Predictive Modeling of Time Spent on Website with Linear Regression and K-means

Student paper

Veljko Lončarević*, Savo Šućurović**
Second cycle students
University of Kragujevac, Faculty of Technical Sciences
Čačak, Serbia
Email: *veljkoloncarevicharry@gmail.com, **savos977@gmail.com

Abstract— **This study investigates the impact of tailoring linear regression models to segmented data subsets using the K-means clustering algorithm, with a specific focus on predicting time spent on a website. While not every subset model displayed a statistically significant reduction in mean absolute error (MAE), the collective results underscore the practical significance of this approach. The average MAE across the subset models is 12.676, reflecting a substantial decrease of 19.36%, revealing the efficacy of cluster-specific feature customization. Distinct features removed during Lasso regularization in individual subsets highlight the importance of tailored model development.**

Keywords - Linear regression, K-means clustering, Lasso regularization, Cluster-specific modeling

I. INTRODUCTION

Websites and online platforms monitor a variety of metrics in order to assess their performance, user engagement, satisfaction and experience, content effectiveness and relevance. Websites often have specific goals, such as capitalizing on conversion opportunities or increasing advertising revenue. These metrics can also help with identifying issues like usability problems, poor content or technical glitches that need to be addressed. Various search engine services may consider user engagement metrics as a factor in determining the relevance and quality of a website's content. One of these metrics is time spent on website, which is being widely used in various types of websites, of which most notable are content-based websites and blogs, news and media websites, educational platforms, e-commerce websites, and social media platforms. The likelihood of making a purchase on an e-commerce platform peaks when an individual spends approximately 50 seconds on the product page [1]. In the context of e-commerce platforms, the duration dedicated to reviewing product-related information, alongside metrics like bounce rates, exit rates, and customer type, holds substantial sway over a customer's decision to make a purchase on the website [2].

With the rise of machine learning algorithms over the past decade, the demand for accurate predictions of these metrics is also growing. These algorithms play a pivotal role in uncovering patterns and trends within large and complex datasets, which can be used to understand and forecast various relevant metrics. For instance, linear regression models are commonly used to predict continuous values by analyzing the relationship between input features and the target variable through training on historical data. The trained model in turn learns to capture the underlying dependencies within the dataset and uses them in order to generalize and predict the outcome on new, heretofore unseen input data. Datasets used to train these models usually have a large number of data points, with the counts in millions, or even billions, while at the same time also having a large number of connected features – datasets which are usually complex and heterogeneous.

When a need arises for grouping unlabeled data into a variable number of groups based on certain features or characteristics, it is common to use clustering algorithms. Clustering algorithms are a type of unsupervised machine learning techniques, whose main purpose is to discover inherent patterns, structures, or relationships within a dataset, when there are no predefined labels or categories. One of the most popular clustering algorithms is K-means, which partitions the data into K clusters by minimizing the sum of squared distances between data points and the centroid of their assigned cluster. In this research paper, we prepare a dataset and train a linear regression model on it, while measuring impact of different features and performance metrics. Afterwards, a clustering algorithm is applied to a dataset in order to segment it into smaller, more meaningful clusters. After dividing the dataset into smaller clusters, a linear model is trained on each of the clusters separately, while performing feature impact analysis, measuring performance metrics, and comparing them to the original model. Our aim is to investigate whether the impact of features on model performance varies across these clusters and, consequently, whether certain features should be selectively retained or eliminated in order to significantly reduce MAE in estimating the total time spent on the site.

II. THEORETICAL FOUNDATIONS

A. Linear Regression

Linear regression is a statistical method used to model the relationship between one or more independent variables (called features) and a dependent variable (i.e., target variable). It is accomplished by fitting a linear equation to the observed data points. The purpose of linear regression is to find the best-fitting hyperplane, or a single line (in the case of one, single feature), by minimizing the sum of the squared differences between the observed values and the values predicted by the linear model. The standard form of a linear regression equation is as follows:

$$y = b_0 + b_1 x_1 + \cdots + b_n x_n \quad (1)$$

where y is the dependent variable, $x_1, \ldots, x_n$ are the independent variables, b_0 represents the intersection with the y axis (i.e., y-intercept), and $b_1, \ldots, b_n$ are coefficients. The linear regression model is typically fitted to the data using a method called the method of least squares. This method aims to minimize the sum of the squared differences between the observed values and the values predicted by the linear model [3]. The coefficients are estimated in such a way that the sum of these squared differences is minimized. Least squares tries to minimize the output of an objective function, which is calculated as the sum of the squared differences between the observed values and the values predicted by the linear model. This method may not always result in an optimal line being fitted to the dataset, as it is sensitive to outliers. Outliers are data points that deviate significantly from the general pattern of the rest of the data. Since least squares minimizes the sum of squared differences, outliers with large deviations can disproportionately influence the resulting model. This sensitivity to outliers can lead to a line which is skewed towards the extreme values.

In order to account for the possibility of outliers and minimize their impact, different regularization techniques are utilized. One such technique is called L1 regularization, also known as Lasso regularization. The L1 regularization (penalty) term is added to the least squares objective function, helping to mitigate the effect of the outliers on the model. While the standard least squares objective function is the difference between observed and predicted values, Lasso regularization objective function adds the following regularization term:

$$\lambda \sum_{i=1}^{p} |b_i| \quad (2)$$

where λ is the regularization parameter, which controls the strength of the regularization, and $|b_i|$ is the absolute value of the coefficient [4]. The regularization parameter λ is also a hyperparameter, which is a parameter not derived from data, but set prior to the training of the model. The minimization of the objective function seeks to find coefficient values that not only minimize the sum of squared differences between observed and predicted values, but also minimize the sum of the absolute values of the coefficients. Some coefficients become exactly zero, effectively excluding certain features from the model. This can be beneficial when dealing with a large number of features, or when there is a suspicion that some features are irrelevant.

MAE is a metric used to evaluate the accuracy of a predictive model, including linear regression. It measures the average absolute differences between the observed actual values and the values predicted by the model. It is calculated in the following way:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (3)$$

MAE provides a straightforward interpretation, the average absolute error across all predictions. Its main advantage is treating all errors equally, unlike other metrics like Mean Squared Error, which can give more weight to larger errors [5]. A metric usually calculated with MAE is Mean Absolute Percentage Error (MAPE), which provides a percentage measure of the average absolute difference between predicted and actual values relative to the actual values. Its formula is:

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100 \quad (4)$$

In addition to MAE and MAPE, a common resampling technique called Cross-validation is applied to assess the performance of the model. The basic idea behind cross-validation is to divide the dataset into multiple subsets, train the model on some of these subsets, and evaluate its performance on the remaining subsets. This process is repeated multiple times, and the results are averaged to obtain an estimate of the model's performance.

B. K-means Clustering

K-means clustering is an unsupervised machine learning algorithm used to partition a dataset into K non-overlapping, distinct subsets, where K is a user-defined number. It groups similar data points based on certain features. The algorithm aims to minimize the within-cluster sum of squared distances, meaning it tries to create clusters in such a way that the data points within a cluster are close to each other. The algorithm chooses K initial cluster centroids by randomly selecting them, although there are more sophisticated methods available like K-means++, which ensures that the initial centroids are well spread out across the dataset [6]. Each data point is assigned to the cluster whose centroid is closest to it, using Euclidean distance. The centroids are afterwards updated, by recalculating them based on the mean of the data points assigned to each cluster. This process is repeated until convergence occurs - when the assignments no longer change significantly, or after a set number of iterations.

However, K-means clustering algorithm has certain drawbacks which should be noted – mainly, it assumes that clusters are spherical, of equal size and density, which might not be the case for most datasets. K-means is also sensitive to the choice of initial centroids, because the algorithm converges to the local minimum, and different initial centroids can lead to different cluster assignments. K-means also require its user to define hyperparameter K before fitting the dataset, which can lead to underfitting (too few clusters, the existing clusters simplify the structure of the dataset) or overfitting (too many clusters, algorithm fits too closely to the data, finds patterns that might not be there), therefore finding an optimal number of clusters is crucial for obtaining meaningful results. One of the most commonly used techniques used to estimate the optimal

number of clusters is the Elbow method. It involves running K-means for a range of K values and plotting the sum of squared distances (inertia) against K. The point where the rate of decrease in inertia sharply changes (forming an "elbow" in the plot) is usually considered the optimal K [7].

III. METHODOLOGY

After analyzing the motives and goals of this research, the following hypothesis was formed:

H: *When trying to predict time spent on a website from a dataset based on user usage data, clustering complex heterogeneous datasets and tailoring features to each cluster individually can significantly reduce MAE of linear regression models.*

A dataset based on metrics extracted from web server logs [8] has been collected during the period from September of 2022 to August 2023, and includes 30 columns and approximately 100,000 rows of data. The target variable is time_spent_on_website (measured in seconds), while the features are shown in Table I. Geodata like city, country, etc. has been obtained by using an IP Geolocation service [9], while data on whether an IP address belongs to a VPN service has been obtained from a VPN detection service [10].

TABLE I. LIST OF COLLECTED FEATURES

#	Feature	#	Feature
1.	session_ID	16.	uses_vpn
2.	page_viewed	17.	uses_proxy
3.	timestamp	18.	previous_visits
4.	browser	19.	device_resolution
5.	IP_address	20.	device_type
6.	referred_by	21.	device_brand
7.	interaction_type	22.	operating_system
8.	conversion_occured	23.	touch_screen_device
9.	exit_page	24.	time_between_sessions
10.	location	25.	connection_type
11.	city	26.	browser_version
12.	zip_code	27.	language
13.	country	28.	cookies_enabled
14.	longitude	29.	engagement_level
15.	latitude		

An example of the first two instances from the dataset (with identifying information excluded for privacy reasons – IP address, session ID, location, city, latitude, longitude, zip code) is shown in Table II.

The dataset used in this research is not publicly available due to privacy considerations and restrictions. The data used for this study may contain sensitive or personally identifiable information, and the authors are committed to protecting the privacy and confidentiality of the individuals or entities associated with the data. As a result, the dataset is not openly shared or accessible. We recognize the importance of promoting transparency and reproducibility in scientific research. However, ethical and legal constraints prohibit the public release of the specific dataset used in this study. For inquiries regarding the dataset or access to related aggregated results, please contact Veljko Lončarević at veljkoloncarevicharry@gmail.com. The authors are committed to addressing any reasonable requests for information that do not compromise the privacy and confidentiality of the individuals or entities involved.

TABLE II. FIRST TWO INSTANCES FROM THE DATASET

Feature	1st instance	2nd instance
page_viewed	item	home
timestamp	21-08-23 12:44	15-02-23 16:49
browser	chrome	safari
referred_by	nan	nan
interaction_type	share	view
conversion_occured	TRUE	FALSE
exit_page	home	nan
country	Serbia	Netherlands
uses_vpn	FALSE	TRUE
uses_proxy	FALSE	FALSE
previous_visits	11	0
device_resolution	1440x2560	750x1334
device_type	smartphone	smartphone
device_brand	Samsung	iPhone
operating_system	android	iOS
touch_screen_device	TRUE	TRUE
time_between_sessions	37	nan
connection_type	5G	4G
browser_version	116	16
language	SR	SR
cookies_enabled	TRUE	TRUE

The experiment is divided into three parts:

1. Training a linear regression model on the entire dataset.
2. Clustering the dataset using K-means into smaller clusters.
3. Training smaller linear models on each of the clusters individually.

The first and the third part, which consist of creating linear regression models, can be additionally divided into phases as shown on Fig. 1.

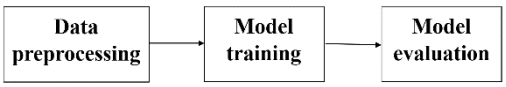

Figure 1. Phases of creating linear regression models

Data preprocessing consists of analyzing each feature and transforming it into a form which is suitable for usage in a linear regression model. First, the amount of missing (null) values was calculated for each feature. In the case of small amounts (less than one hundred values missing) the rows were removed entirely. In other cases, where there was a higher number of missing values, they were swapped with the mean value of the entire feature in the case of numerical features, or with constant "None" in the case of categorical features. Categorical features with smaller numbers (≤5) of unique values were encoded using the One-Hot encoding method, whereas other categorical features were encoded using Target encoding. Numerical values were scaled using a MinMax scaler, which scales the data between 0 and 1, where the smallest data point gets assigned the value of 0, and the largest value of 1, while the scale between data points remains the same. The dataset was divided into five subsets using the train-test split method, and a linear regression model was cross-validated using the training data from each split. Lasso (L1) regularization was employed during the model training. L1 regularization parameter λ has been derived using the Grid Search method. After cross-validation for each subset, MAE and MAPE were calculated, and the coefficients of each feature were recorded.

Subsequently, the dataset underwent K-means clustering with the number of clusters (K) determined using the elbow method. K subsets were extracted from the resulting clusters. Each subset was further split into five train-test splits, and the model cross-validation process was repeated. A new linear regression model was trained on each cluster using Lasso regularization, after which MAE and MAPE, along with the feature coefficients were documented. These values were then compared with those of the original model.

In this study, data processing and preparation were carried out using the Python programming language, leveraging the capabilities of the widely used libraries NumPy and Pandas. The application of machine learning models and their subsequent evaluations (MAE, MAPE calculations and feature impact disparity evaluation) were conducted using the scikit-learn library, which provided a comprehensive suite of tools for classification and regression tasks. The assessment of optimal cluster numbers using the elbow method was visualized using the Matplotlib library.

IV. Results and Discussion

A. Results

After performing Grid Search method in order to find the optimal Lasso regularization parameter on a range of numbers, the best performing value was found to be equal to 0.3, therefore it was set as $\lambda = 0.3$. The optimal number of clusters $K = 5$ was found using the elbow method, as shown in Fig. 2.

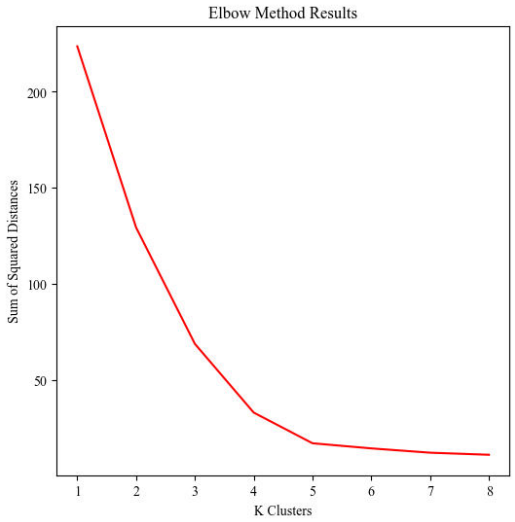

Figure 2. Results of the Elbow method

After training the original linear model on the entire dataset (labeled as OLM), applying the Lasso regularization technique, retrieving coefficients (both before and after applying Lasso) and calculating MAE and MAPE, resulting values have been documented. The same steps have been repeated for every one of the five subsets of the dataset extracted from the clusters after applying K-means clustering (labeled as SLM_1, …, SLM_5). However, since after applying One-Hot encoding to certain features, the resulting dataset had 57 columns, and for that reason only the features whose coefficients were set to 0 (effectively removed) after Lasso regularization will be shown in this paper. The values documented are shown in Table III.

TABLE III. MAE AND REMOVED FEATURES PER LINEAR MODEL

Model	MAE	MAE % Decrease	MAPE	Removed features
OLM	15.72	/	23.86%	touch_screen_device, referral
SLM_1	11.44	27.23%	13.04%	touch_screen_device, exit_page, session_id
SLM_2	15.39	2.10%	21.76%	touch_screen_device, previous_visits, session_id, uses_proxy
SLM_3	12.20	22.39%	16.95%	exit_page, timestamp, session_id
SLM_4	12.53	20.29%	17.12%	browser_version, language, uses_proxy
SLM_5	11.82	24.81%	15.66%	touch_screen_device, exit_page, interaction, cookies_enabled
Average (SLM)	12.676	19.36%	16.91%	/

B. Discussion

While not every linear model trained on segmented data subsets has exhibited a significant reduction in MAE, it is noteworthy that the average MAE across the subset models is 12.676. This represents a notable decrease of 19.36%, a

magnitude that can be characterized as substantial within the context of this analysis. Additionally, while there are common features which were removed after Lasso regularization in most models (touch_screen_device, exit_page, session_id), certain features are distinctive to individual subset models. From these two observations, it can be inferred that customizing features for each cluster distinctly contributes to a substantial reduction in the MAE of linear regression models, particularly in the domain of predicting time spent on the website. Consequently, the hypothesis is deemed confirmed.

C. Comparison with Related Research

In [11], the authors employ a Self-Organizing Maps and Long Short-Term Memory (SOM-LSTM) algorithm to predict the remaining time on an e-commerce website for users. Notably, their model achieves a lower mean value for MAPE across all clusters, specifically measuring at 13.635%. In comparison, our research presents a MAPE of 16.91%. While the SOM-LSTM algorithm in the referenced paper demonstrates superior accuracy in predicting user behavior, it is essential to acknowledge that model selection should be context-dependent. Our research, utilizing a different approach, may offer advantages in instances where interpretability or simplicity is prioritized over nuanced pattern recognition. Linear regression models, such as the one used in this study, are often more interpretable, simple and computationally efficient, making them advantageous in scenarios where a balance between accuracy and model complexity is required. Similarly, in [12], a notable reduction of 14.6% in MAE was achieved through the proposed neural network approach when compared to the baseline linear regression model, underscoring the efficacy of advanced modeling techniques in enhancing predictive accuracy. It is noteworthy that unlike our research, the cited study did not involve clustering, emphasizing the versatility of improved model architectures across various predictive modeling scenarios.

Compared to other research, this paper makes a contribution by demonstrating the implications of clustering on the accuracy of predictions, particularly in the context of user engagement on a website, and offers valuable insights for researchers engaged in predictive modeling tasks; furthermore, it distinguishes itself by proposing a simple solution, addressing the balance between accuracy and model complexity.

V. Conclusion

In conclusion, the findings of this study shed light on the efficacy of tailoring linear regression models to segmented data subsets, particularly in the prediction of time spent on a website. While not every model demonstrated a statistically significant reduction in MAE, the collective impact across the subset models is noteworthy. The average MAE, standing at 12.676, reflects a substantial decrease of 19.36%, underscoring the practical significance of this approach within the scope of our analysis. Certain features removed by Lasso regularization are distinct to individual subset models, suggesting a nuanced and tailored approach to feature selection. This observation implies that customizing features for each cluster significantly contributes to the observed reduction in MAE, emphasizing the importance of considering unique cluster characteristics in model development.

The implications of this study extend beyond the immediate context, offering valuable insights for practitioners and researchers alike. The demonstrated effectiveness of feature customization within clustered subsets underscores the importance of recognizing heterogeneity in datasets. By tailoring models to specific subsets, practitioners can enhance predictive accuracy, especially in scenarios where linear regression is employed for time-related predictions, such as website engagement.

This study contributes to the existing body of knowledge by providing empirical evidence of the impact of cluster-specific feature customization on linear regression models. The identification of both common and distinct features across subsets adds granularity to the understanding of feature relevance, offering practitioners a nuanced perspective in their model development endeavors.

Building on these findings, future research endeavors could explore additional factors influencing the effectiveness of subset-specific model customization. Investigating the interplay between cluster characteristics, data distribution, and the choice of regularization techniques could yield further insights. Additionally, extending this approach to different prediction domains or incorporating advanced machine learning models may broaden our understanding of the generalizability of these findings.

References

[1] X. Niu, C. Li, and X. Yu, "Predictive analytics of e-commerce search behavior for conversion," *AIS Electronic Library (AISeL)*, Aug. 10, 2017.

[2] S.-C. Necula, "Exploring the impact of time spent reading product information on e-commerce websites: A machine learning approach to analyze consumer behavior," *Behavioral Sciences*, vol. 13, no. 6, May 23, 2023.

[3] IA. K. Kuchibhotla, L. D. Brown, and A. Buja, "Model-free study of ordinary least squares linear regression," *arXiv:1809.10538 [math.ST]*, Available: https://arxiv.org/abs/1809.10538, Sep. 27, 2018.

[4] J. Ranstam and J. A. Cook, "LASSO regression," *British Journal of Surgery*, vol. 105, no. 10, pp. 1348, Sep. 2018.

[5] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, pp. 79–82, Dec. 19, 2005.

[6] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07),* January 2007, pp. 1027–1035.

[7] E. Umargono, J. E. Suseno, and S. K. V. Gunawan, "K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula," in *Proceedings of The 2nd International Seminar on Science and Technology (ISSTEC 2019),* January 2020.

[8] Veljko Lončarević's personal website, available at: https://www.veljkoloncarevic.in.rs. Last access to the site: 30th November 2023.

[9] IP Geolocation Service, available at: https://ipgeolocation.io/ Last access to the site: 30th November 2023.

[10] VPN Detection Service, available at: https://www.ipqualityscore.com/ Last access to the site: 30th November 2023.

[11] L.-J. Kao, C.-C. Chiu, H.-J. Wang, and C. Y. Ko, "Prediction of remaining time on site for e-commerce users: A SOM and long short-term memory study," *Journal of Forecasting*, vol. 40, no. 7, pp. 1274-1290

[12] S. Gupta and S. Maji, "Predicting Session Length for Product Search on E-commerce Platform," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020