Identifying the Key Hazards behind Website Drop-Offs by Solving a Survival Problem

Judah Soobramoney *, Retius Chifurira, Knowledge Chinhamu and Temesgen Zewotir

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal,
Durban 3630, South Africa; chifurira@ukzn.ac.za (R.C.); chinhamu@ukzn.ac.za (K.C.); zewotir@ukzn.ac.za (T.Z.)
* Correspondence: judahsoobramoney@gmail.com

Abstract: Within the modern era, corporates are compelled to own an appealing and effective website to survive and thrive within the competitive global digital marketplace. Whilst there are several web metrics to focus on, a key focus area of web analytics is the level of drop-offs. The drop-off rate represents the proportion of visitors that prematurely drop-off a website. Whilst the exact reason behind the drop-off may only be assumed (could be due to the loss of Internet connectivity or dis-interest), this study attempted to identify the triggers behind website drop-offs through a survival problem. Each person entering the website, at a given instance, can view any number of web pages (such as home, contact us, about us, etc.). However, on the studied website, roughly one in five visitors have prematurely dropped-off. The study was conducted on an engineering corporate website with the data collected via the Google Analytics tracking tool. The aim was to determine the key hazards that contributed to the observed drop-off rate through the use of a cox proportional hazard model and a survival random forest model. On the studied website, based on empirical evidence, the online visitors were censored so that those who viewed three or more webpages within the visit were labelled as 'survived'. Visitors who viewed two or less webpages before leaving the website were labelled as 'did not survive'. Thereby, the 'did not survive' observations represented the visits that prematurely dropped off the website. Using the visitor's physical and behavioral characteristics, as tracked by Google Analytics, the cox-proportional hazard and survival random forest models were employed to determine the hazards that influence survival. Visitor's physical characteristics include the device used to access the website, geolocation at the time of the visit, number of previous visits, etc., whilst the behavioral characteristics include the landing page on website, level of engagement, whether entry into the website originated through an organic search or not. Whilst both models have identified similar features as being key hazards, the survival random forest model has been shown to out-perform on the non-linear features relative to the cox proportional hazard model and obtained a higher classification accuracy. During the validation process, the survival random forest model (63%) outperformed the cox model (58%) on classification accuracy. The features that were identified as hazardous indicated that some webpages needed further attention, the visitor's level of engagement with the website (the degree of scrolling and clicks), the distance between a visitor's location and the studied corporate's location, the historic frequency of visiting the website, and if the website entry point was through an organic search. Whilst the study of drop-offs has been a commonly researched problem, this study details the investigation of key hazards through the use of survival models and compares the outcomes of a regression-based model to a machine learning survival model.

Keywords: cox proportional hazard model; google analytics; Kaplan–Meier curve; survival random forest model; webpage drop-offs

1. Introduction

In recent years, corporates rely heavily on digital channels, such as websites, to market products and transact with consumers globally. Accordingly, corporates invest in the careful scrutiny of their website usage to better understand the traffic onto the website and

the corresponding engagement that viewers express whilst on the website. This allows corporates to potentially further optimize the website to better meet customers' needs, study their interests, etc. to ultimately increase market share [1].

An engineering services corporate, TEKmation, experienced a material drop-off rate on its corporate website. Roughly one in five visitors would enter the website and thereafter leave after viewing no more than three webpages of the website (post the exclusion of bounced visits). The data was sourced from the Google Analytics tracking tool. The board of TEKmation thereafter sought to determine the factors that influenced the drop-off rates observed.

Walsh et al. (2020) investigated the occurrence of high drop-offs on a museum website. It was observed that high volumes of viewers would look at only one or two pages within 10 s of the visit, and thereafter, they would drop off. Walsh et al. (2020) probed a better understanding of the type of users who visited the website to explain the high drop-off rate. It was found that the majority of the drop-offs were linked to the understudied general public and non-professional users [2].

Rincon et al. (2022) conducted a study to measure traffic and drop-offs on Colombian banking establishments. Rincon et al. (2022) found that the websites of Colombian banking establishments were well positioned and presented low bounce rates and drop-offs [3].

Dou et al. (2013) claimed that users subconsciously assign a rapid and lasting impression on the attractiveness of the webpage within 50 ms. After which, the user's interest in and engagement with the website is highly influenced by this subconscious assessment. To address and minimize website drop-offs, Dou et al. proposed a deep neural network model to compute and quantify the webpage aesthetics, which has proven to be an effective aesthetics evaluation tool during the web design process [4].

Within this study, the observed drop-offs were treated as a survival problem. The cox proportional hazard model and survival random forest models were employed. The primary research question of the study sought to identify the hazards that have contributed to the observed drop-off rate. The owners of the website invested resources in the development and maintenance of the website and thus found the high drop-off rate concerning. Thereafter, with hazards identified, the owners of the website intended to address these hazards to ultimately boost online engagement. This paper contributes to a unique method of understanding drop-offs at the time of writing; no present literature on website drop-off hazards through a survival problem could be found with the comparison between the cox-proportional hazard model and the survival random forest model. The study of website drop-offs is fairly new and of growing concern as the world becomes more digitally enabled [1]. This study documents a detailed illustration of identifying website drop-off hazards that could be replicated by other corporates experiencing high volumes of website drop-offs. Furthermore, possible recommendations were documented to minimize the observed drop-off rates by addressing the hazards that were identified.

The related research conducted by the authors explored feature selection for unsupervised machine learning model on web analytics data [5] and further explored the use of artificial neural networks to model the COVID-19 web traffic data shift [6].

Within this paper, Section 2 details the methodologies employed within this study and discusses the underlying theoretical framework. Section 3 discusses the data source and the features observed within this study. Section 4 discusses the descriptive statistics of the data and discusses the data censoring prior to the survival model construction. Section 5 discusses the survival models, and the results are discussed in Section 6. The limitations of the study and possible future research areas are discussed in Section 7.

2. Materials and Methods

To investigate the concerning web drop-offs, this study employed two popular survival analysis techniques. The cox-proportional hazard model and the survival random forest model were employed to determine the key hazards that drove web drop-offs on the studied website.

The cox proportional hazard model follows a regression algorithm, whilst the survival random forest model falls within the family of modern machine learning algorithms.

2.1. Drop-Off Literature

The event of people or objects dropping off from given environments has been of great interest across several fields of study. This section discusses previous research on drop-offs from an educational program, sports participation, and musical participation.

Gubbels, van der Put, and Assink (2019) conducted a study to gain further insight into the risk factors associated with school absenteeism and permanent school drop-offs. The study synthesized 75 studies with 635 potential risk factors for school drop-offs. According to the study, factors such as a history of grade retention, low IQ, learning difficulties, and low academic achievements have been shown to hold high significance [7].

Eime, Harvey, and Charity (2019) conducted a study to probe the factors behind people dropping off from sports participation. The study utilized amalgamated data where participants were registered in one of eleven sporting associations. Participants were categorized based on demographics, and a comparison was conducted between registration volumes and participation volumes to better understand drop-offs. The study found that individuals playing multiple sports peaked between ages 5–14 and thereafter diminished as specialization increased. However, the drop-off in community sports participation was a concern during adolescence, and policy recommendations have been made to address the concern [8].

Pitts and Robinson (2016) investigated the individuals dropping off from classical musical participation. Through the use of interviews of current and past members to explore the themes of social acceptance, musical satisfaction, and personal confidence to establish how individual determination competed against the circumstances that would hinder musical activity. The study found that the role of music education was fundamental to lifelong participation. Furthermore, the study discussed the benefits of exposing all children to the experience and the understanding of making music [9].

2.2. Methodology Flowchart

To illustrate the methodology followed within the study, Figure 1 provides a high-level flowchart of the data collection, processing, and modelling.
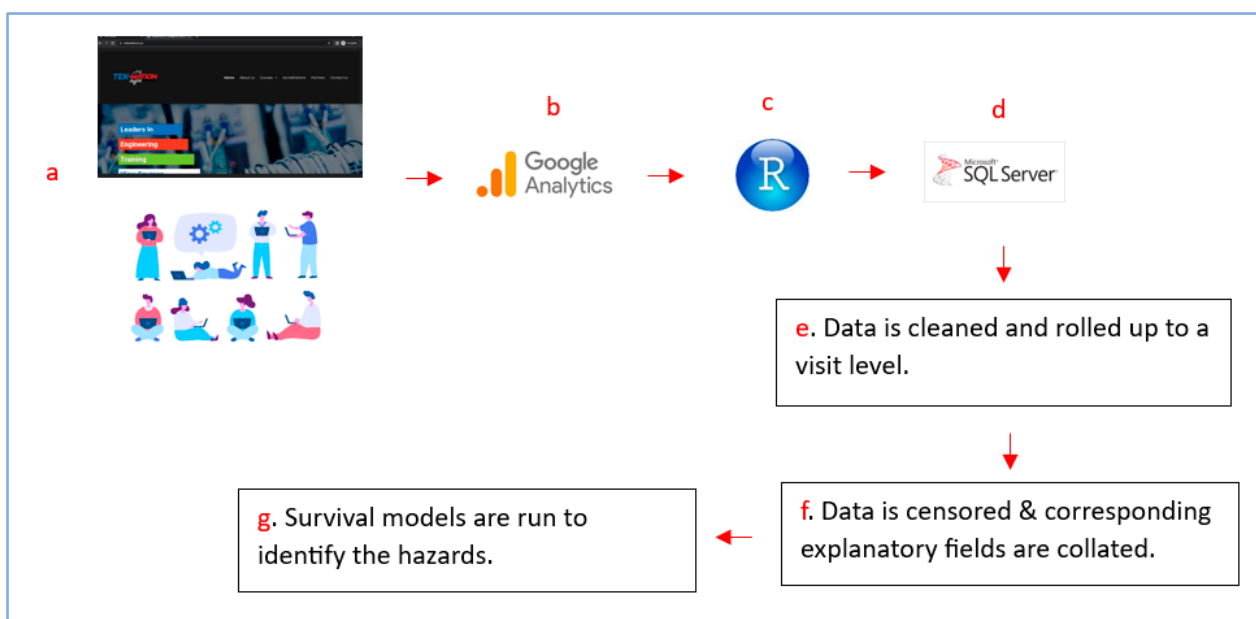
Figure 1. High-level flow chart of the methodology.

At a high-level, eight brief steps can be used to describe the methodology flow:

a. Visitors enter the studied website from their mobile devices, tablets, or computers.
b. The Google Analytics tracking tool tracks each user's activity and records important information, such as the geolocation of the device, the type of device, the detailed activity on the website, etc.
c. Through an API, R (data science programming tool) was utilized to query the data from Google Analytics and shape the data into data frames.
d. The tracking data was then written out to a local SQL environment for further processing.
e. Within SQL, the data was cleaned and rolled up to a visit level to summarize the entire engagement on the studied website per visit.
f. The data was censored to label the visits that were considered survived.
g. Finally, the data then fed into the model build and validation stages.

2.3. Kaplan–Meier Curve

The Kaplan–Meier estimator represents a statistic (non-parametric) that is employed to estimate the survival function of survival (lifetime) data. The Kaplan–Meier estimator of the survival function (S_t), which represents the probability of survival (or life) being longer than time, t, is computed in Equation (1):

$$\hat{S}(t) = \prod_{i:t_i \leq t}\left(1 - \frac{d_i}{n_i}\right), \tag{1}$$

where t_i represents a point in time where at least one event has occurred, d_i counting the number of events that occurred at this point in time (t_i), and n_i denoting the total number of individuals that have survived up to time point (t_i) where $i \,\epsilon\, t$ [10].

2.4. Cox Proportional Hazard Regression

The fundamental theory of the cox proportion hazard model holds that if the proportional hazards assumption is true, then without consideration of the full hazard model, it may be possible to estimate the effect parameters (α_i) as per Equation (2). The cox proportional hazard model holds a hazard function of the form where $X_i = (X_{i1}, \ldots, X_{ip})$ as realized values of the explanatory variables for the subject i {$i\,\epsilon\,1, \ldots, p$},

$$\lambda(t|X_i) = \lambda_0(t)e^{\alpha_1 X_{i1} + \ldots + \alpha_p X_{ip} + \alpha_0}, \tag{2}$$

which yields the hazard function at time point, t, for the subject, i, with the set of explanatory variables, X_i, and intercept term, α_0 [11].

McLernon et al. (2023) attempted to assess the performance and usefulness of predictive survival outcomes through the use of a Cox proportional hazard model. Ultimately, the recommendations of the study propose a set of performance measures that can be used to validate predictions from the survival analysis model [12].

Thiruvengadam et al. (2021) employed a cox-proportional hazards model to identify the factors that influence the duration of hospitalization on COVID-19 patients. The cox-proportional hazard model's covariates that caused longer hospitalization were abnormalities in oxygen saturation, neutrophil-lymphocyte, levels of D-dimer, lactate dehydrogenase, and ferritin. Furthermore, the findings of the study also indicated that patients with more than two chronic diseases had a significantly longer hospital stay [13].

Matsuo et al. (2019) designed a survival problem to predict the survival outcome of cervical cancer patients. The study assessed the performance of the cox proportional hazard model in comparison to deep learning models. The study found that on the observed data, the deep learning model outperformed the cox proportional hazard model. On the studied features, it was apparent that the deep learning models out-performed the cox-proportional hazard model on the non-linear hazards [14].

2.5. Survival Random Forest

The survival random forest model is a machine learning method that is employed to solve survival (time-to-event) problems [15]. The algorithm follows the steps below.

1. The train data set is boot-strapped into n subsets.
2. For each of the n sub-samples, a survival tree is composed where each node is randomly selected with $m \leq p$ where m represents the candidate number of variables considered and p is the total number of predictors. Of the m variables selected, the model would determine the optimal splitting of the variables and split points.
3. The model loops to continue recursive partitioning conditioned on no less than $d_0 > 0$ unique deaths.
4. Finally, compute the hazard function for the terminal nodes of the trees and determine the ensemble cumulative hazard function through a process of aggregation across the trees.

Jin et al. (2020) explored the use of random forests in survival analysis to understand employee attrition. The study proposed a hybrid model based on survival analysis and machine learning which combined survival analysis for censored data processing and training on attrition data patterns. The results of the study proved that the survival analysis model can materially improve the attrition prediction accuracy [16].

Soltaninejad et al. (2018) employed random forests to predict patient survival through the segmentation of brain tumors in multimodal MRI images. Within the study, the classification accuracy, pairwise mean square error, and Spearman rank metrics were all acceptable [17].

Wongvibulsin et al. (2020) conducted a study to assess the prediction of clinical risk for survival using random forest models. The findings of the model highlighted the importance of features, such as the number of preceding heart failure hospitalizations [15].

2.6. Formal Methods

During the deployment of predictive models, formal methods are often implemented to govern the predictive outcomes to minimize the impact of unexpected actions that could be the result of predictive errors [18,19].

Urban and Mine (2021) stressed the importance of software systems to behave correctly and reliably in safety-critical applications. For example, in avionics, the aircraft software has very stringent verification protocols that are mandatory as governed by international standards [20].

Within this study, the predictive survival model for website drop-offs had relatively low-risk consequences in the event of model malfunction. However, to assess the accuracy of the cox-proportional hazard and the survival random forest models, the study randomly split the website visit data into a train subset (80%) and test subset (20%). The models were trained on the training set and thereafter validated on the test subset.

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 the Google Analytics tracking tool. A data pipeline was constructed using R (a data-science programming language) to access the Google Analytics tracking API and imported the data into a local database at a non-aggregated level for further processing. To solve the survival problem, visit data were aggregated as detailed in Table 1.

Features with natural relationships have been omitted from the final dataset used within the survival problem, such as "visitor county: distance" and "DaysSinceLastSession: NewUser-Flag". In such cases, one feature would be omitted to eliminate redundant information.

Table 1. Features explored within the survival problem.

Variable	Description
SessTime	The total time that a visitor spent on the website.
Survived	Flags the visits that has survived. By definition, observations were considered to have survived if they viewed three or more webpages within the visit on the studied website.
Dist	The Euclidean distance between the visitor's coordinates and the corporate's geolocation.
DaysSinceLastSession	The count of the days since the visitor last visited the website. A value of -1 is assigned if the visitor has not visited the website previously.
OperatingSystem	The operating system of the device, e.g., Android, etc.
DeviceCategory	The device used within the visit: mobile device, tablet, or computer.
Browser	The browser type used within the visit (e.g., Google Chrome, Microsoft Edge).
OrganicSearches	A flag to indicate if the visit originated from an organic search or if a link was clicked.
PageCount	The count of the pages viewed within the visit.
Hits	The total number of hits/actions on the webpage during the visit.
H	Flagged as 1 if the first page viewed was the 'Home' page.
CR	Flagged as 1 if the first page viewed was the 'Courses' page.
T	Flagged as 1 if the first page viewed was the 'Trade-test' page.
CU	Flagged as 1 if the first page viewed was the 'Contact-Us' page.
SC	Flagged as 1 if the first page viewed was the 'Short-Course' page.
AP	Flagged as 1 if the first page viewed was the 'Apprenticeships' page.
ET	Flagged as 1 if the first page viewed was the 'Engineering-Trade' page.
EA	Flagged as 1 if the first page viewed was the 'Engineering-Academic' page.
U	Flagged as 1 if the first page viewed was the 'University-of-technology' page.
AC	Flagged as 1 if the first page viewed was the 'Accreditations' page.
CE	Flagged as 1 if the first page viewed was the 'Customised-Engineering' page.
AU	Flagged as 1 if the first page viewed was the 'About-us' page.
L	Flagged as 1 if the first page viewed was the 'Learnership' page.
CL	Flagged as 1 if the first page viewed was the 'Clients' page.

4. Exploratory Analysis

This section discusses the exploratory analysis prior to the construction of the survival models. The data exclusions, definition of censoring, and Kaplan–Meier curves are discussed.

4.1. Webpage Views

The studied data removed "bounce visits" from the analysis. By definition, a bounce visit would represent events of a person entering the website and thereafter, instantly dropping off. This is often a symptom of a person mistakenly entering a website or very quickly determining upon landing on a website that it was not what the visitor was browsing for. In the survival problem, bounce visits would materially increase the model noise [21]. The owners of the studied website (a South African corporate), by design,

intended that visitors should view a minimum of three pages per visit (post exclusion of bounce visits). Figure 2 depicts the censored page-view distribution.
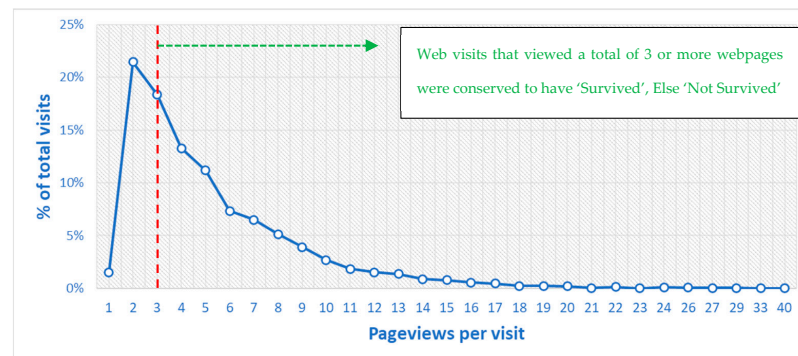
Figure 2. Pageview distribution of the studied website.

Each person entering the studied website, at a given instance, can view any number of webpages (such as home, contact us, about us, etc.). However, according to the observed data, roughly 22% of the visitors would enter the website and only view two webpages within that website and thereafter drop off. This implies that more than one in five visits did not survive to have viewed more than two web pages per visit. This paper sought to identify the hazards that contribute to the material portion of visits that failed to survive through the use of a Cox proportional hazard model and a survival random forest model. Thereby, the definition of survival followed:

visits that viewed in total three or more webpages, we considered survived.

visits that viewed in total less than three webpages were considered a failure to survive.

4.2. Kaplan–Meier Curve

A Kaplan–Meier curve was constructed to visually assess the survival rate as time progressed on the studied website (as depicted in Figure 3).
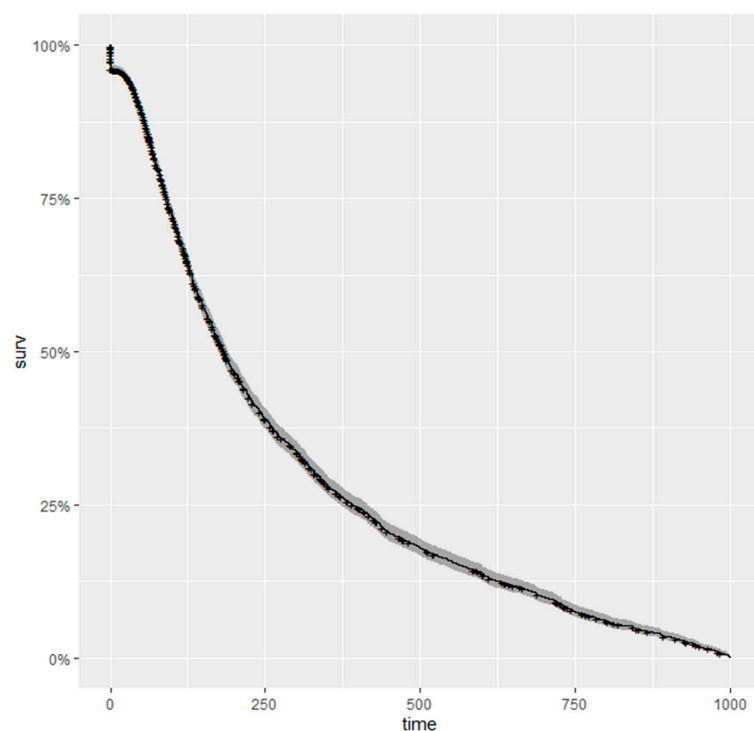
Figure 3. Kaplan–Meier survival curve on the studied website.

On the studied website, the Kaplan–Meier survival curve followed a concave distribution which suggested that survival rates steeply decline within the first 250 s of a website visit. Additionally, after 250 s, a more gradual survival rate has been noted. This implies that the premature drop-offs leave the website shortly after entering, and those that view a greater number of pages have been more engaged with the website.

5. Survival Models

Section 5 details the Cox-proportional hazard model and survival random forest models. The features fed into the models are discussed, the features that the models have identified as hazards are discussed, and the models' classification accuracy is further shared.

5.1. Cox-Proportional Model

A Cox-proportional hazard model was employed to determine the significant hazards that contribute to survival on the studied website. The model sought to solve the problem as expressed in Equation (3) where survived referred to the event of a web visit viewing three or more webpages within the visit:

$$
\begin{aligned}
&\text{surv(sesstime, survived)} \sim \\
&(\text{Hits} + \text{OperatingSystem} + \text{deviceCategory} + \text{Browser} + \text{Dist} + \text{OrganicSearches} + \\
&\text{daysSinceLastSession} + \text{H} + \text{CR} + \text{T} + \text{CU} + \text{SC} + \text{AP} + \text{ET} + \text{EA} + \text{U} + \text{AC} + \text{CE} + \text{AU} + \text{L} + \text{CL})
\end{aligned} \tag{3}
$$

Prior to computing the Cox-proportional hazard model, the categorical features (OperatingSystem, deviceCategory, and Browser) were transformed into indicator variables.

The Cox-proportional hazard model was trained on 80% of the data and tested on 20% of the data for validation. The Cox model recorded a classification accuracy of 58%.

Figure 4 depicts the features and the corresponding hazard level of significant as determined by the Cox-proportional model.
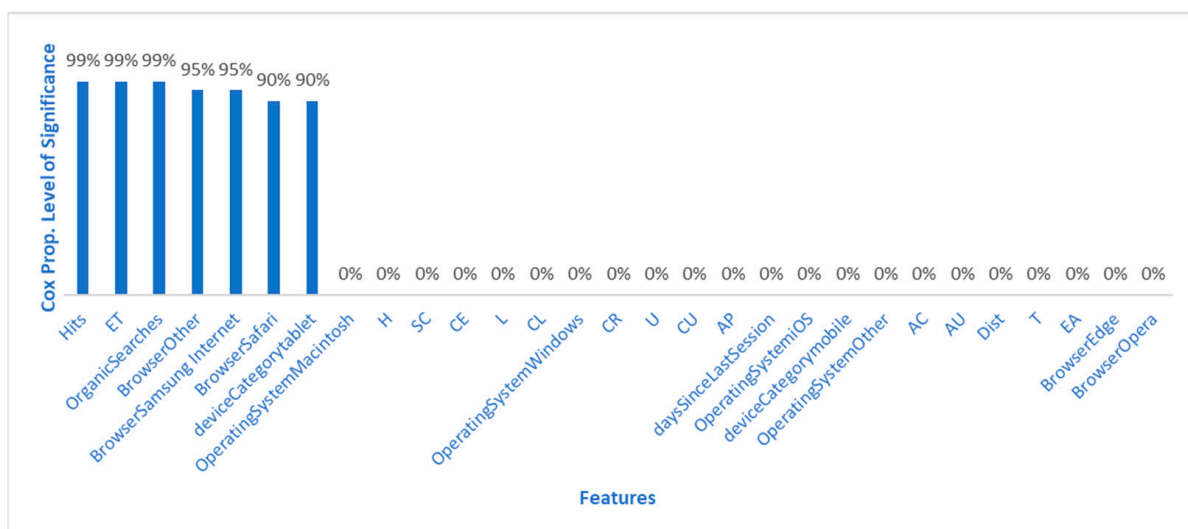
Figure 4. Cox-proportional hazard feature level of significance.

According to the Cox-proportional hazard model, the features, "Hits", "ET", "OrganicSearches", "Browser", and "DeviceCategory" were significant hazards to survival on the studied website. Further descriptive analysis showed that visits that were very engaged with the website (number of hits per page) had a high likelihood of survival. By definition, a hit represents the number of clicks a visitor makes when on a webpage. The feature "ET" represents visits where the first webpage viewed was the "'Engineering-Trade" webpage. According to the data, visits with the first page being the "Engineering-Trade" have been shown to have a high likelihood to drop-off early. This is suggestive that either such visits

get all the information they need from the webpage or are dissatisfied and warranted to drop-off.

The data also informed that visits that originated from an organic search had a high likelihood of survival on the website, and conversely, those visits that were not organically originated have showed a high likelihood to drop-off prior to viewing three or more webpages on the studied website. The Cox-proportional model detected that visits from the browsers, "other", "Samsung Internet", and "Safari" had higher likelihoods of dropping off prior to three page views. This could be due to either a population trait of the users of such browsers (or devices associated with such browsers) or the visual presentation of the corporate website when using these browsers.

5.2. Survival Random Forest Model

The web traffic data has been modelled through a survival random forest model to identify the hazardous features that have contributed to visits not surviving three or more web pages on the studied website. The random forest's relative importance scores are depicted in Figure 5. The survival random forest model was trained on 80% of the data and tested on 20% of the data for validation. The random forest model recorded a classification accuracy of 63%.
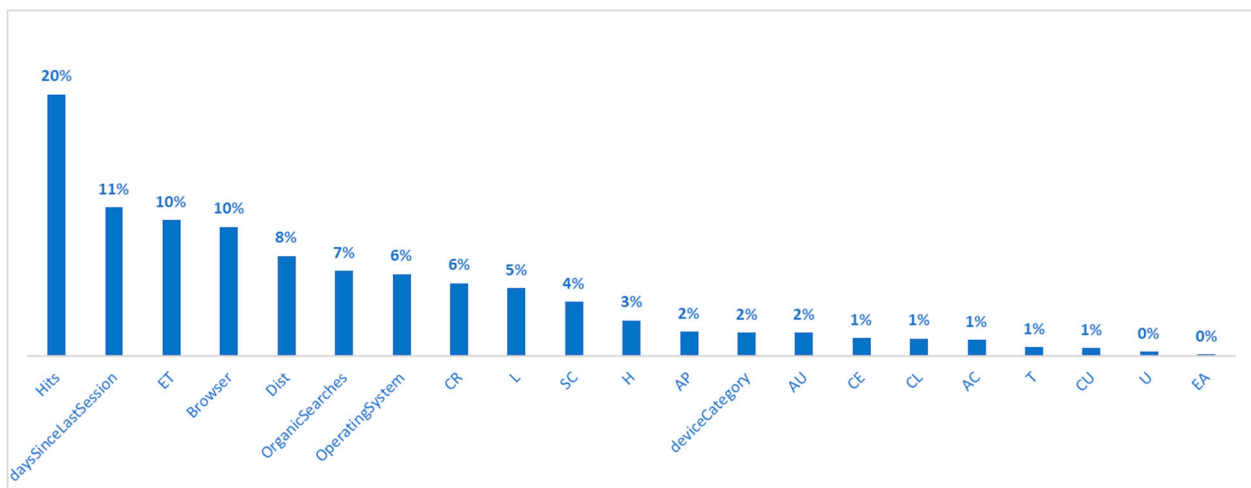
Figure 5. Survival random forest importance scores.

According to the survival random forest model, "Hits", "daysSinceLastSession", "ET", "Browser", "Dist", and "OrganicSearches", were amongst the most important features in determining if a visit would survive or not on the studied website. Empirical analysis has shown that the less engaging a visit was per webpage (measured by the hits: clicks, and scrolls), the lower the chance of survival. The random forest model also detected that visits that were either very frequent or previously visited between 80–100 days ago have been shown to have a low-survival rate. The survival model has also detected that the visitor's Euclidean distance between the device used to browse the studied website and the corporate's coordinates is an important hazard. The further the distance, the lower the survival rate as expressed in the empirical data. This typically represented foreign visits. The feature, "ET", has been identified as a hazard since visits with the first page being "Engineering-Trade" have been shown to have a low-survival rate on the studied data. The feature, "Browser", has been identified with high importance; visits with browsers, such as "Other", "Samsung Internet", and "Safari", have been shown to have a low-survival rate within the data. This could be an indication of either the owners of such devices having an unknown tendency to drop off, or perhaps the visual display of the website on these browsers could contribute to a low survival. The feature, "OrganicSearch", has also been

featured as an important hazard, and according to the empirical data, it can be established that visits that have originated from an organic search have a high-survival rate.

6. Conclusions

The shareholders of the studied website were concerned with the high degree of premature drop-offs. According to shareholder expectations, visitors should view at least three or more web pages per view. A low page view website implies that the marketing cost invested in the website is not being justified according to the intended purpose. This study employed a traditional survival model (through the use of the Cox-proportional hazard model) and a machine learning model (survival random forest model) to identify the underlying hazards that drove the high degree of drop-offs observed. The Cox-proportional hazard model was able to indicate features that were statistically significant in contributing to survival rates, whilst the survival random forest model was able to identify the features that are important in predicting the survival rates. Although the models employed are fundamentally different, the features identified as hazardous were understandable and evident in the empirical data. Table 2 below depicts the features that were identified as hazardous under each model.

Table 2. Web survival hazards comparison between models.

Feature	Cox-Proportional Hazard Model	Survival Random Forest Model
AC		
AP		
AU		
Browser	✓	✓
CE		
CL		
CR		
CU		
daysSinceLastSession		✓
deviceCategory	✓	
Dist		✓
EA		
ET	✓	✓
H		
Hits	✓	✓
L		
OperatingSystem		
OrganicSearches	✓	✓
SC		
T		
U		

Table 2 marks the features that were identified as possible hazards by the Cox-proportional hazard model and the survival random forest model. Whilst both models are very different in theory, it was observed that a material overlap existed in the features identified as hazards. The features, "Browser", "ET", "Hits", and "OrganicSearches", were identified as hazardous by both the Cox-proportional hazard model and the survival random forest model. The survival random forest model has been shown to better identify the hazardous features that were non-linear, such as "Dist" and "DaysSinceLastSession". Furthermore, the Cox-proportional hazard model identified the feature, "deviceCategory", whilst the survival random forest model did not.

The study illustrates how the hazards of the studied website could be successfully identified through a survival problem. The hazard, "Browser" (such as Google Chrome, Explorer, Firefox, etc.), indicates that there potentially could be a compatibility flaw on certain browsers that results in visitors prematurely dropping off. The developers of the website would need to re-test the website loading on the browsers of concern. The feature,

"ET", was detected as hazardous and thereby, indicated that visitors landing onto the website through this webpage were perhaps discouraged by this webpage. Further tests would need to be conducted to isolate the underlying reason (through random AB testing). Visitors could have either dropped off due to the wording, due to the navigation to other webpages from this webpage, due to the imagery, etc. The feature, "daysSinceLastSession", was detected as a non-linear hazard. Exploratory analysis has shown that visitors who were new (first time visitor) or visited a very long while ago have been shown to have had a higher survival rate. This implies that people who recently visited the website previously either entered again mistakenly or came on looking for specific information and thereafter dropped off. The "deviceCategory" feature was selected as a hazard as visitors using a "desktop PC" had a higher tendency to drop off relative to "tablet" and "mobile" device users. Further research needs to be conducted to understand this behavior. Further exploratory analysis has indicated that visitors from foreign countries enter the studied website and shortly dropped off. The studied corporate at the time of analysis had only traded in and around South Africa and thereby, suggested that visits from outside of South Africa were most likely by mistake and thus explained the low survival rate as detected by the "Dist" hazard. The feature, "Hits", has been identified as a hazard to survival. The feature, "Hits", represents the degree of clicks and scrolls that a visitor has performed during the visit. Thus, intuitively, the lower the "Hits", the lower the engagement, and most likely, is an early indication of uninterest. Further testing should be performed, to assess the impact of website prompts that can be thrown at a visitor with low engagement to try increase the interest and engagement with the website to ultimately prevent premature drop-offs. The feature, "OrganicSearches", represents the way in which a visitor found the website, i.e., indicating if the visitor followed a link, or used key words to describe what he is looking for and the search engine yielded its recommendations and thereafter the visitor entered the corporate website. Whilst this hazard may correlate with "first time visitors", further testing can be conducted to throw prompts to visitors who did not enter the website through an organic search to increase engagement and prolong the web journey on the studied website. Ultimately, although both the Cox-proportional hazard model and survival random forest models have performed well in identifying the key hazards, the survival random forest model has outperformed the Cox model due to its ability to better comprehend non-linear features and ultimately attained a superior classification accuracy relative to the Cox model.

Whilst dropping off has been a widely studied phenomenon across several applications, this study uniquely attempts to identify the underlying hazards by solving a survival problem. In a similar manner, the above methods have proved useful, and thus, web-owners across several domains who experience high levels of drop-offs can likewise identify the underlying hazards by solving a survival problem.

Unlike Walsh et al. (2020) who studied the type of visitor that dropped off [2], the use of a survival model can be generalized to any website and the outcome clearly indicates the hazards that need to be addressed. Dou et al. (2013) employed machine learning models to improve the visual appeal of the website to minimize drop-offs; however, visual appeal may not be the only factor that would result in premature drop-offs [4]. For instance, the survival models have identified behavioral elements that may not be influenced by visual appeal alone (such as organically searched visitor, visitor dependency on the days since the last session, etc.).

7. Limitations and Research Directions

A key limitation of the study is the assumption that a person can be generalized to a visitor. Web-tracking tools are only able to track the IP address of a visitor and thereby, track the number of times the visitor enters the website. However, in reality, the same person may visit using his mobile cellphone, later his PC, and tomorrow his tablet. Tracking tools will then see these are three independent persons since the studied website does not require

visitors to log in. Nonetheless, the counter assumption is that this behavior is somewhat rare, and in a big dataset, such occurrences should be normalized.

Given that the study has successfully identified the key hazards that influence premature drop-offs, the initial planned responses are re-design tests (in the case of the "ET" and "Browser" hazards) and website prompts to address the "daysSinceLastSession", "deviceCategory", "Hits", and "OrganicSearch" hazards. However, to optimize these prompts, AB testing would have to be conducted to ensure that the web prompts are effective and minimize premature drop-offs.

Author Contributions: Conceptualization, J.S., R.C., K.C. and T.Z.; methodology, J.S.; software, J.S.; validation, J.S., R.C., K.C. and T.Z.; formal analysis, J.S.; investigation, J.S.; resources, J.S.; data curation, J.S.; writing—original draft preparation, J.S.; writing—review and editing, J.S.; visualization, J.S.; supervision, R.C., K.C. and T.Z.; project administration, J.S.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Awichanirost, J.; Phumchusri, N. Analyzing the effects of sessions on unique visitors and unique page views with google analytics. In Proceedings of the IEEE 7th International Conference on Industrial Engineering and Applications, Bangkok, Thailand, 16–21 April 2020; Volume 1, pp. 1014–1018.
2. Walsh, D.; Hall, M.M.; Clough, P.; Foster, J. Characterising online museum users: A study of the national museums Liverpool museum website. *Int. J. Digit. Libr.* **2020**, *21*, 75–87. [CrossRef]
3. Rincon, J.S.R.; Tarazona, A.R.R.; Martinez, A.M.M.; Acosta-Prado, J.C. Positioning and web traffic of Colombian banking establishments. *J. Theor. Appl. Electron. Commer. Res.* **2020**, *17*, 1473–1492. [CrossRef]
4. Dou, Q.; Zheng, S.Z.; Sun, T.; Heng, P.A. Webthetics: Quantifying webpage aesthetics with deep learning. *Int. J. Hum.-Comput. Stud.* **2019**, *124*, 56–66. [CrossRef]
5. Soobramoney, J.; Chifurira, R.; Zewotir, T. Selecting key features of online behaviour on South African informative websites prior to unsupervised machine learning. *Stat. Optim. Inf. Comput.* **2022**, *11*, 519–530. [CrossRef]
6. Soobramoney, J.; Chifurira, R.; Zewotir, T. Modelling the South African Covid-19 induced web traffic data shift using artificial neural networks. *Appl. Math* **2022**, *16*, 1049–1056.
7. Gubbels, J.; van der Put, C.E.; Assink, M. Risk factors for school absenteeism and dropout: A meta-analytic review. *J. Youth Adolesc.* **2019**, *48*, 1637–1667. [CrossRef] [PubMed]
8. Eime, R.M.; Harvey, J.T.; Charity, M.J. Sport drop-out during adolescence: Is it real, or an artefact of sampling behaviour? *Int. J. Sport. Policy Politics* **2019**, *11*, 715–726. [CrossRef]
9. Pitts, S.E.; Robinson, K. Dropping in and dropping out: Experiences of sustaining and ceasing amateur participation in classical music. *Br. J. Music. Educ.* **2016**, *33*, 327–346. [CrossRef]
10. D'Arrigo, G.; Leonardis, D.; Elhafeez, S.A.B.D.; Fusaro, M.; Tripepi, G.; Roumeliotis, S. Methods to analyse time-to-event data: The Kaplan-Meier. *Oxidative Med. Cell. Longev.* **2021**, *1*, 2290120.
11. Kvamme, H.; Borgan, O.; Scheel, I. Time-to-event prediction with neural networks and Cox regression. *J. Mach. Learn. Res.* **2019**, *20*, 1–30.
12. McLernon, D.J.; Giardiello, D.; van Calster, B.; Wynants, L.; van Geloven, N.; van Smeden, M.; Therneau, T.; Steyerberg, E.W. Assessing performance and clinical usefulness in prediction models with survival outcomes: Practical guidance for Cox proportional hazards models. *Ann. Intern. Med.* **2023**, *176*, 105–114. [CrossRef] [PubMed]
13. Thiruvengadam, G.; Lakshmi, M.; Ramanujam, R. A study of factors affecting the length of hospital stay of COVID-19 patients by Cox-proportional hazard model in a South Indian tertiary care hospital. *J. Prim. Care Community Health.* **2021**, *12*, 21501327211000231. [CrossRef] [PubMed]
14. Matsuo, K.; Purushotham, S.; Jiang, B.; Mandelbaum, R.S.; Takiuchi, T.; Liu, Y.; Roman, L.D. Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. *Am. J. Obstet. Gynecol.* **2019**, *220*, 381.e1–381.e14. [CrossRef] [PubMed]
15. Wongvibulsin, S.; Wu, K.C.; Zeger, S.L. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Med. Res. Methodol.* **2020**, *20*, 1. [CrossRef] [PubMed]
16. Jin, Z.; Shang, J.; Zhu, Q.; Ling, C.; Xie, W.; Qiang, B. RFRSF: Employee turnover prediction based on random forests and survival analysis. *WISE J.* **2020**, *1*, 503–515.

17. Soltaninejad, M.; Zhang, L.; Lambrou, T.; Yang, G.; Allinson, N.; Ye, X. MRI Brain Tumor Segmentation and Patient Survival Prediction Using Random Forests and Fully Convolutional Networks. *Lect. Notes Comput. Sci.* **2018**, *1*, 204–215.

18. Krichen, M.; Mihoub, A.; Alzahrani, M.Y.; Adoni, W.Y.H.; Nahhal, T. Are formal methods applicable to machine learning and artificial intelligence? In Proceedings of the 2nd International Conference of Smart Systems and Emerging Technologies, Riyadh, Saudi Arabia, 9–11 May 2022; Volume 1, pp. 48–53.

19. Raman, R.; Gupta, N.; Jeppu, Y. Framework for formal verification of machine learning based complex system-of-systems. *INCOSE Int. Symp.* **2021**, *31*, 310–326. [CrossRef]

20. Urban, C.; Mine, A. A review of formal methods applied to machine learning. *arXiv* **2021**, arXiv:2104.02466.

21. Marios, P.; Nikos, K.; Sozon, P. Assessing stationarity in web analytics: A study of bounce rates. *Expert. Syst.* **2020**, *37*, e12502.