A Multi-Dimensional Approach to Understanding the Effect of Page Content and Infrastructure on Page Load Time

Daniel A. Oliveira[1], Rosa M. M. Leão[1], Edmundo de Souza e Silva[1]

[1]Computer Systems Engineering – Federal University of Rio de Janeiro (UFRJ)
PO Box 68511, 21941-97 – Rio de Janeiro – RJ – Brazil

{danoliveira,rosam,edmundo}@land.ufrj.br

Abstract. *We study the metric Page Load Time (PLT) which has a significant impact on user experience, search engine optimization, and conversion rates. We explore how page complexity metrics, specifically content and infrastructure, affect PLT. We employ both supervised and unsupervised machine learning models to analyze the influence of these metrics at multiple levels: single page, page category, cluster, and general. Our study shows that the number of bytes, requests, and distinct images are key features in PLT prediction, with the page category model generally outperforming others. The results contribute to a better understanding of the factors influencing PLT and show some insights into how to optimize web pages for better user experiences and business outcomes.*

1. Introduction

Web pages remain an indispensable part of our everyday lives, serving as gateways to a variety of services, such as information, entertainment, and commerce. One crucial aspect influencing the user's browsing experience is page load time (PLT). PLT is the duration taken by a web page to load its entire content, encompassing text, images, videos, and interactive elements, that is, the time it takes from the initial request to the final rendering.

Providing a fast and responsive website is essential for guaranteeing a good experience for end-users, as indicated by substantial research from leading corporations. Over a decade ago, a stufy by Amazon highlighted that every 100 ms of added latency cost 1% in sales. In 2006, Google noted that an additional 0.5 seconds in search page generation time led to a 20% drop in traffic [Gigspaces 2023]. These and other cases demonstrate why companies often invest significant resources in optimizing their websites to avoid such costly delays.

The COVID-19 pandemic increased the importance of delivering a high-quality user experience in web services. Business operations changed, and an effective online presence was critical. Companies that provide low-quality digital experiences have faced significant challenges in maintaining competitiveness and relevance in today's rapidly evolving digital world. As a site becomes less interactive, users increasingly tend to move to a competitor's site. User engagement is critically linked to a website's interactivity, of which PLT is a key determinant.

Page load time emerges as an important factor not only in user experience but also in broader aspects such as search engine optimization and conversion rates. The significance of PLT is recognized by Google, which has integrated this metric into its search engine ranking algorithms [Google 2010]. Moreover, there is a significant correlation between PLT and conversion rates. Pages that load within one second have been found to have conversion rates nearly 2.5 times higher than pages that take five seconds or more to load [Wiegand 2022]. This indicates that faster pages are more effective at converting visitors into customers. Additionally, PLT plays a crucial role in user-perceived quality. Studies like [Hora et al. 2018] employing the Absolute Category Rating (ACR) scale, a self-reported measure of user experience, show a direct correlation between PLT and user satisfaction. Faster loading times are consistently associated with higher ACR ratings, indicating that users perceive websites with shorter load times as more efficient and user-friendly.

In summary, enhancing PLT is not just a technical necessity but a fundamental component of delivering an acceptable user experience. It impacts various dimensions, from search engine rankings and visibility to conversion rates and overall user QoE. Therefore, understanding how page content and infrastructure influence PLT remains crucial. Previous works exploring the relationship between page complexity metrics, namely page content and infrastructure metrics, and PLT have focused either on a few selected pages, with individual analyses being performed for each [Asrese et al. 2019], [Vogel and Springer 2022], or on a diverse group of pages, with an overall analysis conducted for all of the pages simultaneously [Saverimoutou et al. 2019], [Butkiewicz et al. 2011].

In view of the above, this paper investigates the relationship between page complexity metrics and PLT. Specifically, we aim to answer the following question: what are the most important features when it comes to inferring page load time? Supervised and unsupervised machine learning models are employed to provide a broader view of this relationship at multiple levels of granularity: single page, page category, cluster, and "general" (which includes all pages). This approach provides insights at each level, which are then compared and evaluated. The main contributions of this work can be summarized as follows:

- **Multi-dimensional analysis**: In our multi-dimensional analysis, we examined the impact of page complexity metrics on PLT across various levels of granularity. This includes "general," per-category, per-cluster, and per-page analyses, with both supervised and unsupervised models having been utilized.
- **Feature importance**: Our findings reveal that the number of bytes ranks among the top three features for inferring PLT in all models under study. In the models categorized by page types, both the number of bytes and the number of requests emerged as the top two features for most categories. These observations from the page category models are consistent with findings from the individual page model, where the three most important features were identified as the number of bytes, the number of requests, and the number of distinct images.
- **Effectiveness of different models**: We individually tested the general model, the page category model, and the cluster model for each page. The page category model outperformed both the cluster and general models for ten out of the 15 pages. Notably, in most models, the root mean squared error (RMSE) is less than one, indicating a prediction error in estimating PLT of less than one standard deviation.

2. Related Work

Extensive research has thoroughly examined page performance metrics, such as page load time (PLT), time to first paint, and above-the-fold time, in relation to the perceived quality of web browsing experiences. Studies involving passive and active user interactions with pages [Salutari et al. 2019, Hora et al. 2018, Gao et al. 2017, Egger et al. 2012] carried out evaluations based on several scales. These range from bad/neutral/good to a 5-point ACR scale, or involve choosing preferred pages from a group [Salutari et al. 2019, Hora et al. 2018, Gao et al. 2017]. Subsequently, these metrics are correlated with user scores using either expert analysis or machine learning models [Egger et al. 2012, Hora et al. 2018, Hoßfeld et al. 2018, Jahromi et al. 2018].

In addition to studying the relationship between page performance metrics and the web users' QoE, it is also important to analyze how these metrics are influenced by various factors such as page structure/content, content provision strategies, and network conditions. In [Avram et al. 2014], the authors introduce a new metric known as the latency amplification factor, which measures the impact of latency on page load times. To study this metric, they collect the page's dependency graph and estimate the overall effect on page load time through artificially added latency. In [Vogel and Springer 2022], the authors report that 70% of JavaScript and 90% of CSS scripts are loaded as render-blocking code, often only utilized after the page has finished rendering. This highlights a good opportunity for optimization. For social media and news pages, our findings indicate that the number of CSS objects and the number of JavaScript objects are, respectively, the most important features for predicting page load time. This could be partly attributed to the inefficient loading of this content. In [Saverimoutou et al. 2019], an analysis of time to first visual rendering showed lower RTTs (Round Trip Times) and fewer requests in "good response" navigations under various conditions, underscoring the significance of the number of requests in page load time prediction. Similarly, [Butkiewicz et al. 2011] finds a strong correlation between the number of bytes and PLT, while also identifying the number of requests as the best predictor of PLT.

Our findings align with prior research, highlighting the importance of certain features in predicting page load time, such as the number of requests and the number of bytes. Notably, our study shows the significant impact of image-related features, particularly in the context of single pages, and the importance of the number of servers in a diverse group of pages, with respect to page load time prediction. Additionally, category-specific relationships are also found, providing a more comprehensive understanding of how these metrics influence page load time than previous works.

3. Methodology

To analyze the relationship between page complexity metrics and page load time, it is essential to employ a software solution capable of navigating automatically to various web pages and gathering all relevant metrics for each session. For the automatic navigation, Node Puppeteer was used, instrumenting the Google Chrome browser, which was loaded with a plugin developed in [Hora et al. 2018]. This plugin collects important page loading timing information as well as page complexity metrics and was modified to send this data to our collection server. The complexity metrics collected include: the total number of servers contacted, the number of bytes, the number of requests, the number of CSS objects, the number of JavaScript (JS) objects, the number of image objects, the number of distinct images, and the number of image pixels.

We collected over two months of data, from July 5, 2023, to September 18, 2023. The experiments were conducted using five Raspberry-Pi 4 Model B, each equipped with 4GB of RAM and a 1500MHz CPU with 4 cores. These Raspberry-Pi devices were located across three different ISPs. Custom data-collecting software [1] was installed in the Raspberry-Pi´s and each unit was placed in the home of a volunteer, connected directly to the residential router via an Ethernet cable. After each web page is accessed the data from that navigation is sent in JSON format to a collection server for later analysis.

The structure of some web pages can change frequently. Therefore, to obtain a statistically significant number of samples for each page structure profile, the sampling interval varied according to an exponential distribution with a mean of 30 minutes. In each test, all selected pages were navigated in sequence. This approach limits the number of pages that can be analyzed, which was defined as 20. If a page did not fully load within 18 seconds, the navigation was canceled, and no data was collected for that page in that test period. This methodology was adopted in [Saverimoutou et al. 2019]. Consequently, each test had a maximum duration of approximately 400 seconds, which is just under seven minutes. The order in which the pages were selected followed a sampling-without-replacement strategy, ensuring that each page was tested exactly once before being excluded from subsequent selections.

The 20 pages were selected from a curated list of the top 100 most visited websites in Brazil during 2022[2]. These pages were chosen based on the following criteria: preference for higher ranked pages, broad representation of web page categories, and preference for pages where the landing page contains the content users usually consume on the site.

One limitation of this study is that it only uses landing pages and not internal web pages. This issue was addressed by [Aqeel et al. 2020], who found that two thirds of the analyzed papers required at least minor revisions due to this limitation. The final selection criterion was chosen to partially mitigate this issue, expecting that this limitation would less impact web pages with similar landing and internal pages. Out of the 20 selected pages, 18 received sufficient navigation data on each of the Raspberry-pis. The resulting 18 pages, along with their respective ranks and categories, are presented in Table 1. From the top 34 most visited pages, we selected 18 based on the criteria described above. In this group of 18 pages, the representation of each category was significant, with categories being represented at levels ranging from 50% to 100% of those found in the list of the top 34 pages.

Following data collection, the next phase was data analysis. To determine the most critical page complexity metrics for inferring PLT, various feature importance methods were utilized. Specifically, these methods included recursive feature elimination, forward and backward sequential feature selection, and Gini importance. These methods were applied in conjunction with traditional regression models. Interpretable regression

[1] wptagent-automation: `https://github.com/danielatk/wptagent-automation`

[2] Top 100 most accessed sites in Brazil [2022 Edition]: `https://pt.semrush.com/blog/top-100-sites-mais-visitados/`

Table 1. Pages Tested

Rank	Page	Category
21	123movies.net	streaming
19	amazon.com.br	e-commerce
30	americanas.com.br	e-commerce
11	caixa.gov.br	government
4	globo.com	news
35	letras.mus.br	music
26	magazineluiza.com.br	e-commerce
14	mercadolivre.com.br	e-commerce
28	olx.com.br	e-commerce
7	pornhub.com	adult
22	reddit.com	social media
29	shopee.com.br	e-commerce
18	spankbang.com	adult
32	tiktok.com	social media
34	twitch.tv	streaming
9	twitter.com	social media
5	uol.com.br	news
3	xvideos.com	adult

models, such as decision trees and random forests, were also employed, offering additional insights into the significance of the selected features. The analysis was conducted at different levels of granularity: individual page level, page category level, and general level.

Additionally, an unsupervised approach was employed to cluster pages into groups based on similar page complexity metrics, with analyses also performed at this level of granularity. This approach involved using tensor decomposition to identify the principal relationships among the page complexity metrics in a low-dimensional space. Following this, the k-means algorithm was applied to cluster the data, enabling the impact of complexity metrics on PLT within each cluster to be examined.

4. Results and Discussion

4.1. Feature Importance

As a first step to understanding the effect of complexity metrics on PLT, feature importance was analyzed. This gave an understanding of which metrics had a significant impact on PLT. For this, the following methods were used: Pearson correlation, recursive feature elimination (RFE), forward and backward sequential feature selection (FSFS and BSFS, respectively) and Gini importance.

Linear regression, decision tree, and random forest models were employed, complemented by the various feature selection methods. For each combination of model and feature selection method, 5-fold validation was used. In this process, the data was divided into training and testing groups, with 80% used for training and 20% for testing in each fold.

Table 2 displays the features selected by each method. The Gini importance measure, which was employed together with the random forest model, selected three features, by calculating the RMSE for each number of features and applying the elbow method (for feature selection). The Pearson correlation coefficient method also chose three features. The RFE, FSFS, and BSFS methods utilized all features, except in the case of the decision tree combined with the BSFS method, where only four features were selected. Number of servers was the only feature selected by all methods.

Table 2. Feature importance per method

feature	Gini	Corr.	RFE	FSFS	BSFS+LR	BSFS+DT
number of servers	**0.321**	**0.488**	**1**	**1**	**1**	**1**
number of CSS objects	**0.166**	0.102	**1**	**1**	**1**	**1**
number of bytes	**0.149**	-0.167	**1**	**1**	**1**	0
number of JS objects	0.100	**0.501**	**1**	**1**	**1**	**1**
number of image pixels	0.094	**-0.283**	**1**	**1**	**1**	0
number of distinct images	0.087	0.232	**1**	**1**	**1**	0
number of requests	0.066	0.125	**1**	**1**	**1**	**1**
number of image objects	0.016	-0.097	**1**	**1**	**1**	0

Table 3. Results of PLT prediction using different models

Model	Feature Set	RMSE	MAE
Linear Regression	Gini	0.814	0.645
Linear Regression	Corr.	1.001	0.909
Linear Regression	RFE	0.739	0.575
Linear Regression	FSFS	0.739	0.575
Linear Regression	BSFS	0.739	0.575
Decision Tree Regression	Gini	0.483	0.260
Decision Tree Regression	Corr.	0.410	0.235
Decision Tree Regression	RFE	0.423	0.218
Decision Tree Regression	FSFS	0.423	0.218
Decision Tree Regression	BSFS	0.412	**0.214**
Random Forest	Gini	**0.409**	0.230
Random Forest	Corr.	0.415	0.232

The RMSE and mean absolute error (MAE) of the PLT predictions for all models are shown in Table 3. For linear regression the feature selection performed with recursive feature elimination yielded the best results. For decision tree regression the results using the different feature sets were all very similar, with the lowest MAE obtained when using RFE and the lowest RMSE obtained when using the correlation features. With random forest the results were comparable whether using correlation metrics or features identified as most important by the method. The best performing model out of all of them, in relation to RMSE, was random forest, using the feature set obtained via Gini importance. The features selected via this method were number of servers, number of CSS objects, and number of bytes.

The interpretation of random forest models is quite limited given that it is an ensemble model that uses many individual decision trees. One way of circumventing this is through the use of a global surrogate; i.e., a simpler, usually more interpretable model that can be used to reproduce the behaviour of a more complex model [Molnar 2019]. In this case, a decision tree model was employed as a global surrogate to the random forest model. This can be done by training the decision tree with the original training set, but replacing the labels with the output of the random forest model. The tree for the surrogate model is shown in Figure 1. The maximum depth was set at 2 for better visualization. The root divides the tree according to number of servers, using 16 as the cut-off point, with two thirds of the samples being below that. For those samples, the next discriminator is number of CSS objects, with 47 as the cut-off point, with just under a tenth of the samples being above that. The samples with higher number of CSS objects had an average PLT roughly double that of the samples with lower values of CSS objects. Counterintuitively, samples with number of servers above 16 and number of kBytes above 1086 resulted in a low PLT of around 3.6 seconds. However, this occurred in only a small fraction of the samples (approximately 0.2%), indicating that such instances should be considered special cases.

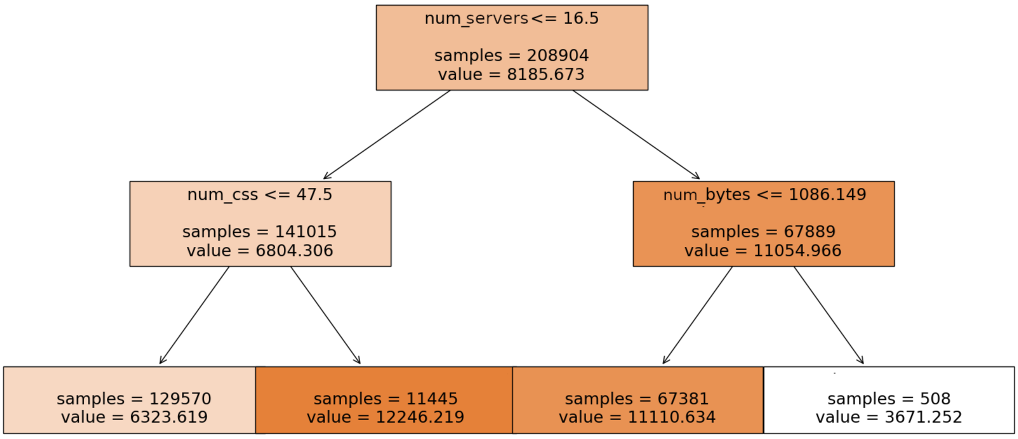

Figure 1. Surrogate decision tree for random forest model

The same models were applied in a per-page fashion and, as was the case in the general analysis, for most pages random forest was the best performing model. Table 4 presents the top three features selected by the random forest model for each page, along with their corresponding Gini importance values.

As observed, number of bytes and number of requests were the features chosen in most cases, with these being prominent in 14 out of the 18 pages due to their high Gini importance value. In 11 out of these 18 pages, the features including number of bytes, number of requests, and one image-related aspect (such as number of image objects, number of distinct images, or number of image pixels) were selected. However, only five pages did not include an image-related feature. Number of distinct images was chosen as frequently as the combined selection of number of image pixels and number of image objects, suggesting that it may better represent the impact on PLT compared to other image-related features.

The analysis was also conducted on a per-category basis, yielding the results presented in Table 5. Interestingly, while number of servers was among the least important features in the per-page analysis, it emerged as the most significant feature

Table 4. Features selected by random forest for each page (Gini Importance value)

web page	bytes	servers	imgs.	JS	CSS	requests	dist. imgs.	pixels
reddit	0.110	—	—	—	—	0.187	**0.513**	—
magazineluiza	**0.395**	—	0.236	—	—	0.121	—	—
pornhub	**0.707**	—	—	—	—	0.049	0.166	—
globo	0.184	—	—	—	—	**0.490**	—	0.095
shopee	0.139	—	0.072	—	—	**0.659**	—	—
olx	**0.661**	0.176	—	—	—	—	0.039	—
twitch	0.361	—	—	0.090	—	**0.364**	—	—
xvideos	**0.345**	—	—	0.168	—	0.151	—	—
letras	0.184	—	—	**0.458**	—	—	0.168	—
americanas	0.155	—	—	—	—	0.208	—	**0.287**
caixa	**0.773**	—	—	—	—	0.098	—	0.037
tiktok	—	—	—	**0.251**	0.194	0.141	—	—
amazon	—	—	0.170	—	—	0.147	**0.343**	—
mercadolivre	**0.394**	—	—	—	—	0.172	0.156	—
123movies	0.245	—	—	—	—	**0.468**	0.110	—
uol	**0.358**	—	—	—	—	0.126	—	0.180
spankbang	**0.619**	0.092	—	—	—	0.156	—	—
twitter	**0.580**	—	—	0.059	—	0.267	—	—

in the e-commerce category. This suggests its crucial role in determining PLT for e-commerce pages, though it appears to be less relevant for individual page differences. The fact that over a third of the measurements originate from e-commerce pages explains why number of servers was identified as the most important feature in the overall analysis. Analogously, number of CSS objects was the key feature in the social media category, while number of JS objects was the most important feature in the news category. Consistent with the per-page analysis, number of bytes, number of requests, and number of distinct images continued to be the three most important features across these categories.

Table 5. Features selected by random forest for each page category

category	bytes	servers	imgs.	JS	CSS	requests	dist. imgs.	pixels
e-commerce	0.145	**0.606**	—	—	—	—	0.114	—
adult content	**0.843**	—	—	—	—	0.055	0.037	—
social media	0.074	—	—	—	**0.737**	—	0.062	—
news	0.162	—	—	**0.477**	—	0.113	—	—
streaming	0.205	—	—	—	—	**0.601**	0.057	—
government	**0.773**	—	—	—	—	0.098	—	0.037

4.2. Unsupervised Analysis

Analyzing how page complexity metrics affect PLT across different categories proved useful, as demonstrated in Section 4.1. However, a potentially more insightful analysis could involve examining pages that share similar complexity metrics. This approach may reveal some nuances or patterns that are not apparent when comparing across broader categories. To explore this possibility, we propose an unsupervised method that groups pages based on their complexity metrics.

As a first step, we modeled the data as a three-way tensor. The first mode represented the "page-raspberry" combination, encompassing all possible pairings of pages and Raspberry-Pis. The second mode was dedicated to complexity metrics, and the third mode to the hour of the day. For this last mode, we calculated the average value of each complexity metric for every page-raspberry pair for each hour, which formed the tensor's values. We then normalized the data using standard scaling. Following this, a PARAFAC algorithm was employed. To determine the tensor's rank, split-half validation was performed and the total explained variance of each tested rank value was calculated. A rank value of 5 was selected, as the rank-5 tensor explained over 90% of the total variance in the data.

Figure 2 presents the factor matrix for the page complexity metrics mode. The values in the matrix are called loadings, which are a way to measure how much each complexity metric contributes to each factor, allowing for a representation of the metrics in a low rank space. Considering the first three factors, all of the complexity metrics had loadings higher than 0.45. This high value suggests that every complexity metric significantly influenced at least one of these factors. Additionally, these factors are arranged in order, wit the first factor explaining the most about the variations in the data, the second factor explaining the next most, and so on. Since the first three factors had high loadings for all the metrics, it means they were the most significant in terms of explaining differences in the data.

Figure 3 displays the page-raspberry factor matrix. For clarity, only five page-raspberry combinations are depicted. As observed, the loading values for each page-raspberry pair are similar for a given factor. This similarity suggests that the patterns of page complexity metrics are consistent across different page-raspberry pairs. Given that these Raspberry-Pis were connected to three distinct ISPs, this finding implies that network variations do not significantly influence these metrics, a conclusion that aligns with observations in other studies [Huet et al. 2021]. The first factor primarily accounts for the variation observed in the complexity metrics of amazon.com.br, as evidenced by the high positive loading values for the metrics: number of bytes, number of JavaScript objects, and number of distinct images (Figure 2). In contrast, caixa.gov.br showed negative loadings for the second factor, while TikTok had positive loadings for the same factor. The second factor was mostly associated with a high number of servers, distinct images and image objects. As expected, TikTok showed positive loadings for these metrics. Conversely, Caixa, being a government website, tended to have fewer images and servers.

The data points were then clustered using the loadings from the complexity metrics mode. K-means clustering was performed with $K = i, i \in [2, 10]$. The silhouette scores were then computed for each value of K, with the elbow method being used to choose the final number of clusters to continue the analysis, which yielded a value of 4.

Figure 4 displays the distribution of navigations for each web page, categorized by the assigned cluster. Note that almost half of all the pages had all their navigations mapped exclusively to a single cluster. Additionally, every page had at least 60% of its navigations

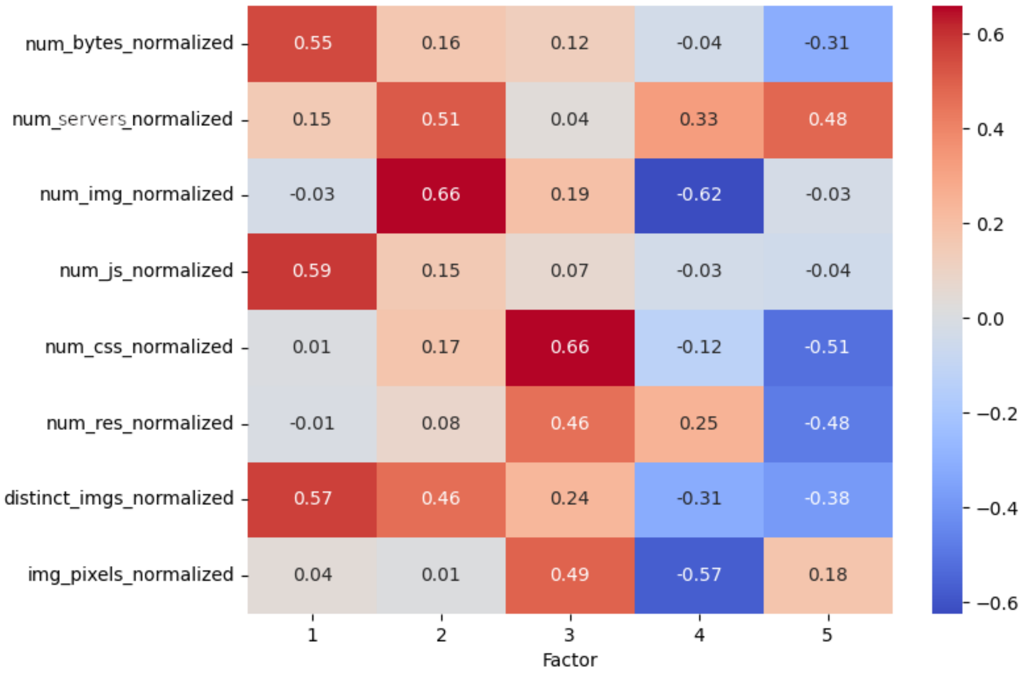

Figure 2. Factor matrix for complexity metrics mode

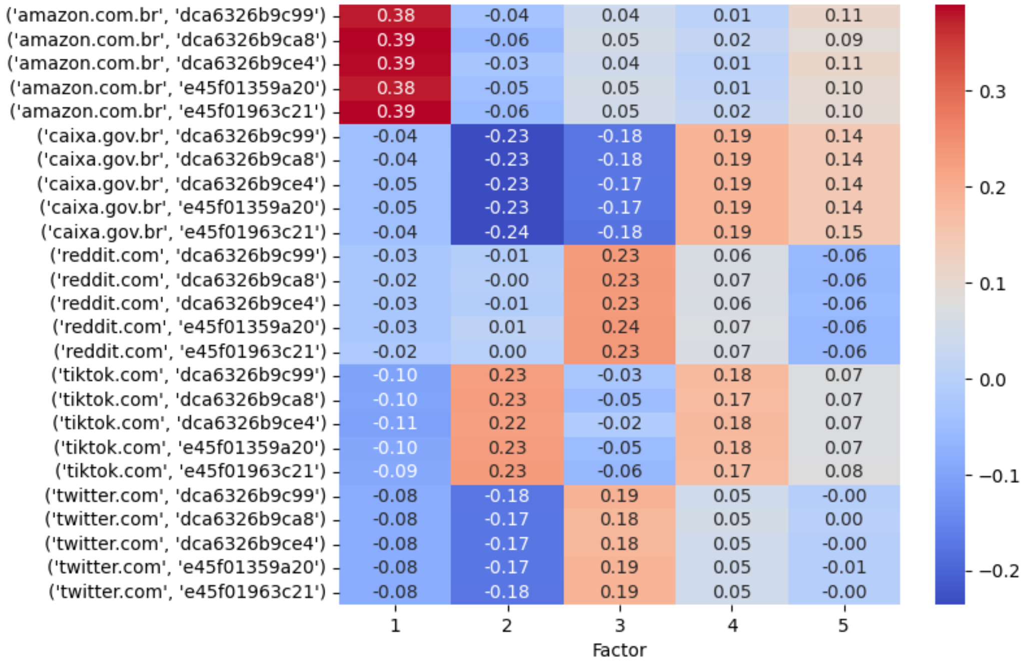

Figure 3. Abbreviated factor matrix for page-raspberry mode

mapped to one predominant cluster. A notable correlation exists between the category of a web page and its corresponding cluster. Navigations to web pages featuring adult content were exclusively mapped to cluster 0, while cluster 1 predominantly grouped news and e-commerce web pages. Cluster 2 mainly grouped e-commerce and social media pages. Notably, cluster 3 was unique in containing only navigations to amazon.com.br. (Note that not all Raspberry Pis collected data from the Pornhub page, thus precluding its evaluation via unsupervised analysis.)

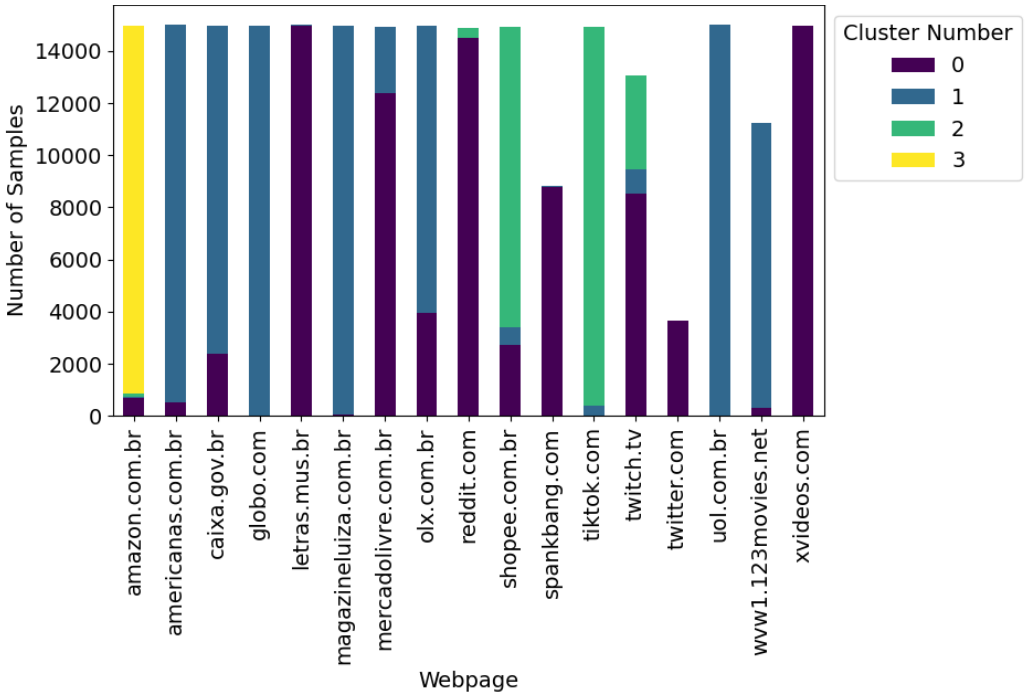

Figure 4. Cluster distribution per web page

A random forest analysis was applied to each of the clusters, calculating the Gini importance for each feature. The results are presented in Table 6. Similar to the per-category analysis, the number of bytes is one of the top three features for all clusters. Notably, the number of requests was among the top three features only in cluster 3, which exclusively comprises navigation instances from amazon.com.br. As expected, cluster 1, predominantly consisting of e-commerce and news pages, identified the number of bytes, number of servers and number of JavaScript objects as the three top features. Recall that in Section 4.1 the number of bytes and number of servers were the two most important features for the e-commerce category, and the number of bytes and number of JavaScript objects are the two most important features for the news categories.

Table 6. Features selected by random forest for each cluster

cluster	bytes	servers	imgs.	JS	CSS	requests	dist. imgs.	pixels
0	**0.287**	—	—	—	0.160	—	0.111	—
1	**0.658**	0.063	—	0.131	—	—	—	—
2	0.069	**0.743**	—	—	0.063	—	—	—
3	0.139	—	—	—	—	**0.189**	0.183	—

To compare the effectiveness of the different models, the general model, the category model, and the cluster model were tested on each page individually. For cluster models, the assigned cluster for each page was the one containing over 60% of that page's navigations. All data points were normalized using standard scaling, considering the entire dataset, before fitting the models.

The models were adapted to train on all the pages except the one being analyzed. Taking the Shopee page as an example, the general model used all pages except Shopee for training, selecting the top three features. This model was then exclusively tested on Shopee data. In the category model, an e-commerce model was trained (since Shopee is an e-commerce site) using all e-commerce pages except Shopee, and then tested only on Shopee data. For the cluster model, we allocated Shopee to cluster 2, as it encompasses most of its navigations. This model was trained using all data from cluster 2, excluding Shopee, and tested on all Shopee data, including navigations on the Shopee page that did not belong to cluster 2. The features were selected for each page/model combination, according to the training data, which, in certain cases resulted in different feature sets from those previously reported.

Table 7. RMSE per page per model

page	category	cluster	category model	cluster model	general model
magazineluiza	e-commerce	1	2.488	1.888	**0.537**
shopee	e-commerce	2	**0.841**	1.351	1.242
olx	e-commerce	1	0.870	**0.752**	0.813
americanas	e-commerce	1	**0.718**	1.219	0.912
amazon [1]	e-commerce	3	**0.992**	—	1.369
mercadolivre	e-commerce	0	**1.342**	2.140	1.799
reddit	social media	0	**1.173**	1.344	1.641
tiktok	social media	2	**1.074**	1.429	1.734
twitter	social media	0	0.896	**0.811**	1.133
globo	news	1	**0.669**	0.868	0.766
uol	news	1	**0.717**	0.831	0.942
xvideos	adult	0	3.486	2.920	**1.207**
spankbang	adult	0	**0.420**	0.698	0.468
pornhub	adult	—	**0.283**	—	0.324
twitch	streaming	0	1.014	**0.792**	0.982
123movies	streaming	1	**0.698**	1.181	1.416
caixa [2]	government	1	—	**0.578**	1.695
letras [2]	music	0	—	**0.829**	0.954

Table 7 shows the RMSE for each page. For ten out of the 15 pages, where the category and the cluster models were evaluated, the category model outperformed both the cluster and general models. This highlights the significance of incorporating page categories in analyzing PLT. Note that, in most models, the RMSE value suggests an error in estimating the PLT of less than one standard deviation. In future research, a larger number of pages will be analyzed to gain a more comprehensive understanding of category-specific trends.

5. Conclusion

In this work, we examined the relationship between page complexity metrics and page load time using both supervised and unsupervised models. The Random Forest model

[1] There are no results from the cluster model because cluster 3 comprises solely of Amazon data.

[2] The category model results are absent for caixa and letras, as these were the only websites in the government and music categories, respectively.

demonstrated the best performance among the supervised models. For the unsupervised analysis, a tensor decomposition approach was employed. This method enables the representation of data in a multidimensional space, which facilitates subsequent clustering. This clustering was used to analyze how the complexity metrics affect page load time within each identified cluster.

We trained three Random Forest models: a general model (encompassing all pages), a specific page model, and a page category model, to analyze the most important features for each. The results indicate that the number of bytes is one of the top three features for the majority of models, with the only exceptions being the models for Amazon and TikTok pages.

In the page category models, the number of bytes and the number of requests emerge as the top two features for most categories. However, distinct trends appear depending on the page category. For the e-commerce category, the number of servers is the most crucial feature for determining page load time. In the news category, it is the number of JavaScript objects, likely due to the ads loaded through JavaScript. For the social media category, the number of CSS objects is most significant. These findings from the page category models align with the analysis from the specific page model, where the three most important features were identified as the number of bytes, the number of requests, and the number of distinct images.

We obtained four clusters from the unsupervised analysis. Interestingly, number of bytes is one of the top three features for all clusters, aligning with the findings from the supervised analysis. Each cluster predominantly groups navigations to web pages within specific categories: cluster 0 exclusively contains pages with adult content; cluster 1 primarily includes news and e-commerce web pages; and cluster 2 features a combination of e-commerce and social media pages. Remarkably, cluster 3 is unique, exclusively encompassing navigations to amazon.com.br.

To assess the effectiveness of different models, we individually tested the general model, the page category model, and the cluster model on each page. The page category model outperformed both the cluster and the general models for ten out of the 15 pages. This indicates that page category models are significant in predicting page load time. It should be noted that, in most models, the RMSE value suggests an error in estimating the page load time of less than one standard deviation.

References

Aqeel, W., Chandrasekaran, B., Feldmann, A., and Maggs, B. M. (2020). On landing and internal web pages: The strange case of jekyll and hyde in web performance measurement. *IMC '20: Proceedings of the ACM Internet Measurement Conference*, page 680–695.

Asrese, A. S., Eravuchira, S. J., Bajpai, V., and Sarolahti, P. (2019). Measuring web latency and rendering performance: Method, tools, and longitudinal dataset. *IEEE Transactions on Network and Service Management*, 16:535–549.

Avram, C., Salem, K., and Wong, B. (2014). Latency amplification: Characterizing the impact of web page content on load times. *Lecture Notes in Computer Science*, pages 20–25.

Butkiewicz, M., Madhyastha, H. V., and Sekar, V. (2011). Understanding website complexity: measurements, metrics, and implications. *IMC '11: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, page 313–328.

Egger, S., Reichl, P., Hoßfeld, T., and Schatz, R. (2012). "time is bandwidth"? narrowing the gap between subjective time perception and quality of experience. *2012 IEEE International Conference on Communications*, pages 1325–1330.

Gao, Q., Dey, P., and Ahammad, P. (2017). Perceived performance of top retail webpages in the wild: Insights from large-scale crowdsourcing of above-the-fold qoe. *Internet QoE '17: Proceedings of the Workshop on QoE-based Analysis and Management of Data Communication*, page 13–18.

Gigspaces (July, 2023). Amazon found every 100ms of latency cost them 1% in sales. https://www.gigaspaces.com/blog/amazon-found-every-100ms-of-latency-cost-them-1-in-sales, Accessed: 2024-01-12.

Google (April, 2010). Using site speed in web search ranking. https://developers.google.com/search/blog/2010/04/using-site-speed-in-web-search-ranking, Accessed: 2024-01-12.

Hora, D. D., Asrese, A. S., Christophides, V., Teixeira, R., and Rossi, D. (2018). Narrowing the gap between qos metrics and web qoe using above-the-fold metrics. *PAM 2018 - International Conference on Passive and Active Network Measurement*, pages 1–13.

Hoßfeld, T., Metzger, F., and Rossi, D. (2018). Speed index: Relating the industrial standard for user perceived web performance to web qoe. *2018 Tenth International Conference on Quality of Multimedia Experience*, pages 1–6.

Huet, A., Saverimoutou, A., Houidi, Z. B., Shi, H., Cai, S., Xu, J., and Mathieu, B. (2021). Deployable models for approximating web qoe metrics from encrypted traffic. *IEEE Transactions on Network and Service Management*, 13:3336–3352.

Jahromi, H. Z., Delaney, D. T., and Hines, A. (2018). How crisp is the crease? a subjective study on web browsing perception of above-the-fold. *2020 6th IEEE Conference on Network Softwarization*, pages 43–50.

Molnar, C. (2019). *Interpretable Machine Learning*. Lulu.com.

Salutari, F., Hora, D. D., Dubuc, G., and Rossi, D. (2019). A large-scale study of wikipedia users' quality of experience. *WWW '19: The World Wide Web Conference*, page 3194–3200.

Saverimoutou, A., Mathieu, B., and Vaton, S. (2019). A 6-month analysis of factors impacting web browsing quality for qoe prediction. *Computer Networks*, 164.

Vogel, L. and Springer, T. (2022). An in-depth analysis of web page structure and efficiency with focus on optimization potential for initial page load. *International Conference on Web Engineering*, pages 101–116.

Wiegand (April, 2022). Site speed is (still) impacting your conversion rate. https://www.portent.com/blog/analytics/research-site-speed-hurting-everyones-revenue.htm, Accessed: 2024-01-12.