation of predictors should factor this in as a minimum prediction accuracy to achieve.

Given that both the prediction attributes and the target attribute have similar distributions, the correlation between the attributes were observed next. From Figure 2.2, two relatively significant correlations for Exit Rates and Page Values can be observed in relation to revenue. One should expect to see that Exit Rates and Page Values play in important role in predicting the target attributes, which will be explored further in later steps of the analysis.

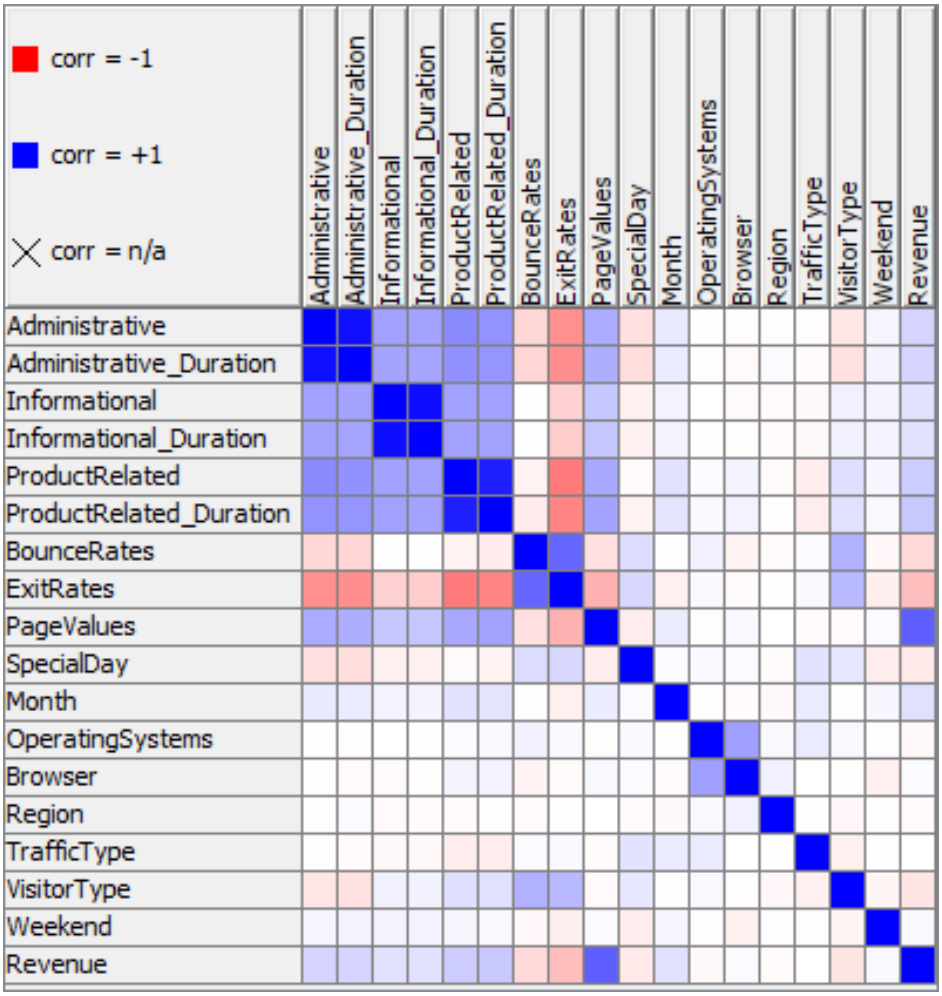

Figure 2.2: Rank correlation matrix of the attributes

As far as why the data is distributed this way, one could reason that a majority of visitors to Nozama's website tend to not interact much with it and don't make any purchases, leading to the high number of low values for each numerical attribute.

3 Attribute Selection and Ranking

To improve clustering results and prediction accuracy, a tree ensemble leaner was used to filter out unimportant attributes. The math formula node and sorter node helped determine which attributes could be removed based on their importance in level 0 of the decision trees in model. The procedure used involved running the workflow, removing attributes with 0 importance (level 0), and repeating the steps until all attributes had non-zero importance. The results of the attribute selection and ranking can be summarized in Table 3.1 and Table 3.2 respectively.

Table 3.1: Selection of attributes based on importance (level 0).

Important Attributes	Unimportant Attributes
Administrative	Browser
Administrative Duration	Informational
Bounce Rates	Informational Duration
Exit Rates	Month
Page Values	Operating Systems
Product Related	Region
Product Related Duration	Special Day
	Traffic Type
	Visitor Type
	Weekend

The important attributes from Table 3.1 will be the only ones used during cluster analysis and prediction model building.

Qualitatively, "Informational" type pages don't really help with sales. Neither do the attributes relating to the time of visiting nor "how" someone is visiting the site. It seems like the main reasons leading to a purchase lies in the content of certain pages and the metrics for measuring a visitor's journey through Nozama's website.

Table 3.2: Ranking of attributes based on importance (level 0).

Attributes	Importance (level 0)	Rank
Page Values	1.00	1
Exit Rates	0.85	2
Product Related Duration	0.61	3
Product Related	0.47	4
Bounce Rates	0.30	5
Administrative	0.24	6
Administrative Duration	0.11	7

From Table 3.2, the two most important attributes seem to be “Page Values” and “Exit Rates”, which agrees with the observations from chapter 2.

4 Clustering

4.1 k-Means

- # clusters: 2-5?
- Conclusions there!

4.2 EM

- # clusters: 2-5?
- 1 cluster captures who buys, try to target other clusters

4.3 SVD (Singular Vector Decomposition)

- use matlab script to make it christmas and comment on clumpiness
- Bounce rate and Exit rate are super correlated.

5 Predictors

5.1 Random Forests

- 500 trees good number
- Compare filtered vs unfiltered attributes

5.2 Support Vector Machines (SVMs)

- rule out svm due to low accuracy
- Consider baysian since it's suppose to be used with all attributes when you don't know which ones are important

5.3 k-NN

- Consider based on clustering in previous section

6 Actionable Conclusions

- Draw conclusions on attributes associated with eventual buyers, and suggest improvements to the website

References

[1] A. J. Rohm and V. Swaminathan, "A typology of online shoppers based on shopping motivations", *Journal of Business Research*, Marketing on the web - behavioral, strategy and practices and public policy, vol. 57, no. 7, pp. 748–757, Jul. 1, 2004. [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S014829630200351X Last accessed: Nov. 15, 2021.