IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A COMPARATIVE ANALYSIS ON WEB-METRICS EVALUATION MODELS USING MACHINE LEARNING

Susmit Sekhar Bhakta, Deep Narayan Chaudhuri, Ayon Roy, Sana Tasneem, Ananya Sarkar

Student, Student, Student, Student, Student
Computer Science and Engineering(CSE)
Techno International Newtown, Kolkata, India

Abstract: In the days of the Internet, millions of users are using billions of websites in every fraction of seconds. The performance of each and every website depends upon some crucial factors. These factors are called web-metrics. It is crucial to evaluate them for better performance, security etc. These web-metrics consist of web-page load time, number of redirections, total number of pages, First Contentful Paint(FCP), Largest Contentful paint(LCP), Speed Index, DOM size, QLS etc. By evaluating these features we can understand how a website will perform during user usage. Also by evaluating them we can make necessary changes to improve the performance of the website. A huge amount of study is already going on by various researchers. In our paper we made a comparative study by using machine learning models like linear regression, SVM, Random Forest. Also we proposed a model by imposing an ensemble learning method with ML models which can predict the website's load time. Our model can also be able to give information about other web-metrics by evaluating this predicted web-page load time and the corresponding website URL. Our proposed model is directly trained by extracting web-metrics from the input URL. This hybrid model can be used by developers, Administrators to optimize and improve the performance of the corresponding website. Our proposed model is achieving a highly accurate(98%) predicted load time for websites and can successfully predict the other parameters of web-metrics. Also the proposed model is highly flexible to add or remove other web-metrics.

Index Terms **- Internet, web-metrics, FCP, LCP, speed index, DOM size, OLS, machine learning, linear regression, SVM, random forest, ensemble learning, URL.**

I. INTRODUCTION

Web-metrics evaluation has become very crucial for website admins, because poor website performance may lead to the negative user reactions. This can badly affect the website reports which includes higher website bounce rates, less user engagement which results in decreased conversion rates. So, getting a positive user experience is very much important to stay in this comparative market. To get a positive user experience it is necessary to make a website which can perform well in some constraints also. The website performance directly lies upon web-page load time, FCP, LCP, Speed Index, DOM size etc. From that the webpage load time is most critical as high load-time may distract the user from getting the information within time.

To resolve this issue by evaluating various web-metrics a lot of researchers are going on. Machine learning models are being used to predict web-page load times. Linear regression was used to predict page load time based on several web-metrics like page size, number of requests and content type[1]. Also by using SVMs it is possible to predict the load time by evaluating page size, number of requests, and server response time[4]. The Random Forest algorithm is very much effective to predict average load time[2].

Meanwhile researchers had found that hybrid modeling can be effectively able to predict load times with higher accuracy. By extracting web-metrics, random forest and neural networks can be able to give higher accuracy in prediction[3]. But there are some more percentages to improvements. There are also a variety of web-metrics present today.

In this paper, we proposed a hybrid model consisting of Linear regression, support vector machine(SVM) and random forest. We have used ensemble learning techniques(Voting classifier) to merge them. Our model can effectively predict load time by extracting a lot of web-metrics like number of headings, total page size, redirection number, videos number etc and can be able to give FCP, LCP, OLS, DOM size etc. In our paper we have shown a comparative analysis on individual models and ensemble models and achieved highest accuracy in our proposed model. The claim of highest accuracy is based on compared to other models we have shown and comparing the output values with the Google Analytics report of the corresponding website. Please note that a website load may vary time-to-time as more contents are adding or deleting or improvements are going on.

II. FEATURE SELECTION AND DATA SOURCE

For this study, we have selected web-metrics like number of pages, total page size, total redirects, totals videos present etc. These features are directly linked to the page load time of a website. The data source is a user input URL(should be valid).

III. THEORETICAL FRAMEWORK

In this paper we have various machine learning models like Linear Regression, Support Vector Machine(SVM), Random Forest and ensemble learning technique which are discussed below →

3.1. ***Linear Regression***– This prediction model is very much famous among researchers for its simplicity and flexibility. It is basically a statistical method used to make predictions on various fields like economics, Science and technology, medical science etc. This model works on a method of relationship between a dependent variable and one or more independent variables. The most simplest form of Linear regression is simple linear regression. The simple linear regression model can defined as,

$$P = \beta_0 + \beta_1 I + \epsilon \ldots\ldots\ldots\ldots (1)$$

In Equation1, P is the dependent variable, I is the independent variable, β_0 is the p-intercept of the line, β_1 is the slope of the corresponding line and ϵ is the error term. So, the Simple Linear regression model predicts a continuous dependent variable P based on a single independent variable I. The main purpose of the model is to find the best-fit line which is a straight line that minimizes the sum of the squared differences between the actual and predicted values of P. If there is a presence of multiple number of independent variable then the formula becomes,

$$P = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \cdots + \beta_n I_n + \epsilon \ldots\ldots\ldots\ldots\ldots (2)$$

In Equation2, $I_1, I_2, \cdots, I_n$ are the independent variables and their corresponding coefficients are $\beta_1, \beta_2, \cdots, \beta_n$.

3.2. ***Support Vector Machine(SVM)*** – This is a supervised learning model which can be used for both classification and Regression problems. It is also a well known model for prediction purposes. This model is able to directly learn a decision boundary which separates the different classes in the dataset. The decision boundary is known as Hyperplane. This algorithm attempts to maximize the margin i.e. the distance between the hyperplane and the closest data points from each class. The closest points to the hyperplane are called support vectors. The SVM can be defined as,

$$f(x) = sign(\sum_{m=1}^{n} \alpha_m y_m K(x_m, x) + A) \ldots\ldots\ldots\ldots\ldots (3)$$

In Equation3, x is the input vector, $y_m$is the class label of the jth training set, α_m are the Lagrange multipliers, K(x_m,x) is the kernel function which measures the similarity between $x_m$and x in the feature space, A is the bias term, $f(x)$ is the bias function.
The sign function returns the sign of the sum of the kernel evaluations plus the bias term, which determines the predicted class label of the input vector x.

3.3. ***Random Forest***– This is also a famous machine learning model which can be used for both classification and regression problems. This ensemble learning method is used for improving accuracy and reducing overfitting by combining multiple decision trees.
The algorithm constructs a forest of decision trees by extracting random subsets of available features and a random subset of the training data from the training dataset. For predicting a new data point, the algorithm passes the data down each decision tree in the forest and collects the predicted output from each decision tree. The final prediction will be made by aggregating the outputs using either majority voting for classification problems or averaging for regression problems. It can be defined as,

$$f(x) = \frac{1}{Y}\sum_{y-1}^{Y} T_y(x) \ldots\ldots\ldots\ldots\ldots (4)$$

In Equation4, x is input data, $f(x)$ is predicted output, Y is the number of trees present in the forest, T_y is the y-th decision tree in the forest and the sum of the outputs from all the trees in the forest is $\sum_{y-1}^{Y}$.

3.4. ***Ensemble Learning***– Ensemble Learning techniques like voting, bagging, boosting, stacking etc. are very popular among researchers as these are used to merge models which results in higher accuracy and optimized performance.
In our paper, this voting classifier combines the predictions of the models(linear regression, SVM, Random Forest). These models must be trained independently before applying this voting classifier. Each of the models generates their own predictions and the final prediction is made by taking a vote(here mode or most common result) from the models. It can be defined as,

$$P = mode(predicted_{class}[i] \; For \; i \; in \; range \; (model_n)) \ldots\ldots\ldots\ldots\ldots (5)$$

In Equation(5), P is the final predicted class, $model_n$ is the total number of models, $predicted_{class}[i]$ is the predicted class of the i-th model.

IV. RESEARCH METHODOLOGY

In this paper we have introduced a hybrid model consisting of Linear regression, Random Forest and SVM. These all are ensemble with voting technique. There are a lot of web-metrics that can affect the web-page load time directly. Also a lot of research is going on about this topic in the field of machine learning. After an in-depth study, we figured out some important web-metrics like total page size, total number of redirects, total number of videos, FCP, LCP, DOM size, speed index, QLS. These indexes can directly affect the total page load time.

After figuring out these web-metrics, the second vital task is to model choosing. We chose linear regression, random forest and SVM as we saw that these can individually give a pretty good accuracy. In this paper we have also shown a comparative study on the models that we have discussed.

In our proposed model, the model requests for an URL from the user then it extracts the web-metrics which can affect the load time. It makes a dataset array by using it. Then split it into train and test sets. After passing it through the hybrid model it predicts the load time of the webpage. Then by using it, the model is able to give information about FCP, LCP, speed index, QLS and DOM size of the corresponding web-URL. The workflow of our model is shown in Figure 4.1.

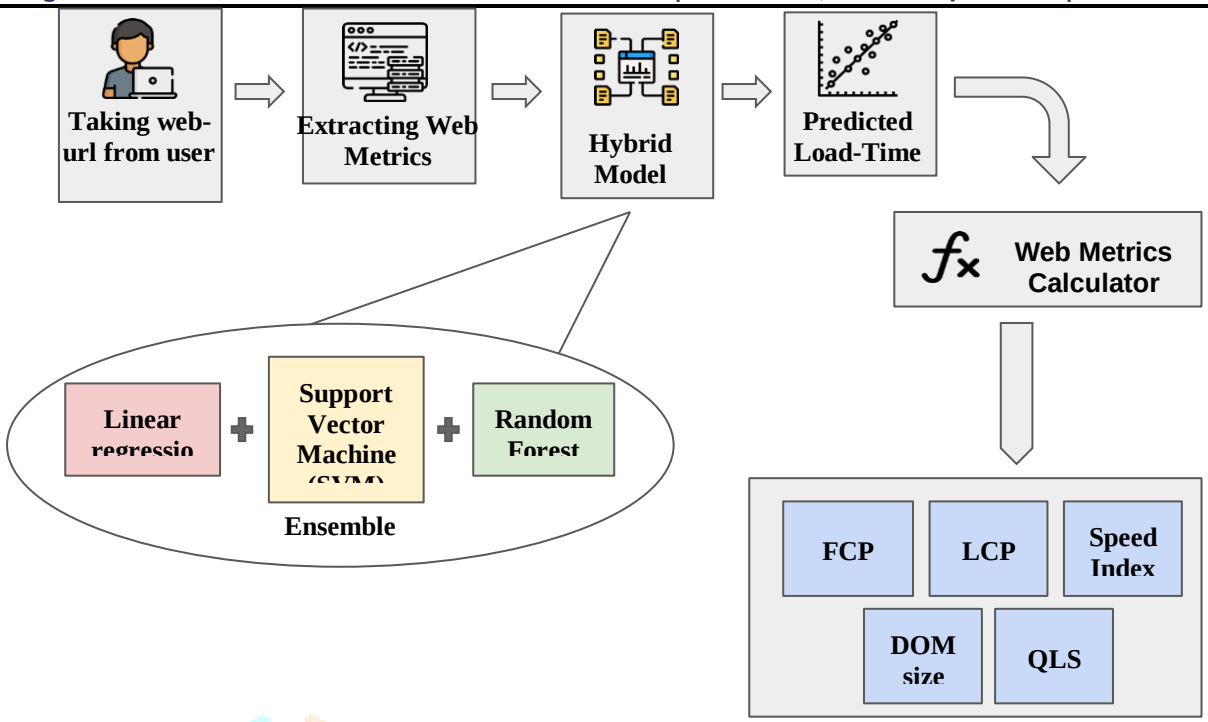

Figure 4.1: Proposed Hybrid model

V. RESULT AND DISCUSSION

As we have discussed earlier, our hybrid model is the combination of Linear regression(LR), Random Forest(RF) and SVM. So, Our proposed model can be defined as,

$$ensemble_{voting} = (LR, RF, SVM) \ldots\ldots\ldots (6)$$

We have done tedious work to show the comparison between various model combinations. Also we have compared the predicted load time with the actual load time present on the experiment day(it may vary as web-pages are updating day by day). Also we have calculated approximated accuracy for better understanding. The detailed comparison is shown in Table 5.1.

Models	Predicted web-page load time(s)	Actual load time(s)	Accuracy(%)
Linear regression	0.049	0.060	81.67
Random Forest	0.056		93.33
Support Vector Machine(SVM)	0.051		85.00
Linear regression + Random Forest	0.053		88.33
Linear Regression + SVM	0.055		91.67
SVM + Random Forest	0.058		96.67
Linear regression + Random Forest + SVM ← proposed model	0.059		98.33

Table 5.1: Comparative Analysis of Different Models

From Table 5.1, we can see that our proposed model has outperformed other models and achieved a high accuracy of 98.33 %. Also as the predicted load time is highly accurate this model can also provide information about other web-metrics like FCP, LCP, Speed Index, DOM size and QLS with high accuracy.

VI. REFERENCES

[1] Wang, S., Zheng, L., Liu, J., & Liu, Z. (2019). Website Page Load Time Prediction Based on Linear Regression. IEEE Access, 7, 103850-103860.

[2] Xue, Y., Han, D., Zhang, Y., Li, W., Zhang, Z., & Li, C. (2017). A Random Forest Regression Model for Web Performance Prediction. Journal of Physics: Conference Series, 974(1), 012043.

[3] Zhang, M., Chen, L., Yang, X., & Cao, H. (2020). A Hybrid Model for Predicting Web Performance Metrics. Wireless Personal Communications, 115, 3219-3234.

[4] Agrawal, R., & Singh, R. (2021). Prediction of Web Performance using Machine Learning. International Journal of Advanced Computer Science and Applications, 12(4), 411-417.