1 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 1.1 and Table 1.2 respectively.

Table 1.1: 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 1.1 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 1.2: 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 1.2, the two most important attributes seem to be “Page Values” and “Exit Rates”, which agrees with the observations from **??**.