STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. 11, March 2023, pp 519–530.
Published online in International Academic Press (www.IAPress.org)

Selecting key features of online behaviour on South African informative websites prior to unsupervised machine learning

Judah Soobramoney [1,*], Retius Chifurira [1], Temesgen Zewotir [1]

[1] *School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa*

Abstract The main aim of the study was to explore the feature selection process of online web data prior to unsupervised machine learning models. At the time of writing, no such literature could be found reporting the use of feature selection in this context. Feature selection was determined by inspecting the variability and association between features. The variability of numeric features were quantified using the variance, mean absolute difference and dispersion ratio metrics whilst the coefficient of unalikeability was employed for categorical features. To quantify association, correlation matrices were used for numeric features, chi-squared independence tests between categorical features and box-and-whisker plots between mixed features. The main findings showed the variance, mean absolute difference, dispersion ratio and coefficient of unalikeability metrics have successfully highlighted features with very low variability within the observed data. Whilst the correlation matrix, chi-squared test for independence and box-and-whisker plots highlighted possible redundancy, natural relationships and insightful relationships between the features thereby suggesting features to be considered for omission prior to unsupervised modelling. The proposed methods and findings can be applied to various other applications of feature selection and exploration.

Keywords Feature Selection, Google Analytics Tracking, Online Behaviour, Unsupervised Machine Learning

AMS 2010 subject classifications 62H20, 62H30, 62J10, 62P99

DOI: 10.19139/soic-2310-5070-1139

1. Introduction

1.1. Context and framework

The world wide web is soon becoming a market common place. Upon which, customers may browse a wide repertoire of products available globally [15]. With the rise of such technology, corporates are pulled into a more competitive environment [16]. As a result, the need for deep analytics of online user behaviour emerges. Such analytics enable corporates to further apply targeted marketing strategies to optimize online market share [18]. There are tools available that allow tracking and storage of a vast range of information on each user and their corresponding activity on a particular website. The data tracked includes detailed information on the activity online (such as point of entry, browse path, duration, clicks, etc.), historic activity (such as previous times visited, and previous engagement each time), geographic information and device specific information (type of device, operating system, web browser, etc.). Whilst such online tracking tools make such vast data available, this study provides methods to identify the key features that explain online behaviour from an unsupervised perspective.

*Correspondence to: Judah Soobramoney (Email: judahsoobramoney@gmail.com). School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

1.2. The case study

The case study was taken from a corporate called TEKmation who hosts an informative website. An informative website posts information regarding the company, location, history, its products, etc – however, with no online purchase functionality. Website traffic often entail numerous users from several different devices performing a wide variety of tasks and of varying engagement levels. Thus, attempting to summarize web traffic activity is considered a highly complex task. A simple but important, yet very difficult question to answer is: "What are people doing on my website?". For instance, some users spend a few seconds on the website whilst others perhaps hours; some users visit the website once whilst others several times; users have several different entry points onto the website and follow unique page paths [12]. Therefore, due to this complexity, data scientists often employ unsupervised machine learning techniques (clustering algorithms) to assist in further aggregating the data based on patterns within the data. Prior to unsupervised machine learning , the process of feature selection is described as an important yet challenging problem [17].

There are several researchers that have described methods of feature selection for unsupervised machine learning techniques, however, this study focuses specifically on methods of feature selection prior to the unsupervised modelling on web traffic data from a South African informative website. The literature discussed provides a few examples of previous research on unsupervised machine learning feature selection (UFS).

The discussed literature details the methods and applications of feature selection conducted by a few previous researchers. However, none provide an in-depth exploration of feature selection on such potentially big, robust and detailed data on web traffic (that could be found thus far). This research aims to provide a generalized methodology that could be used by any corporate hosting a similar website to determine the key features that construct online user behaviour groups. The web tracking tool employed within this study is Google Analytics Tracking. Whilst there are several studies that have employed clustering algorithms to achieve significant behaviour groups of online users, this study illustrates an in-depth method on how the under-pinning features can be identified. The article further discussed the methodology in Section 2, the data in Section 3 and the empirical results in Section 4.

2. Methodology

In this section, the background theory on methods that can be utilized to select the key features for unsupervised machine learning models, to attain accurate online user behaviour groups are discussed.

2.1. Related work

In an application on microarray data, Wang and Zhu [17] proposed an unsupervised feature selection (UFS) technique that separates data points into clusters, and based on cluster contribution, features are selected. The framework of the study that Wang and Zhu [17] conducted focused on penalized model-based clustering. Wang and Zhu [17] have found that the proposed methods have efficiently removed the non-informative features. Fraiman et al. [7] proposed UFS procedures targeted at identifying "noisy" non-informative features and multicollinearity between features that are appropriate to the forward-backward clustering algorithm employed. The methods were based on a "two variable selection" process and "conditional means". Fraiman et al. [7] found that the proposed methods did not work well for high-dimensional data. Fop and Murphy [6] classify UFS for model-based clustering into four broad groups: Bayesian, penalization and model selection approaches which have been applied to mortality data. Fop and Murphy [6] concluded that feature independence is crucial and the aim is to discard both redundant and uninformative features. Maugis et al. [11] proposed an UFS process that classifies each feature as a relevant clustering feature, an irrelevant clustering feature dependent on a part of a relevant feature or an irrelevant clustering feature totally independent of all relevant features. The selection technique of identifiability and consistency proved to be established [11]. Maugis et al. [11] assessed random waveform data to illustrate the proposed feature selection process. Chormunge and Jena [2] proposed the use of correlation assessment to aid in UFS of high dimension data. Chormunge and Jena [2] discuss first feature elimination through k-means clustering and thereafter identification of non-redundant features through correlation measures from each cluster. Chormunge and Jena [2] state the experiment results, using microarray data, yielded accuracy and efficacy using the proposed

method. Guerif [8] proposed UFS through a combination of multiple rankings. The experiment data showed that the approach yields effective and stable results [8].

2.2. Initial feature selection

The success of supervised and unsupervised machine learning models depend highly on the features used for data modelling. It has been proven, that the set of features chosen can improve or reduce the performance of statistical models [4]. Feature selection would also determine the computational costs and run-time associated with training models. Furthermore, feature importance scores are often used to interpret models and thus including the appropriate features is necessary [4]. The initial set of features considered to be included within the model depend on the available features, the model's intuition, and research on similar models and applications. The features considered could be sourced directly from the data or inferred (indirectly) obtained using the available information [5].

2.3. Features relevance

During the feature selection process, the variability of the features within the consideration set need to be assessed. Features with no variability ought to be removed whilst those with little variability need to be further analysed. For instance, suppose a feature (say, age) has only a single value (age = 32) across all observations within the dataset (in an extreme case). This would imply that the feature "age" is non-discriminant enough to be included within the model. Features with low variability can be included if the observations that differ could potentially provide insight to the unsupervised learning model. However, data scientist may choose to omit features with relatively low variance levels (although risking a potential loss of information to the model) in the attempt to optimize runtime in certain applications of machine learning.

2.3.1. Variance

The variance provides a measurement of how far spread observations are from the mean. For a random variable X, Equation 1 formulates the variance, where μ_i is the expectation (E) of X_i:

$$Var(X) = E[(X_i - \mu_i)^2]. \tag{1}$$

Whilst the variance metric is relative to the unit of measure of a particular feature, the higher the variance metric, the more spread observations are within the feature. Features with variance = 0 indicate that the feature observations are all identical. The coefficient of variation (CoV) of a random variable X is computed as the square root of the variance (standard deviation denoted by σ_X) divided by the mean (μ_X) of a feature (Equation 2)

$$CoV(X) = \frac{\sigma_X}{\mu_X}. \tag{2}$$

When the coefficient of variation is less than 1, the feature is said to have very low variability between the observations with the feature [1]. However, this rule of thumb (coefficient of variance less than 1) has shown to be unreliable for very small mean values. This is driven by the calculation of the metric, with the denominator being the mean value, thus the closer the mean value is to 0, the larger the coefficient of variance metric will be. In this paper, integer valued features with mean values less than 0.2 and a coefficient of variance greater than 2.5 is also be used to identify features that had little variance about an approximate zero mean.

2.3.2. Mean absolute difference

The mean absolute difference is a measure of statistical dispersion within a numeric feature. The mean absolute difference of a variable X_i, is computed as per Equation 3

$$MAD(X_i) = \frac{1}{n}\sum_{j=1}^{n} |X_{ij} - \bar{X}_i|, \tag{3}$$

where j represents each observation within X_i. The mean absolute difference provides an indication of the spread of the observations from the mean. The larger the mean absolute difference, the greater the variability within a feature. A mean absolute difference of zero, implies that all observations within the feature are identical. Whilst there are no supporting literature to define a cut-off point to identify features with very low variability, this paper identifies features that have a mean absolute difference within 5% of the mean to be a low variability feature.

2.3.3. Dispersion ratio

The dispersion ratio of a variable X_i represents the ratio between the arithmetic and the geometric mean of the variable as per Equation 4:

$$Dispersion\ Ratio(X_i) = \frac{\mu_i}{GM_i}, \tag{4}$$

where $\mu_i = \bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$ and $GM_i = (\prod_{j=1}^{n} X_{ij})^{\frac{1}{n}} = e^{[\frac{1}{n}\sum_{i=1}^{n} \ln a_i]}$.

The closer a dispersion ratio is to 1, the lower the dispersion between the observations within a feature. In theory, the arithmetic mean would be larger than the geometric mean. However, the geometric mean is impacted by zero or missing values. Hence, the geometric mean (in cases of zero values) is often computed by excluding the 0 or missing data points which would, in-turn, alter the expected relationship between the arithmetic and geometric means [3]. Presently, there is no literature to advise on a dispersion ratio value that could be used as a benchmark to identify low variability features. Thus, this paper isolated features with a dispersion ratio between 0.8 and 1.2 as potential low variability features where zero data points were omitted from the geometric mean computation. However, depending on the scale of the variables and distribution of the data, an appropriate low variability interval would need to be determined in other applications.

2.3.4. Coefficient of unalikeability

The coefficient of unalikeability (u) computes how frequently observations differ from each other within variable X_i. It is often used as a pseudo measure of variance for categorical features of n observations by comparing one observation x_i with another observation within the same variable x_j for $i, j \epsilon n$ and $i \neq j$ (Equation 5).

$$u(X_i) = \frac{\sum_{i \neq j} c(x_i, x_j)}{n^2 - n}, \tag{5}$$

where:

$$c(x_i, x_j) = \begin{cases} 1, & x_i \neq x_j \\ 0, & x_i \neq x_j \end{cases}.$$

The coefficient of unalikeability computes a measure between 0 and 1 where the measures closer to 1 indicate the data within X_i are more unalike [10]. There is no supporting literature to advise on the point at which the coefficient of unalikeability indicates very low variability. As a result, this paper employed a rule of thumb that features with a coefficient of unalikeability less than 0.2 should be considered as possible low variability features.

2.4. Association between features

Within the feature selection process, it is important to inspect the association between features. Whilst measures with high or significant association should be included, this process will highlight potential information redundancy and features with potential natural relationships [5]. Thus if features share a high association, further exploratory analytics is required to decipher if the association indicates redundancy or insight. If features are highly associated indicating redundancy then one of the two features should be omitted from the unsupervised machine learning models. Similar to variability, data scientists are often required to produce models that are light-weight in terms of run-time. Thus, during the feature selection process, associated variables may also be omitted (however potential loss of valuable insight needs to be inspected first). It is also important to note, association may or may not be driven by a causal relationship.

2.4.1. Correlation matrix

A correlation matrix can be used to establish the association between numeric features with each other. The correlation $\rho_{(X,Y)}$ between two random variables X and Y, where X has a mean values μ_X and standard deviation σ_X. Suppose Y has a mean value μ_Y and standard deviation σ_Y is computed as per Equation 6:

$$\rho_{(X,Y)} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}. \tag{6}$$

The correlation matrix provides a metric ranging from -1 to +1. Features that share a correlation close to -1 imply that these features share a strong opposite relationship (when one in high, the other is low and vice-versa). Similarly, any two features sharing a correlation close to +1, share a strong positive relationship (when one is high, the other is also high). Features sharing a correlation close to 0 imply that these features have no relationship with each other. Features with an absolute correlation of between 0.68 and 1 are considered to share a strong or high correlation. Whilst features that share an absolute correlation of between 0.9 and 1 are considered to be very highly correlated [13].

2.4.2. Chi-squared test for independence

The chi-squared test for independence (χ^2) can be used to assess the association between categorical features with each other. Whilst not all dependent variables are a concern, this method will highlight possible redundancy. Suppose a random sample, having n observations, which are classified into k groups (that are mutually exclusive), having observed numbers x_i ($i = 1, 2, \ldots, k$). Taking p_i as the probability that an observation falls into the ith class and taking the expectation defined as m_i=np_i, the χ^2 statistic is calculated as per Equation 7

$$\chi^2 = \sum_{i=1}^{k} \frac{{x_i}^2}{m_i} - n. \tag{7}$$

It is important to note, that if significant dependencies do occur, features need not be removed. Rather, the significant dependencies will highlight areas for further investigation. For instance, if features are naturally dependent, then one of the two features ought to be removed. For example, the features “country” and “region” (or “international” and “country”) will naturally have a dependency and thus one ought to be removed. The null hypothesis that tests if features are independent follow a significance test where the null hypothesis is rejected if the p-value is less than or equal to the level of significance (Alpha) [9].

2.4.3. Box-and-whisker plot

A box and whisker plot can be used to visualize the association between numeric and categorical features with each other. This visual representation can inform on the possible level of redundancy as well as the variability between the two assessed features. A box and whisker plot, is a visual representation that depicts the minimum value, maximum value, 25th percentile, 50th percentile and 75th percentile of the numeric features spread across the levels within the categorical feature [14]. For a categorical feature X with k levels and numeric feature Y, a box-and-whisker plots will indicate a high level of association if within the k levels of X, the observations contain the same value of Y (max-min = 0) and the Y value differs for the k levels or groupings of the k levels of X. The box-and-whisker plots within the k categories have non-overlapping interquartile ranges.

3. Data

The underlying data represented within this study reflected web traffic data of a South African SMME (small, medium or micro enterprise) informative website. The online user tracking was conducted via Google Analytics Tracking. A data-pipeline was constructed using R (a data-science programming language) to access the Google Analytics Tracking API and imported the data onto a local database at a non-aggregated level for further processing. The methods illustrated within this article describe the feature selection process ahead of any

unsupervised machine learning models apart from any assumption validation requirements that are specific to particular models. Table 1 details the features explored within this study. The features tabulated in Table 1 reflect those that are sourced from the Google Analytics tracking tool. Whilst there are several other features available, the researcher selected these features as the most informative features that could be used within unsupervised machine learning models. Off the 25 features considered, 3 were binary, 5 categorical and 17 numeric data types.

Table 1. Features explored within the study.

Feature Name	Data Type	Feature Description
Accreditations	Numeric	Count of visits the user made to this page within each session.
Apprenticeship	Numeric	Count of visits the user made to this page within each session.
Bounces	Binary	Flags if the session was a single-page visit only.
Browser	Categorical	Web browser used to access the website (Google, Explorer, etc.).
Contact-us	Numeric	Count of visits the user made to this page within each session.
Country	Categorical	The country that the user accessed the website from.
Courses	Numeric	Count of visits the user made to this page within each session.
Customised-engineering-trading	Numeric	Count of visits the user made to this page within each session.
daysSinceLastSession	Numeric	The number of days a user is returning to the website.
DeviceCategory	Categorical	Indicating if a tablet, mobile or desktop device was used.
Distance	Numeric	The Euclidean distance between the user's co-ordinates and the company's co-ordinates (owner of the website).
Engineering-academic-studies	Numeric	Count of visits the user made to this page within each session.
Engineering-Trade	Numeric	Count of visits the user made to this page within each session.
Hits	Numeric	Represents any action on a webpage that results in data being sent to Google Analytics (such as page clicks, etc.).
Home	Numeric	Count of visits the user made to this page within each session.
International	Binary	Flags if the user is South African or not.
OrganicSearches	Binary	Flag to indicate if the user organically constructed a search that resulted in landing onto the webpage (no a web URL clicked)
Pageviews	Numeric	the number of instances a page was loaded (or reloaded)
Region	Categorical	The regions that the user accessed the website from.
sessionCount	Numeric	An indicator of the nth time the user has accessed the website.
SessionDuration	Numeric	The duration of the session (seconds).
Short-courses-skilled-programmes	Numeric	Count of visits the user made to this page within each session.
Trade-test-arpl	Numeric	Count of visits the user made to this page within each session.
University-of-technology-uot	Numeric	Count of visits the user made to this page within each session.
UserType	Categorical	Indicates if the user is a new user or returning user.

These are the typical features that are available to use on web traffic data with an information website. Some features are merely counts of activity on each web page (such as 'trade test arpl') which will be specific to the studied website. However, other tracking features (such as sessions, hits, bounce rates, etc.) are standard metrics supplied by the Google Analytics tracking tool. For modelling purposes, this study analysed the data at a session level. A session (or visit) simply represents the group of interactions (pages viewed, duration, etc.) a user made while on the website in that particular instance. A user may have multiple sessions if the user visited the website several times.

4. Empirical Results

This section discusses acceptable approaches to gauge the variability within features and the correlation between features across various data types.

4.1. Measure of variability

The variability of numeric features were quantified using the variance, mean absolute difference and dispersion ratio statistics as reported in Table 2. The non-numeric features were assessed using the coefficient of unalikeability as shown in Table 3.

Table 2. Mean and measures of variability within numeric features.

	Mean and measures of variability						
Numeric Features	Mean	Variance	Coefficient of variance	Mean Absolute Difference	Mean* 95%	Mean* 105%	Dispersion Ratio
Accreditations	**0.1419**	0.1698	2.904	0.2496	0.1348	0.149	0.1262
Apprenticeship	**0.1521**	0.1856	2.8314	0.2647	0.1445	0.1597	0.1368
Bounces	0.3966	0.2429	1.2428	0.48	0.3768	0.4164	0.3954
Contact-us	**0.1949**	0.2665	2.6492	0.3291	0.1851	0.2046	0.1672
Courses	1.0812	2.2372	1.3834	**1.0988**	1.0271	1.1353	0.6165
Customised-engineering-trading	**0.0966**	0.1369	3.8316	0.1785	0.0918	0.1014	0.0814
daysSinceLast Session	1.7043	64.2067	4.7017	2.9873	1.6191	1.7895	0.2996
Distance	11.3806	1253.9402	3.1115	17.0262	10.8116	11.9496	22.1475
Engineering-academic-studies	**0.0051**	0.0154	24.1764	0.0102	0.0049	0.0054	0.0032
Engineering-Trade	**0.0154**	0.0442	13.6725	0.0305	0.0146	0.0162	0.0115
Hits	3.8376	18.338	1.1159	2.9534	3.6457	4.0295	1.5478
Home	1.1325	**0.7857**	0.7827	0.5831	1.0759	1.1891	**0.8934**
OrganicSearches	0.4726	0.2512	1.0604	0.4993	0.449	0.4963	0.4721
Pageviews	3.8342	18.3455	1.1171	2.9538	3.6425	4.0259	1.549
sessionCount	2.1487	13.5279	1.7117	1.6609	2.0413	2.2562	1.4899
SessionDuration	235.5274	249615.2537	2.1213	288.9087	223.751	247.3037	1.4289
Short-courses-skilled-programmes	**0.0103**	0.0204	13.9343	0.0204	0.0097	0.0108	0.0088
Trade-test-arpl	0.2068	0.3969	3.0458	0.3596	0.1965	0.2172	0.1477
University-of-technology-uot	**0.1838**	0.3571	3.2521	0.3245	0.1746	0.1929	0.1324

The values in bold font highlight the features that were detected to be low variability features according to the respective metrics as discussed in section 2.

Table 3. Measures of variability within categorical features.

Categorical features	**Coefficient of unalikeability**
International	**0.142**
Country	**0.1473**
UserType	0.4167
Region	0.5018
Browser	0.5045

From Table 3 the coefficient of unalikeability indicated that the "home" feature contained low variance. Furthermore, for features with near zero mean values, the coefficient of variance highlighted that the features "accreditations", "apprenticeships", "contact-us", "customised-engineering-trading", "engineering-academic-studies", "engineering-trade", "short-courses-skilled-programmes" and "university-of-technology-uot" held very

low variability. Similarly, the feature "courses" had a mean absolute difference within a 5% interval from the mean and the dispersion ratio identified the "home" feature to be a low variance feature. With regards to the categorical features, the coefficient of unalikeability (Table 3) indicated that features "international" and "country" show little variability between observations. Exploratory analysis explained that most website visits are primarily from South Africa and a minimal portion of all visits are from elsewhere thus the low variability within the "international" and "country" features.

4.2. Measure of association

To assess the measures of association between features, the employed methods were a correlation matrix, chi-squared test for independence and box-and-whisker plots.

4.2.1. Numeric to numeric features

The correlation matrix illustrated in Figure 1 expressed the association between the numeric features with each other.

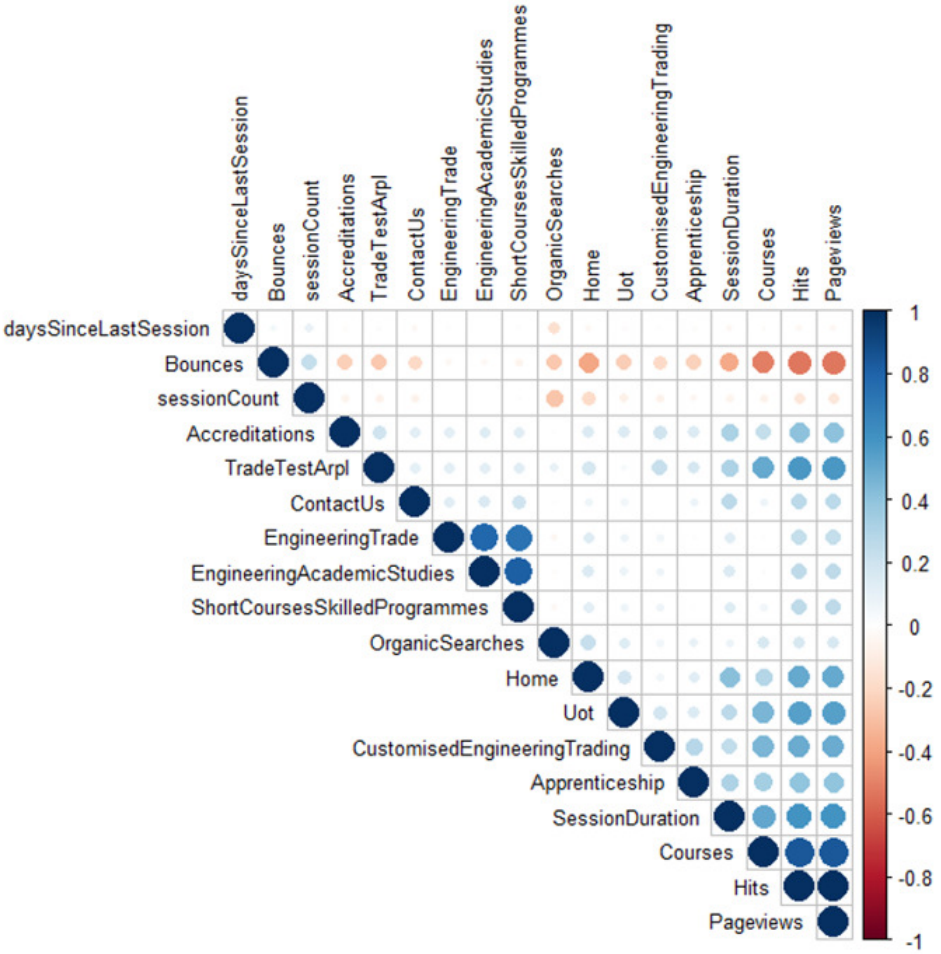

Figure 1. Correlation matrix of numeric features.

It is observed that a strong positive correlation exists between visits to certain web pages (“engineering trade”, “engineering academic studies and “short courses skilled programmes”). Whilst this suggests a strong positive, it is not advised for such features to be removed due to the correlation as this relationship may be insightful.
A somewhat strong negative correlation appears between the bounces and several other numeric features. Naturally, the higher the bounce rate, the less the interaction with the website. Since user bouncing is an important behaviour to monitor, the “bounce” feature should be included within the unsupervised machine learning model.
However the features “hits” and “pageviews” show to be very highly associated and to avoid redundancy, one of these two features should be omitted from unsupervised machine learning models. This was driven by the case study website, by design not encompassing much engagement per page with users. Thus, maintaining a pageview to hit ratio of 1:1.

4.2.2. Numeric to unordered categorical features

To inspect the association between the numeric features and the categorical features, box-and-whisker plots were constructed. Figure 2 depicts a few of the box-and-whisker plots between the categorical features and numeric features within the study. Table 4 labels the categorical features that have shown to have a high level of association with the numeric features.

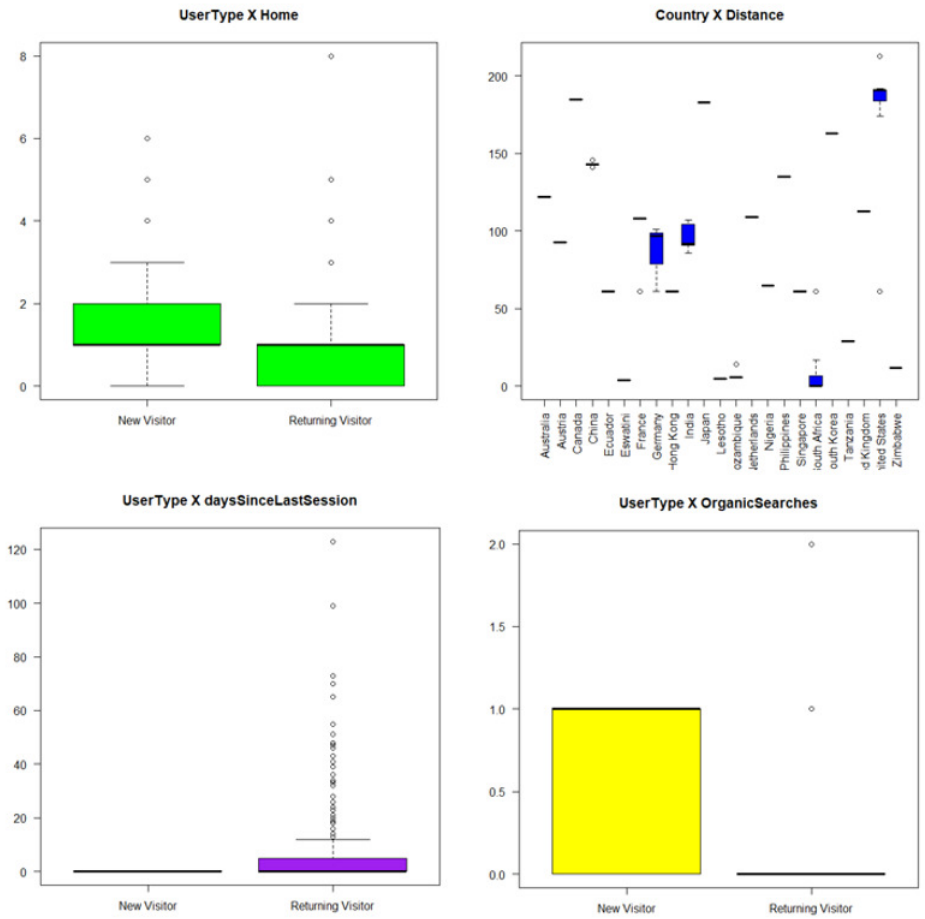

Figure 2. Box and whisker plots.

Table 4. Box-and-whisker plot levels of associations.

	Categorical features			
Numeric Features	**Browser**	**Country**	**Device Category**	**UserType**
Accreditations	Low	Low	Low	Low
Apprenticeship	Low	Low	Low	Low
Bounces	Low	Low	Low	Low
Contact-us	Low	Low	Low	Low
Courses	Low	Low	Low	Low
Customised-engineering-trading	Low	Low	Low	Low
daysSinceLastSession	Low	Low	Low	High
Distance	Low	High	Low	Low
Engineering-academic-studies	Low	Low	Low	Low
Engineering-Trade	Low	Low	Low	Low
Hits	Low	Low	Low	Low
Home	Low	Low	Low	High
OrganicSearches	Low	Low	Low	High
Pageviews	Low	Low	Low	Low
sessionCount	Low	Low	Low	High
SessionDuration	Low	Low	Low	Low
Short-courses-skilled-programmes	Low	Low	Low	Low
Trade-test-arpl	Low	Low	Low	Low
University-of-technology-uot	Low	Low	Low	Low

The box-and-whisker plots illustrated in Figure 2 highlight potential measures of high association. The high association found between "user type and dayssincelastsession", "user type and sessioncount" and "country and distance" can be attributed to the natural relationships between these features. For features with natural relationships, one of the two features ought to be omitted due to redundancy of information. The numeric features are often chosen over categorical to avoid possible loss of information. Thus, "usertype" should not be included within the same model as "dayssincelastsession" and "sessioncount". Whilst "country" and "distance" have shown to have a naturally strong relationship, both these features may be included within the unsupervised model due to potentially insightful variation that occurs within a country. For instance, South Africa has a wider distribution of distances that users accessed the website from. Furthermore, a high association has been detected between "usertype and home" and "usertype and organicsearches". This was driven by the tendency of new visitors to view the "home" page of the website more often than returning visitors. Returning visitors sought specific information on the website upon return. Returning visitors show to rarely search organically whilst this could be due to the device browser capability to conveniently route to sites previously visited. Since these relationships are insightful rather than an indication of redundancy, such features can be included within an unsupervised machine learning model together.

4.2.3. Categorical to categorical features

Chi-squared test's for independence have been used to establish the association between categorical features with each other. Table 5 presents the p-value measures of the chi-squared test for independence between the categorical variables.

Table 5. chi-squared test for independence: p-values.

Chi-Squared Test For Independence	Browser	Country	Device Category	Inter-national	Region	UserType
Browser		0.0000*	0.0000*	0.0000*	0.0000*	0.0036*
Country	0.0000*		0.0109*	illogical	illogical	0.1921
DeviceCategory	0.0000*	0.0109*		0.0000*	0.0094*	0.0708
International	0.0000*	illogical	0.0000*		illogical	0.0001*
Region	0.0000*	illogical	0.0094*	illogical		0.4594
UserType	0.0036*	0.1921	0.0708	0.0001*	0.4594	

There are several variables that show significant dependency on each other (α = 0.05). Taking for instance "browser" and "country", the dependency here is due to certain web browsers being more widely used within certain countries and since this is insightful, these measures should not be omitted from the unsupervised machine learning model due to this significant dependency. In cases where runtime is a concern, perhaps such features could be considered. User type (new or returning visitor) shows to be the least dependent variable except for international visitors whom have a tendency to only visit once on the case study website. The features "browser" and "usertype" share a dependency due to returning visitors being primarily from South Africa, and South Africans most often use Google Chrome as the device browser of choice. However, of the features with significant dependencies identified, such relationships have shown to be insightful and thus none will be omitted due to these relationships.
Thereby, of the 25 features considered, for reasons of low variability and high association (such as natural relationships), 13 features are considered for omission from an unsupervised machine learning models (a 52% degree of reduction on this study). Of which, there was one binary feature ("international"), two categorical ("Region", "UserType") and ten numeric features ("Accreditations", "Apprenticeship", "Contact-us", "Courses", "Customised-engineering-trading", "Engineering-academic-studies", "Engineering-Trade", "Hits", "Short-courses-skilled-programmes", "University-of-technology-uot").

5. Conclusion

In this article, methods for feature selection prior to unsupervised machine learning models on web traffic data have been explored. The evaluated methods focused on two important concepts: variability of the features and the association between the features. The features considered within the study were of various data types and an appropriate method had to be applied accordingly. Using the metrics variance, mean absolute difference, and dispersion ratio indicated that "accreditations", "apprenticeships", "contact-us", "courses", "customised-engineering-trading", "engineering-academic-studies", "engineering-trade", "short-courses-skilled-programmes" and "university-of-technology-uot" should be omitted from an unsupervised machine learning model on the account of low variability. Whilst "home" was also detected to contain low variability, the feature had insightful relationships with other features when the feature did vary. As discovered by the box-and-whisker plots and chi-squared tests, the features "usertype", "international" and "region" ought to be excluded due to natural relationships which would result in redundancy. Furthermore, the feature "hits" shared a very high correlation with "pageviews" and thus to eliminate redundancy the feature "hits" should be omitted. Whilst the features "engineering-academic-studies", "engineering-trade" and "short-courses-skilled-programmes" show to have a fairly strong positive correlation despite having low variability. This suggests that when these pages were viewed (although minimal), the pages were often viewed together.
Although it was found that within the study, the variability metrics employed adequately highlighted features of concern. However, it was noticed that all three metrics (dispersion ratio, mean absolute difference, variance) did not yield the same results. Future work can perhaps study these three quantitative measures used to understand the ideal environments or types of data distributions for each. Furthermore, the study resulted in a 52% degree of reduction, although omission of features were logical and scientific, perhaps further research can propose methods to evaluate potential data loss resulting from cases of harsh feature omission.
The outcome of this study is of tremendous value to data scientists and corporates building online behaviour

models. Such models are on the rise as the digital market continues to expand globally. However, much of the study would further contribute to unsupervised machine learning feature selection across several different applications.

REFERENCES

1. C. E. Brown, *Coefficient of Variation*, Springer, Berlin, Heidelberg, 1998.
2. S. Chormunge, and S. Jena, *Correlation based feature selection with clustering for high dimensional data*, Journal of Electrical Systems and Information Technology, vol. 5, pp. 542–549, 2018.
3. R. de la Cruz, and J. Kreft, *Geometric mean extension for data sets with zeros*, arXiv, 1806.06403 , 2019.
4. G. D. Dy, and C. E. Bordley, *Feature Selection for Unsupervised Learning*, Journal of Machine Learning Research, vol. 5, pp. 845–889, 2004.
5. A. J. Ferreira, and M. A. T. Figueiredo, *Efficient feature selection filters for high-dimensional data*, Pattern Recognition Letters, vol. 33, no. 13, pp. 1794–1804, 2012.
6. R. Fob, and T. Murphy, *Variable selection methods for model-based clustering*, Statistical Surveys, vol. 12, pp. 18–65, 2018.
7. R. Fraiman, A. Justel, and M. Svarc, *Selection of Variables for Cluster Analysis and Classification Rules*, Journal of the American Statistical Association, vol. 103, no. 483, pp. 1294–1303, 2008.
8. S. Guerif, *Unsupervised Variable Selection: when random rankings sound as irrelevancy*, Journal of Machine Learning Research-Proceedings, vol. 4, pp. 163–177, 2008.
9. D. Holt, A. J. Scott, and P. D. Ewings, *Chi-Squared Tests with Survey Data*, Journal of the Royal Statistical Society: Series A (General), vol. 143, no. 3, pp. 302–320, 1980.
10. G. D. Kader, and M. Perry, *Variability for Categorical Variables*, Journal of Statistics Education, vol. 15, no. 2, 2007.
11. C. Maugis, G. Celeux, M. Martin, and L. Magniette, *Variable selection in model-based clustering: A general variable*, Computational Statistics and Data Analysis, vol. 53, pp. 3872–3882, 2009.
12. J. Steven, and M. L. L. S. Turner, *A Study of Web Mining Application on E-Commerce using Google Analytics Tool*, International Journal of Computer Applications, vol. 149, no. 11, pp. 975–8887, 2016.
13. R. Taylor, *Interpretation of the Correlation Coefficient: A Basic Review*, Journal of Diagnostic Medical Sonography, vol. 6, no. 1, pp. 35–36, 1990.
14. C. Thirumalai, M. Vignesh, and R. Balaji, *Data analysis using box and whisker plot for lung cancer*, Innovations in Power and Advanced Computing Technologies (i-PACT), 10.1109/IPACT.2017.8245071, 2017.
15. Y. Thushara, and V. Ramesh, *A Study of Web Mining Application on E-Commerce using Google Analytics Tool*, International Journal of Computer Applications, vol. 149, no. 11, pp. 975–8887, 2016.
16. K. Venkatram, and G. A. Mary, *Review on Big Data and Analytics – Concepts, Philosophy, Process and Applications*, Cybernetics and Information Technologies, vol. 17, no. 2, 2017.
17. S. Wang, and J. Zhu, *Variable Selection for Model-Based High-Dimensional Clustering*, The International Biometric Society, vol. 64, no. 2, pp. 440–448, 2008.
18. W. Xing, R. Guo, G. Fitzgerald, and C. Xu, *Google Analytics based Temporal-Geospatial Analysis for Web Management: A Case Study of a K-12 Online Resource Website*, International Journal of Information Science and Management, vol. 13, no. 1, pp. 87–106, 2015.