CMPE 251 Project Report

Online Shopping Website Analysis

Bryan Hoang

2021-11-24

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 [2], 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 fig. 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 fig. 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.

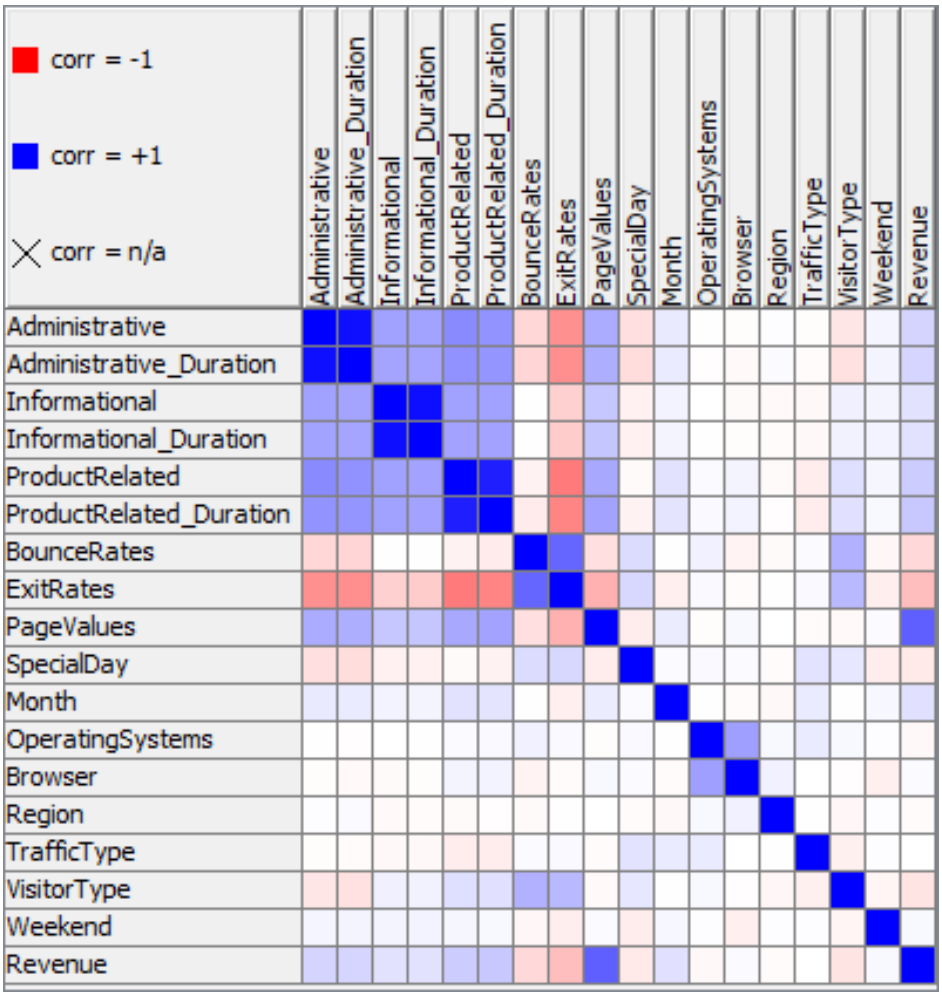


Figure 2.2: Rank correlation matrix of the attributes.

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 (fig. 3.1), removing attributes with 0 importance (level 0), and repeating the steps until all attributes had non-zero importance.

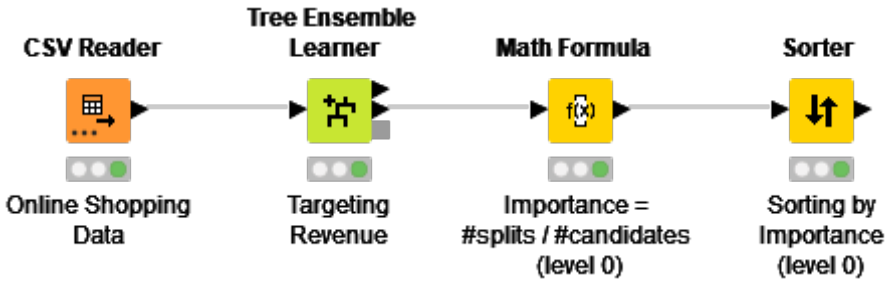


Figure 3.1: KNIME workflow used for attribute selection and ranking.

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 to be described in the following chapters.

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

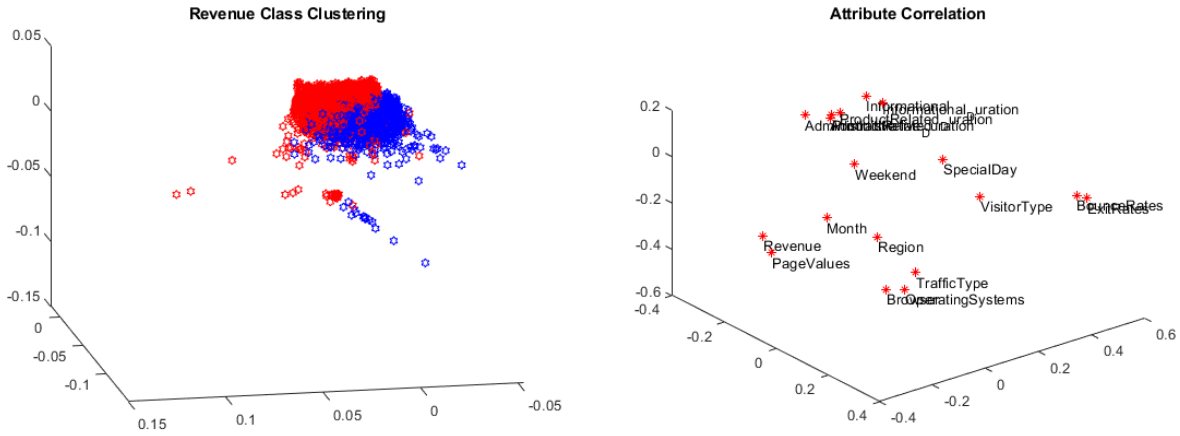
From table 3.2, the two most important attributes seem to be "Page Values" and "Exit Rates", which agrees with the observations from chapter 2.

4 Clustering

Both k-means and Singular Value Decomposition (SVD) methods of clustering were used to determine more properties from the data, using only the important attributes discussed in chapter 3.

4.1 Singular Value Decomposition

SVD clustering was attempted first to help visualize potential clusters and to further analyze the correlation between different attributes. From fig. 4.1a, there is an evident separation between data points in either of the Revenue's classes. So clustering using k-means should be feasible given the observed clumpiness, but may require multiple clusters due to how close the groups are. The Figure also shows promise for an accurate prediction model given the grouping of the two classes.



(a) SVD clustering of data points coloured based on Revenue class.
Blue = TRUE, Red = FALSE.

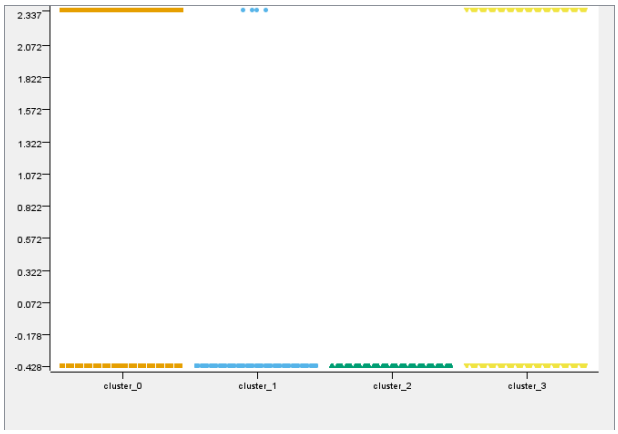
(b) SVD of the attributes.

Figure 4.1: SVD figures generated from a modified version of the `showsvd.m` MATLAB script.

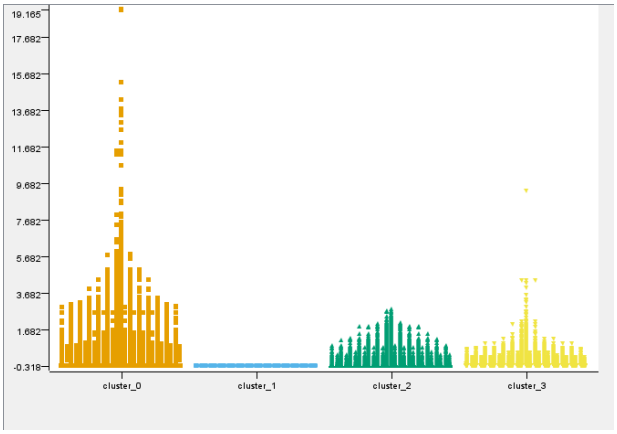
In addition, fig. 4.1b shows that the Revenue and Page Values attributes are highly correlated, which reinforces the correlation analysis conducted in chapter 2 and the attribute ranking done in chapter 3. There is also some strong correlation between the important attributes located in different "correlation" clusters. The more notable of those clusters is the Bounce Rate and Exit Rate attributes being isolated and correlated similarly to the Revenue and Page Values attributes. This has some implications for the website's page structure, which will be further discussed in chapter 6.

4.2 k-Means

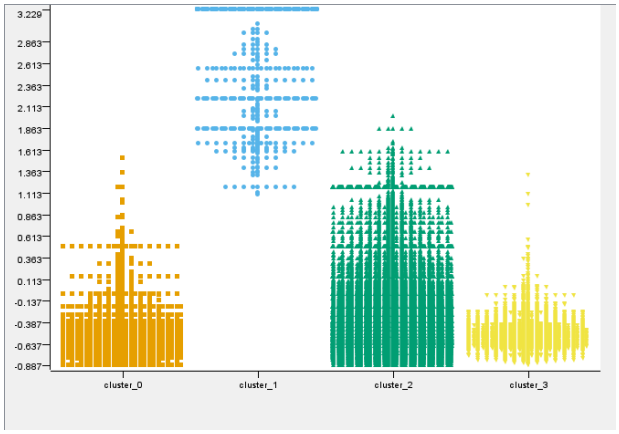
After trying various values of k in the range of $[2, 7]$, the number of clusters that provided the most useful insight was 4. From fig. 4.2, we can see that the clusters 0 and 3 reveal that revenue generating visitors do tend to have high Page values and low Exit Rates associated. The reverse trend can be observed in clusters 1 and 2, which reinforces the correlation analysis conducted in chapter 2.



(a) Clusters vs. Revenue



(b) Clusters vs. Page Values



(c) Clusters vs. Exit Rates

Figure 4.2: Scatter plots of the k-means clusters against significant attributes.

One may want to target clusters 1 and 2 to determine what would make them purchase something from Nozama's website, which will be expanded on in chapter 6.

5 Predictors

With knowledge of the dataset's characteristics, different types of predictors were built to help determine whether a visitor will generate Revenue by making a purchase at Nozama's website. Three predictors, namely Random Forest (RF), Support Vector Machine (SVM), and k-nearest neighbours (k-NN).

5.1 Random Forests

A RF predictor was built first, as seen in fig. 5.1, since its reliability gives a good baseline to evaluate other predictors on.

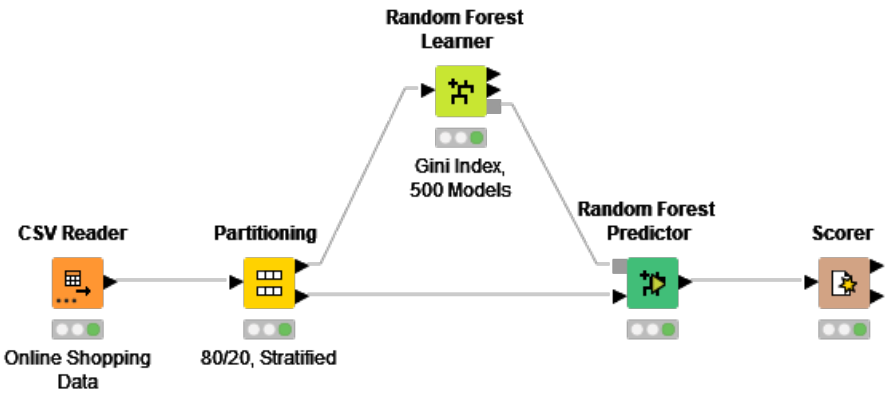


Figure 5.1: KNIME workflow used to build the RF predictor.

Revenue \ Prediction (Revenue)	FALSE	TRUE
FALSE	1996	88
TRUE	160	222

Correct classified: 2,218 Wrong classified: 248

Accuracy: 89.943 % Error: 10.057 %

Cohen's kappa (κ) 0.584

Figure 5.2: Confusion matrix of the RF predictor built.

The prediction accuracy of 89.9 % seen from the confusion matrix shown in fig. 5.2 is greater than the minimum prediction accuracy of 84.5 % determined in chapter 2. Then the RF predictor does provide value as an improvement over naively guessing that no visitor will generate Revenue.

5.2 Support Vector Machines

An SVM predictor was built next, as seen in fig. 5.3, since the prediction attribute has two classes. But using any of the 3 kernel functions (polynomial, hypertangent, radial basis), results in a predictor that's less accurate than the RF predictor and sometimes doesn't meet the 84.5 % threshold of naively guessing. This may be due to the lack of separation between the classes seen in fig. 4.1a, resulting in the difficulty of a kernel function to map the data points into a space where there is sufficient separation between the classes.

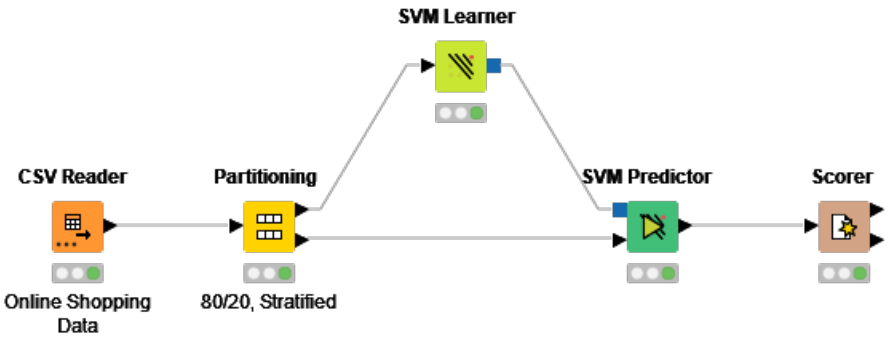


Figure 5.3: KNIME workflow used to build the SVM predictor.

5.3 k-nearest neighbours

k-NN was the last prediction method to consider, as built in fig. 5.4, since cluster results from chapter 4 indicated that it might be feasible to for a k-NN to have a higher prediction accuracy than the RF predictor.

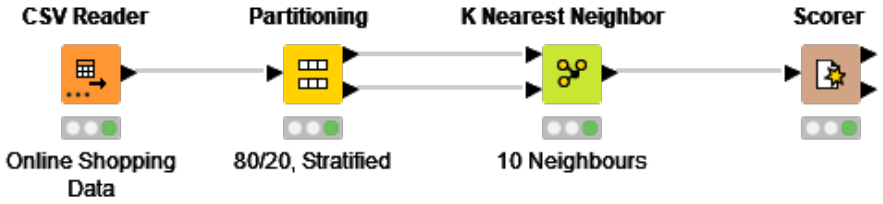


Figure 5.4: KNIME workflow used to build the k-NN predictor.

Revenue \ Class [kNN]	FALSE	TRUE
FALSE	2043	41
TRUE	279	103

Correct classified: 2,146 Wrong classified: 320

Accuracy: 87.024 % Error: 12.976 %

Cohen's kappa (κ) 0.335

Figure 5.5: Confusion matrix of the k-NN predictor built.

The prediction accuracy of 87.0 % observed in the confusion matrix shown in fig. 5.2 is better than naively guessing all FALSE, but doesn't improve in the RF predictor. In retrospect, the non-existent separation between the blue and red data points in fig. 4.1a imply that that the k-NN predictor would have a difficult time dealing with data points close tho the boundary between the two revenue classes.

6 Actionable Conclusions

6.1 Predicting Who Will Make a Purchase

Due to the nature of the dataset, a RF predictor built with is the best model to use to determine if a visitor will make a purchase. By using Gini Index for the split criterion in conjunction with 500 models, its prediction accuracy of 89.9 % is an improvement over the prediction accuracy of 84.5 % from naively guessing that every visitor will not make a purchase. Nozama may want to consider using the RF predictor to help them test improvements to their website during testing.

6.2 Properties of Visitors Who Make a Purchase

Analysis of the data reveals that Nozama can capitalize on their Page Value, Exit Rate, and Bounce Rate metrics to increase the number of visitors who will make a purchase on their website.

For instance, Nozama should restructure their website so that pages with high Page Values should be easier to navigate to. e.g., feature the pages prominently in navigation elements. Since pages with high Page Values are visited by potential buyers, making sure they're exposed to as many people as possible will increase the likelihood of a visitor making a purchase.

Pages with high Exit Rates lead to make visitors not purchase anything. So Nozama should investigate why visitors are leaving that page, and may reconsider the design, layout, and content of similar pages.

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, 1st Jul. 2004, ISSN: 0148-2963. DOI: 10.1016/S0148-2963(02)00351-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S014829630200351X. [Accessed 15/11/2021].

[2] C. O. Sakar, S. O. Polat, M. Katircioglu and Y. Kastro, 'Real-time prediction of online shoppers purchasing intention using multilayer perceptron and LSTM recurrent neural networks,' *Neural Computing and Applications*, vol. 31, no. 10, pp. 6893–6908, October 2019, ISSN: 0941-0643, 1433-3058. DOI: 10.1007/s00521-018-3523-0. [Online]. Available: http://link.springer.com/10.1007/s00521-018-3523-0. [Accessed 15/11/2021].