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

- no missing values
- Mention dataset used (header/noheader, numeric vs categorical)

2.1 Percentage of Revenue classes

- High percentage of Revenue classes are false, which will affect the evaluation of certain predictors

2.2 Distribution

- A lot of then numerical attributes take on a monomodal distribution (right skewed)
- Given that a lot of the revenues are false, the peaks in the data make indicate correlation with the revenue 84%

2.3 Correlation

- Rank/linear correlation (numeric data)
- Comment on ratio/percentage of true/false for the revenue's class.

3 Attribute Selection and Ranking

- Filter out attributes as done in class
- 8/9
- Rank based on important from math node

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.