IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 11, November 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 11, November 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1011001 |

A Software Cost and Time Estimation Model Using Machine Learning Techniques

Tawanda Laston Makombe, Monika Gondo

MTech Student, Dept. of Software Engineering, Harare Institute of Technology, Harare, Zimbabwe

Lecturer, Dept. of Software Engineering, CCIS, Harare Institute of Technology, Harare, Zimbabwe

ABSTRACT: No matter what software application a company or an individual sell, the price that they charge the clients or customers will have a direct impact on the success of their business. According to the literature there are various models that can be used to cost software development. Many Software developers and project managers are finding it difficult to accurately attach a price to the software applications prior development. It has been challenging to consider every aspect that affects software development costs. The aim of this study is to accurately anticipate the cost of software development using machine learning techniques and taking into account all the aspects that have an impact on software development. Since explanatory research studies trends across time, the researcher utilized an explanatory study design. SEERA dataset was used which has all the attributes that are for software development. This research considered only three Machine Learning techniques which are K-Nearest Neighbor Regression, Linear Regression, and Random Forest Regression. Random Forest Regression performed better than the other two. This resulted in it being suitable for implementation in the web-system that computes and asserts the software application cost.

KEYWORDS: Software development costs; Machine Learning; SEERA; dataset; Random Forest Regression

I. INTRODUCTION

Software is built and maintained via a systematic, iterative process called software engineering. Problems with the software life cycle are resolved via software engineering. The following stages make up the software life cycle:[1] inception stage, requirement stage, design stage, construction stage, testing stage, deployment stage and maintenance stage.

During conception phase, the project team set project goals, performing various project estimations, and determining the project's scope. User demands are investigated, and technical and functional requirements are discovered, during the requirement or planning phase. Architecture is chosen while keeping needs in mind at every stage of the design process.

The project's execution is done during the construction phase. During the construction phase, a prototype model is created. The working model of the prototype is then put into use. Finding bugs, mistakes, and defects is done during the testing phase, and then the software's quality is evaluated. The software then moves on to the deployment phase, when it is made available to the environment for use by end users, following a successful testing period. The end-users' feedback is gathered and software enhancement is carried out by the developers during the maintenance phase.[1]

Initial phase involves the process of approximating the cost using a variety of unclear and unreliable data. Estimates utilized as input for budget analysis, investment analysis, iteration planning, project planning, and other activities. Estimation is used to determine the size of a system or product and also time, effort, human skill, money, and resources required to develop the product. Although various models have been established over the past 20 years, measuring effort is still a challenging problem since the early stages of software development involve more unreliable and unclear data. Software effort is measured in person-months or hours.

The literature analysis deduced that analogy-based assessment, expert judgment, function point investigation, machine-learning techniques comprising of regression techniques, classification approaches and clustering methods, neural network and deep learning models, fuzzy-based approaches and ensemble methods are used for SDECE. [1]

Cost estimation is a tool for evaluating resource allocation, planning, and budgeting for software projects. Prior the SDECE of a project, it is essential to first comprehend the software`s actual requirements, level of complexity, and

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 11, November 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1011001 |

other cost-influencing factors such as product attribute, project attribute, personal attribute and hardware attribute. These attributes or factors serve as the process' input for cost calculation. [2]. As a result, the method typically returns three results: effort, development duration, and resources.

The main obstacles to software development and commercialization are under-pricing, over-pricing, overbudgeting, and underbudgeting; as a result, accuracy in software effort estimating is constantly required. The complexity of software projects has significantly increased during the past few decades. [3] and this led to the creation and application of several approaches for calculating the time and expense involved in developing software, including both conventional methods and machine learning techniques. Future software development has two major challenges: overestimating and underestimating, therefore precision in software effort prediction is always needed.

However, the study by Noor Azura Zakaria and Amelia Ritahani Ismail [4] indicates that in the Machine Learning techniques there is need of performing ensemble stacking also known as blending, in order to optimize the predictive models.

According to the research by Lalit V. Patil et al [5], The process of approximating the cost of software is dependent on factors like the number of team members, the programming language and software tools being used, the salaries and overhead expenses of the development team, the size of the database being used, the cost of training, unintentional rework, the policy being used within the organization, the cost of shared facilities like a restaurant or library, and the use of resources like light and networks, among others. Each of these elements helps to clarify software cost calculation.

The research by Hareton Leung [5] stated that in addition to software size, software effort is influenced by many other cost attributes. COCOMO II model proposes and employs the most complete list of price considerations. These cost considerations are classified into four categories:

- **Computer attributes:** execution time limitation; main storage limitation; computer turnaround limitations; platform instability.
- **Personnel attributes:** analyst competence; application knowledge; programming competence; platform knowledge; language and tool knowledge; personnel endurance.
- **Product attributes:** required trustworthiness; product intricacy; database size used; required reusability; documentation match to life-cycle needs.
- **Project attributes:** multisite development; use of software instrument; required development plan.

The attributes listed above are not always self-determining and of which many are difficult to quantify. Other aspects appear in combination in many models, while others are simply ignored. Other factors take discrete values are not in continuous form, resulting in a piece-wise form for the estimation function.[5]

Although some of the factors mentioned above are difficult to quantify, this research proposes to take into account all of the factors that influence the software development process. It aims to use machine learning techniques to quantify all of the factors that influence software development.

II. RELATED WORK

In 2019, Mahmood Mohd Al Asheeri and Mustafa Hammad published Machine Learning Models for Software Cost Estimation. The goal of this paper was to address the challenge of creating an accurate cost estimation model for software project development. [6] The dataset that was used is Usp05-ft and Usp05 and the solution was the Random Forest. It had 0.8441 R squared. More datasets must be included to provide a broader perspective and a greater variety of inputs, resulting in better estimation and more accurate results. To incorporate all machine learning techniques currently in use and increase accuracy, further machine learning algorithms can be tested and added.

A. SaberiNejad and R. Tavoli discovered K-Nearest Neighbor as the solution to the problem of cost estimation that had plagued systems analysts, project managers, and software engineers for years in their paper titled A Method for Estimating the Cost of Software Using Principal Components Analysis and Data Mining in 2018. They used COCOM and NASA data sets which resulted 94.74% accuracy in their solution. The gap that was not filled by their research is of applying different learning algorithm of machines and a different software work and also to use different methods such as wrapper in order to improve software cost estimations. [13]

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 11, November 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1011001 |

This research proposes a solution in which it uses other machine learning techniques and compare to find out the best.

III. METHODOLOGY/ APPROACH

Since explanatory research studies trends over time, the researcher utilized explanatory research design. [7]. As with any other explanatory research design, the researcher began by reviewing all prior literature on software cost estimation and projections. The explanatory research design has been regarded as the best research design for projects involving a high level of uncertainty and ignorance about the subject.

Dataset used is called SEERA dataset [8] and was obtained from the published data source Zenodo with the website www.zenodo.org. Jupyter notebook is the main software that was used for data analysis and also Microsoft-word was used to write the document.

The attributes of the dataset used are divided into eight (8) categories of which six (6) are: general information, size, users, developers, project and product. The other two (2) attributes are effort representing the effort attributes and environment representing local attributes.

On the modelling the data was modelled using the three machine learning algorithms that are Random Forest Regression, K-Nearest Neighbour Regression and Linear Regression. All of techniques considered are supervised machine learning algorithms. Random Forest Regression employs the ensemble learning technique, which consolidates forecasts from multiple machine learning algorithms to come up with an improved accurate prediction than a single model. KNN regression belongs to a non-parametric method that averages observations in the same neighbourhood to approximate the connection between independent variables and continuous outcomes. While KNN can be used to solve either regression or classification problems, it is most commonly used as a classification algorithm, assuming that similar points can be found nearby.

All the three machine learning algorithms were modelled using the following steps [9]:

Step 1: Data collection or gathering the dataset
Step 2: Identify the dependent (y) and independent variables (X) from the dataset
Step 3: Split the dataset into the Training set and Test set
Step 4: Training the model on the whole dataset and in this paper the dataset used is called SEERA dataset.
Step 5: Predicting the Test set results

The models were then assessed using the following metrics: MAE, MSE, RMSE and R squared.

The MAE calculates mean magnitude of errors in a set of estimates without taking into account their path. [10] The formula for MAE is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$$

MSE calculates the average squared difference between observed and predicted values. The MSE is zero when there is no error in a model. As model error increases, so does the model's value. [11] The formula for MSE is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2$$

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 11, November 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1011001 |

The RMSE is calculated using the square root of the average absolute moment of deviations between predicted and observed values, or the quadratic mean of these deviations. [10] The RMSE syndicates the sizes of prediction errors for different data points to produce a single measure of predictive power. The RMSE formula is shown below.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2}$$

By comparing it to the average line of the dependent variable, the R-squared shows how good the model is fitted to the data. [12] Below is the formula for R-squared

$$\text{Rsquared} = 1 - \frac{\sum_{i=1}^{n}(x_i - x)^2}{\sum_{i=1}^{n}(x_i - x)^2}$$

IV. RESULTS AND DISCUSION

The dataset used in this research had 120 entries and 72 attributes. The attributes are categorized into: general information, size, users, developers, project, product and effort.

The dataset was divided to have a subset in which the researcher focused on. The factors or attributes of the whole dataset which were not considered in the subset are the ProjID ProjID.1 year and Organization id. They were not considered because the researcher regarded them as they do not contribute to the cost estimation of the project.

Jupyter from Aanconda 3 has been used to evaluate the used algorithms. The dataset was divided into two parts (2) in which 70% was to train the data and the remaining 30% was used to test the predictions until the whole data (100%) has used.

In the Jupyter tool, three (3) types of Machine Learning techniques were used to determine the best algorithm to use for Software Cost Estimation. The used ML algorithms are Random Forest Regressor, KNN Regressor and Linear Regression.

Table 1 below shows the presentation of the results obtained from the implementation of the three algorithms on the SEERA dataset and under the same conditions.

Table 1: Results of ML Algorithms used

Algorithm Used	MAE	MSE	RMSE	R squared
Linear Regression	5099.4732	54678089.5528	7394.4634	0.6062
KNN Regression	6365.2083	156307016.5486	12502.2805	0.2268
Random Forest Regression	3178.285	37890909.6126	6155.5592	0.7271

Random Forest Regression outperformed all other algorithms in the criterion used to test and evaluate the error rate, as shown in Table 1, in which it had the least error rate for the three types (MAE, RMAE, and RAE) and the highest correlation coefficient R-squared.

Nevertheless, other algorithms showed low performance with the worst results coming from the Linear Regression. Because of this, they were unable to get accurate prediction values while taking into account all 68 parameters from the dataset.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 11, November 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1011001 |

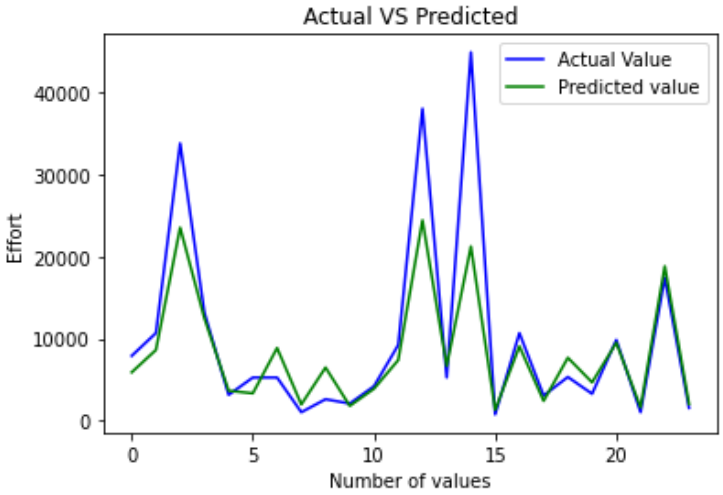

Figure 1: Random Forest Regressor algorithm Diagram

The value distribution for the top outcomes determined by the Random Forest Regressor algorithm is depicted in Figure 1. The x-axis represents the dataset's frequency, and the y-axis represents the estimated and actual values. The corresponding points in both figures show the variances between the actual effort points and those predicted by the Random Forest Regression model. The predicted values match the actual values almost exactly. Certain predicted values, on the other hand, differ significantly from the dataset's effort values, increasing the error rate.

V. CONCLUSION

Overestimation and underestimation are both major challenges for future software development; as a result, there is a constant need for accuracy in software effort estimation. Therefore, appropriate models for software development cost are of great interest to project and operational managers, software development companies, investors and financial institutions.

Using the SEERA dataset, the researcher ran certain checks on the data using Python software. All three (3) machine learning algorithms taken into consideration in this study. The dataset was divided into two parts (2) in which 70% was to train the data and 30% is for testing. The researcher also did model building and performed same model evaluation on all the tree (3) algorithms considered. In terms of the R-squared, Random Forest Regression had 0.7271, KNN Regression had 0.2268 and Linear Regression had 0.6062. This resulted in the Random Forest Regression considered to be the top amongst the three.

Based on the conclusions drawn from this study project, it is possible to develop a website design that uses the Random Forest Regression technique to forecast the cost and price of the software.

The future researchers are recommended to design a model that do the estimation of the cost of software applications using the Deep Learning techniques. It is also recommended to us other datasets not only to be limited to the SEERA dataset which was used in this research. Other researchers are also encouraged to do more researches which contribute to the increase of more datasets of different countries and continents which enable researchers to do researches that are more aligned to their countries.

REFERENCES

[1] P. V. A. G, A. K. K, and V. Varadarajan, “Estimating Software Development Efforts Using a Random Forest-Based Stacked Ensemble Approach,” pp. 1–21, 2021.

[2] A. Wicaksana, *Project Management ToolBox*. 2016. [Online]. Available: https://medium.com/@arifwicaksanaa/pengertian-use-case-a7e576e1b6bf

[3] L. McLeod and S. G. MacDonell, “Factors that affect software systems development project outcomes: A survey of research,” *ACM Comput. Surv.*, vol. 43, no. 4, pp. 24–56, 2011, doi: 10.1145/1978802.1978803.

[4] N. A. Zakaria, A. R. Ismail, A. Y. Ali, N. H. M. Khalid, and N. Z. Abidin, “Software Project Estimation with Machine Learning,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 726–734, 2021, doi: 10.14569/IJACSA.2021.0120685.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 11, November 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1011001 |

[5] H. Leung, "Software Cost Estimation (W02)," 2001, Accessed: Sep. 08, 2022. [Online]. Available: https://www.computing.dcu.ie/~renaat/ca421/report.html

[6] M. Mohd and A. Asheeri, "Machine Learning Models for Software Cost Estimation," no. September, 2019, doi: 10.1109/3ICT.2019.8910327.

[7] T. Boru, "CHAPTER FIVE RESEARCH DESIGN AND METHODOLOGY 5 . 1 . Introduction," *CHAPTER FIVE Res. Des. Methodol. 5.1. Introd.*, no. December, p. 41, 2018, doi: 10.13140/RG.2.2.21467.62242.

[8] E. I. Mustafa and R. Osman, "SEERA: A software cost estimation dataset for constrained environments," *PROMISE 2020 - Proc. 16th ACM Int. Conf. Predict. Model. Data Anal. Softw. Eng. Co-located with ESEC/FSE 2020*, pp. 61–70, 2020, doi: 10.1145/3416508.3417119.

[9] P. Dönmez, "Introduction to Machine Learning, 2nd ed., by Ethem Alpaydın. Cambridge, MA: The MIT Press2010. ISBN: 978-0-262-01243-0. $54/£ 39.95 + 584 pages.," *Nat. Lang. Eng.*, vol. 19, no. 2, pp. 285–288, 2013, doi: 10.1017/s1351324912000290.

[10] W. Wang and Y. Lu, "Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 324, no. 1, 2018, doi: 10.1088/1757-899X/324/1/012049.

[11] Z. Wang and A. C. Bovik, "Mean Squared Error : Love It or Leave It ?," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, 2009, [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4775883

[12] D. B. Figueiredo Filho, J. A. Silva Júnior, and E. C. Rocha, "What is R2 all about?," *Leviathan (São Paulo)*, no. 3, p. 60, 2011, doi: 10.11606/issn.2237-4485.lev.2011.132282.

BIOGRAPHY

Tawanda Laston Makombe is a Master of Technology student in the Software Engineering Department in the School of Information Sciences and Technology at the Harare Institute of Technology. He finished the Bachelor of Technology degree in 2017 at the Harare institute of Technology. His research interests are focused in the Machine Learning, Software Pricing, etc.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details