Risk Prediction with Regression in Global Software Development using Machine Learning Approach: A Comparison of Linear and Decision Tree Regression

1,2Asim Iftikhar, 1Shahrulniza Musa, 1,2Muhammad Alam, 4Rizwan Ahmed, 1Tariq Rahim Soomro and 3Mazliham Mohd Su‘ud

1College of Computer Science and Info. Sys, Institute of Business Management (IoBM), Korangi Creek, Karachi, Pakistan
2Malaysian Institute of Information Technology, Universiti Kuala Lumpur, (UniKL MIIT), Kuala Lumpur, Malaysia
3Chancellery, Universiti Kuala Lumpur, Kuala Lumpur, Malaysia

Abstract: Software development through teams at different geographical locations is a trend of modern era, which is not only producing good results without costing lot of money, but also productive in relation to its cost with low risk and high return. This shift of perception of working in a group rather than alone is getting stronger day by day and has become an important planning tool and part of their business strategy. Due to this phenomenal shift the development processes have become complex and chances of risks have been increased. The utilization of Machine learning to manage risk is helpful when taking care of and evaluating data. In this research regression approaches like Linear Regression and Tree Regression have been implemented to predict the responses of risks involved in global software development. Comparative analysis has also been performed between these two algorithms to determine the highest accuracy algorithms. The results indicate that Fine tree regression, which is one of techniques of decision tree regression, gave better results in terms of goodness of fit measures as compared to linear regression model fitted to examine the relationship of cost, time and resource related risk with the overall risk of global software development projects.

Keywords: Global Software Development, Risk Management in Global Software Development, Machine Learning, Linear Regression, Decision Tree Regression

Introduction

During recent times in order to offer advantages over conventional techniques the software development environment has shifted to a distributed environment from unified one (Al-Zaidi and Qureshi, 2017). Gradual accomplishment depends upon its utilization as a competitive weapon. Since past Ten years, many software firms began to discover or test with the distributed software development facilities and with sub-contracting to search for cheaper and skilled resources (Prikladnicki et al., 2003). As a result, software development is a multsite, diverse and globally a distributed work. At various levels the designers, engineers, managers and officials have to face challenges from being social and cultural compared to specialize (Herbsleb and Moitra, 2001; Prikladnicki et al., 2003). Such groups are named remotely dispersed at number of locations by different experts or Global Software Development (GSD) environment (Iftikhar et al., 2018a).

Global Software Development

The GSD is well known amongst IT organization and is gaining traction as more and more employees with relevant skills and experience are grabbing the opportunities of global assignments, as it offers lucrative perks, irrespective of the assignment duration (Arunmugam and Kaliamourthy, 2016). Reasons of GSD popularity includes unwavering product development, consistent reduction in ‘time to market ratio’ to outpace outdated products and services, low cost labor, gradual enhancement in product quality and access to economically viable skilled resources (Al-Zaidi and Qureshi, 2014; 2017; Anjum et al., 2006). In GSD environment distributed teams are still facing many challenges during GSD process such as cultural issues, strategic issues, Inadequate communication, distance, different backgrounds and project and process management issues (Casey and Richardson, 2009; Herbsleb and Moitra, 2001) as shown in Fig. 1.
Global software development projects are generally operational on a broader scale and it leads to a remarkably increased hurdle which leads to intense danger. Offshore projects are generally not productive because of “cultural differences, time constraints, stakeholder and organizational distances deleteriously clout communication and lacks knowledge exchange amongst onshore and offshore project team members” (Fabriek et al., 2008; Verner et al., 2014). When a software project is carried out in a distributed environment within several countries, then the software project manager should address operational risks, such as those related to communication, coordination, time zone differences, project setting and infrastructure (Casey, 2009; Hossain et al., 2009a, 2009b; Verner et al., 2014).

GSD Risk Management

The risk management procedure is the methodical function of management strategies, measures and practices towards the actions of interactive, referring launching of the context. It further includes identifying, analyzing, evaluating, handling, inspecting and reviewing risk (Chadli et al., 2016; Iftikhar et al., 2018b). There are some threats involved in it also when managing a Global Software Development project because of the team located in different localities. The risk factor can be confronted including organizational, cultural, temporal, language, political and geographical obstacles. These obstacles can arise when dealing with a Global Software Development project present as risks (Galli, 2018).

In this research Linear Regression and Tree Regression approaches have been implemented to predict the risks involved in GSD environment. Results have been presented using predicted vs actual plot. Comparative analysis has also been performed between these two regression algorithms to determine the best performance.

This study comprises of five (5) sections. The first section covers the introduction of this research study. Related work with respect to research will be explained in section 2. The Machine learning and its algorithms utilized in this research be depicted in section 3. Research methodology will be elaborated in section 4. Section 5 will talk about the Results and findings and the last segment will conclude this research.

Related Work

An experimental study (Rathore and Kumar, 2016) was conducted on five different open-source projects to determine the capability of capabilities of Decision Tree Regressions (DTR) for predicting the number of faults in intra-release and inter-release scenarios. The PROMISE repository used in the experiment included data of nineteen releases. The authors used various measures such as Absolute Error, Relative Error, Prediction at Level 1 and Goodness-of-Fit to evaluate the accuracy of DTR. Based on the results the prediction accuracy of DTR was found to be significant in predicting the number of faults with better accuracy for inter-releases faults across all datasets.

Researchers in (Myrtveit et al., 2005) have proposed a model based on experience factory approach for well-organized and effective knowledge and experience management for software development industry. The study has identified a set of predictors for the proposed model. The relationship between the identified predictors was verified through a correlational survey research. Reliability and Regression analysis were carried out to validate various relationships. According to successful analysis, an experienced factory organization is far more optimal in terms of Experience goals as opposed to Project organization. Simultaneously, due to discipline in relationship, no serious consequences in the model were experienced.

Researchers in (García-Floriano et al., 2018) attempted to predict software enhancement by applying...
prediction accuracy of two types of Support Vector Regressions (ν-SVR and ε-SVR) and analyzing both techniques. Several machine learning kernels used for both ν-SVR and ε-SVR in the study include: Kernel Function linear, a polynomial, radial basis function and sigmoid kernels. In the same study, the researchers further tested prediction accuracies for ε-SVR and ν-SVR against those of association rules, statistical regressions, decision trees and neural networks found with 95% confidence that the polynomial kernel ε-SVR was statistically better than all that they compared.

The planning, development and controlling of the software are very crucial goals for software engineering. The effort estimation is one of the important parameters to predict the amount of effort needed to develop or maintain a software. This is one of the most difficult areas of software engineering as good predictive models are hard to come by. A good study has been done by a set of researchers (Jayaram et al., 2018) who did the effort estimation for small-scale visualization projects. These visualization projects were developed by postgraduate students in pure academic settings however the projects themselves were very relatable to even general audience. The projects include The Cricket Game, Sudoku, Blackjack on the gaming side. While other projects were related to science and engineering with very important real-life applications. They used seven novel software parameters they list in their abstract as Cumulative Grade Point Average (CGPA), Lines of Code (LOC), New and Changed code (N&C), Reuse code (R), Cyclomatic Complexity (CC), Functional Points (FP) and Algorithmic Complexity (AC) which are considered important in software development effort. They report the Mean Magnitude of Error Relative to estimate (MMER) to be 0.006 for multiple linear regression, 0.002 for nonlinear regression and 0.002 for neural network models. The authors argue that since the differences are marginal in the error estimates, hence the three models can be alternatively used for visualization projects.

**Machine Learning**

Artificial Intelligence has many applications one of which is machine learning. Machine learning develops itself and enhances its performance automatically using only previous experiences without taking aid from computer based programs. Concentration of machine learning for the most part is the training of computer programs that can extract data and make it valuable through rigorous analysis. If we think about coming times the idea behind the machine learning is to sustain and improve quality decision making that is based upon analyzing data, observing patterns, previously received instructions or examples or any direct experience. Concisely underlying objective is to permit computers to learn, unlearn and relearn on itself (Van Liebergen, 2017).

Since the 1950s the ‘thinking machines’ has been the topic of interest to computer scientists as well as mathematicians. The last 30 years have been filled with tremendous progress both in the laboratory setting as well as in the commercial settings (Jordan and Mitchell, 2015). ML is a very well-chosen method within the field of Artificial Intelligence (AI). Within AI, it is used to develop useful software systems for many applications such as speech recognition and processing, computer vision and robotics among many other applications too numerous to count here. ML capabilities are added to a system via software systems with ML components and frameworks via tools and libraries that provide ML functionalities as discussed elsewhere in more details (Wan et al., 2019).

**Linear Regression**

Linear regression has been a mathematical tool known since the early 19th century, when it was used by famous mathematicians like Legendre and Gauss to study planetary system. It simply models the relationship between a dependent variable (usually denoted as y) and an independent variable (usually denoted by x) as shown in Fig. 2. The dependent variable can be studied with more than one independent variable and it is a modelling technique which has been studied very extensively in both academic and commercial environments. Due to its simplicity in usage and well-known behaviors, the Linear Regression is used a lot for algorithms in Machine Learning (Tanner, 2020).

The linear regression formula as shown in Eq. 1 and 2 where j is dependent variable and I is independent variable:

\[
m = \frac{(\sum j)(\sum i^2) - (\sum i)(\sum ij)}{n(\sum i^2) - (\sum i)^2} \tag{1}
\]

\[
n = \frac{n(\sum ij) - (\sum i)(\sum j)}{n(\sum i^2) - (\sum i)^2} \tag{2}
\]

The Interaction linear regression formula as shown in Eq. 3 where an interaction effect can be described as the changes in the effect of an independent variable on a dependent variable:

\[
\hat{y} = n_0 + n_1X_1 + n_2X_2 + n_3X_1X_2 \tag{3}
\]

The Stepwise linear regression formula as shown in Eqs. 4 to 6 where Stepwise regression basically performs multiple regression number of times and removes the weakest correlated variable each time:

\[
\hat{y} = n_0 + n_1X_1 \tag{4}
\]
\[ \hat{y} = n_0 + n_1X_1 + n_2X_2 \]  
\[ \hat{y} = n_0 + n_1X_1 + n_2X_2 + n_3X_3 \]

The Eq. 7 shows the Robust Linear Regression formula where Robust regression is an iterative procedure that finds to recognize outliers and minimize their effect on the coefficient estimates:

\[ \min_{\beta} \sum_{i=1}^{N} \rho(y_i - \bar{x}, \beta) = \min_{\beta} \sum_{i=1}^{N} \rho(e_i) \]

**Decision Tree Regression**

Decision Tree Learning (DTL) is a modelling technique used when one wants to predict the value of a certain target variable based on a number of input variables. Decision Trees when the target variable can take continuous values (for example involving real numbers) is called Regression Trees. The technique develops regression or classification models based on the tree structures by breaking down a dataset into smaller subsets. It divides the dataset into smaller subsets that contains homogenous values. These tree structures are named as Fine Tree, Medium Tree and Coarse Tree. The measure of Mean Square Error (MSE) is used as a decision criterion to split a node into a number of subnodes. The regressors which are the natural generalization of decision trees for regression problems are of interest to us as discussed by [put lead author’s name here] due to the efficiency of these regressors. As they show in their paper that instead of a class label being associated to every node, a real value of some of the inputs is all what they need to use in order to predict the value of the output (Dobra and Gehrke, 2002).

The decision tree produces a model that may have the rules that can be interpreted with logical statements. Also, the axis parallel decision surfaces produced in decision trees is an important characteristic of this approach that make it superior to other techniques (Jamal and Nodeli, 2017).

Some studies compared the relative efficiencies of statistical methods with data mining techniques. Using RMSE as a measure, it was found that in case of continuous independent variables, linear regression gives better results as compared decision tree and ANN. Linear regression was also found best for continuous and categorical independent variables if the number of categorical variables is one (Kim, 2008).

In Fine tree regression, there are many leaves to make many fine distinctions between classes and it is considered more complex structure with maximum 100 splits. The medium tree regression is based on medium number of leaves for finer distinctions between classes with maximum 20 numbers of splits. Coarse tree regression has maximum 4 splits and it is considered as the simplest structure tree regression. Considering the complexity of the model, Fine tree regression has the highest value of \( R^2 \) with the least RMSE value as compared to the Medium Regression Tree, whereas the Coarse Regression Tree has the least \( R^2 \) with the highest RMSE value.

![Fig. 2: Linear regression (Tanner, 2020)](image_url)
Hypothesis 2: Risk related to cost has a significant effect on the overall risk of a GSD project

Hypothesis 3: Risk related to resource has a significant effect on the overall risk of a GSD project

The hypotheses stated above are to be tested using the p values corresponding to the goodness of fit measure ($R^2$) of the fitted regression models using the techniques of linear regression and decision tree regression. The best model is to be chosen out of seven different models fitted on the data. The measures of goodness of fit ($R^2$), RMSE, MSE, MAE, Prediction Speed and Training Time are used to select the best model.

**Research Methodology**

Under umbrella of Linear Regression Algorithm (Linear Regression, Interactions Linear Regression, Robust Linear Regression and stepwise Linear Regression) and under umbrella of Tree Regression (Fine Tree Regression, Medium Tree Regression and Coarse Tree Regression) algorithms implemented for regression to manage this risk in software project.

The three types of risks namely time risk, cost risk and resource risk are collectively responsible for an overall risk in a project related to GSD. But their weightage or effect on an overall risk may not be equal. Therefore, there is a need to examine if the three types of risks have an equal and significant effect on an overall risk or some risk factor may affect the overall risk more as compared to the other factors. To address this perspective, three hypotheses have been developed that are to be tested to compare the relationship between each determinant of risk with the overall risk.

**Hypothesis 1:** Risk related to time has a significant effect on the overall risk of a GSD project

**Hypothesis 2:** Risk related to cost has a significant effect on the overall risk of a GSD project

**Hypothesis 3:** Risk related to resource has a significant effect on the overall risk of a GSD project

Table: Sample dataset

| Country | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 | Q12 | Q13 | Q14 | Q15 | Q16 | Q17 | Q18 | Q19 | Q20 | Q21 | Q22 | Q23 | Q24 | Q25 | Q26 | Q27 | Q28 | Q29 | Q30 | Output |
|---------|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| USA | 1 | 0 | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 1 | 1 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 1 | 4 | 4 | 4 | 3 | 0 | 3 |
| AUS | 0 | 0 | 0 | 1 | 1 | 2 | 3 | 3 | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 4 | 1 | 3 | 4 | 4 | 1 | 4 | 4 | 4 | 3 | 0 | 2 |
| PAK | 0 | 0 | 0 | 1 | 1 | 2 | 3 | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 4 | 1 | 3 | 4 | 4 | 1 | 4 | 3 | 1 | 0 | 0 |
| USA | 3 | 2 | 0 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| AUS | 1 | 1 | 1 | 0 | 1 | 2 | 2 | 2 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 1 | 3 | 4 | 3 | 3 | 4 | 4 | 3 | 4 | 2 | 3 | 2 | 1 | 2 |
| PAK | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 1 | 3 | 4 | 3 | 3 | 4 | 4 | 3 | 4 | 2 | 3 | 1 | 2 | 3 |
| USA | 3 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| USA | 0 | 0 | 0 | 1 | 1 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| USA | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| AUS | 1 | 1 | 1 | 0 | 1 | 2 | 2 | 2 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 1 | 3 | 4 | 3 | 3 | 4 | 4 | 3 | 4 | 2 | 3 | 1 | 2 | 3 |
| PAK | 3 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| USA | 3 | 0 | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| USA | 1 | 1 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| USA | 2 | 1 | 1 | 1 | 2 | 2 | 1 | 4 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| USA | 2 | 3 | 2 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| USA | 2 | 1 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
Results and Finding

To make the data ready to fit the models, three variables time risk, cost risk, resource risk and overall risk, the average of the response score of the questions pertinent to each variable is created. To test the hypotheses stated in the study, three different combinations of predictor and response variables are used i.e., cost risk versus overall risk; time risk versus overall risk and resource risk versus overall risk. Linear Regression and Tree Regression parameters (Root Mean Square Error (RMSE), Mean Squared Error (MSE), R-Squared, Mean Absolute Error (MAE), Training Time and Prediction Speed) have been calculated in all variants of Regression. Response plot of both regression techniques has also been shown in Fig. 3 to 23. RMSE, R-Squared, MSE and MAE can be calculated using the Formulas as shown in Eqs. 8 to 11.

The RMSE is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed (Chai and Draxler, 2014):

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (P_i - O_i)^2} 
\]  \hspace{1cm} (8)

R-squared ($R^2$) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model (Heinzl and Mittlböck, 2003):

\[
R^2 = 1 - \frac{n\sum{ij} - \left(\sum{i} \right) \left(\sum{j} \right)}{n\sum{i}^2 - \left(\sum{i} \right)^2} \sqrt{\frac{n\sum{j}^2 - \left(\sum{j} \right)^2}{n\sum{i}^2 - \left(\sum{i} \right)^2}} 
\] \hspace{1cm} (9)
Fig. 5: Robust linear response plot of time related risk

Fig. 6: Stepwise linear response plot of time related risk

Fig. 7: Fine tree response plot of time related risk
DOI: 10.3844/jcssp.2021.67.89

Fig. 8: Medium tree response plot of time related risk

Fig. 9: Coarse tree response plot of time related risks

Fig. 10: Linear regression response plot of cost related risks
Asim Iftikhar et al. / Journal of Computer Science 2021, 17 (2): 67.89
DOI: 10.3844/jcssp.2021.67.89

Fig. 11: Interaction linear response plot of cost related risks

Fig. 12: Robust linear response plot of cost related risks

Fig. 13: Stepwise linear response plot of cost related risks
Fig. 14: Fine tree response plot of cost related risks

Fig. 15: Medium tree response plot of cost related risks

Fig. 16: Coarse tree response plot of cost related risks
Fig. 17: Linear regression response plot of resource related risks

Fig. 18: Interaction linear response plot of resource related risks

Fig. 19: Robust linear response plot of resource related risks
Fig. 20: Stepwise linear response plot of resource related risks

Fig. 21: Fine tree response plot of resource related risks

Fig. 22: Medium tree response plot of resource related risks
The MSE is the average squared difference between the estimated values and the actual value (Heinzl and Mittlböck, 2003):

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(10)

MAE is a measure of difference between two continuous variables (Chai and Draxler, 2014):

$$MAE = \frac{1}{n} \sum_{j=1}^{n} |y_j - \hat{y}_j|$$

(11)

Table 2, 3 and 4 shows the result comparison of Linear Regression types (Linear Regression, Interaction Linear Regression, Robust Linear Regression, Stepwise Linear Regression) and Tree Regression (Fine Tree Regression, Medium Tree Regression, Coarse Tree Regression). After comparison, results proved that fine tree regression outperformed linear regression in all three regression models with seven alternatives that indicates the existence of nonlinearity or relatively different degree of variability across different segments of the dataset. Considering the case of time related risk versus overall risk presented in Table 5, the Fine Tree Regression achieved minimum RMSE 2.9192 and highest R-Squared value of 0.74. The findings are similar in case of cost related risk when linked with the overall risk and Fine Tree Risk gives the RMSE 2.4229 with R-Squared value of 0.82. Furthermore, comparing resource related risk with overall risk exhibits the minimum RMSE 4.1241 and 0.48 as the value of R-squared. In these three cases the Robust Linear regression and Coarse Tree regression approaches attained maximum MSE with minimum R-squared values. Using the criterion of explained variation, the cost related risk ($R^2 = 0.82$) can be considered strongly linked with the overall risk followed by the time related risk ($R^2 = 0.74$) whereas, the resource related risk has the least value of R-squared ($R^2 = 0.48$). These results revealed that the cost related risk contributes to the highest degree to the overall risk associated with Global Software Development projects, whereas time related risks is another important factor that has an effect on the overall risk. The resource related risk does also have the effect on the overall risk of the project but it appeared the least. The $R^2$ values associated to each factor indicate that the $p$ values associated to the three best fitted models are less than 0.05 and found significant due to a sample size of 274, which is sufficiently large to support these. Therefore, it can be concluded that all three hypotheses about the risk pertinent to cost, time and resources are supported and implies that all three types of risks have a strong influence on the overall cost.

The predicted vs actual graph has been plotted to determine the strength of relationship between predicted and actual variables in predictive model. The predicted model would be considered to be more accurate relatively if the dots are located closer to the 45° line. Predicted vs actual plot of both regression techniques related to project time, cost and resource risks has been shown in Fig. 24 to 44 in Appendix. The more you closer to the value of 1 it would be considered as more accurate predictive model. The plot demonstrated that the fitness of the predictive model between predicted and actual. The dots showed the predicted class values and intersecting line which is linked with all the dots shows the fitness of the model with respect to predicted values.
Acknowledgement

I wish to offer my genuine thanks to my research companion Mr. Talha Ahmed Khan and my mentors whose consistent help and directions made this research possible.

Funding Information

I also would like to thanks Universiti Kuala Lumpur’s (UniKL) and Institute of Business Management (IoBM) who partially funded this research.

Author’s Contributions

Every author has equal contribution in this research.

Ethics

This research paper is genuine and all authors have read it thoroughly and approved that it does not contain any material which is already published. In this article no ethical issues are involved.

References


Casey, V. (2009, July). Leveraging or exploiting cultural difference?. In 2009 Fourth IEEE International Conference on Global Software Engineering (pp. 8-17). IEEE.


Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?- Arguments against avoiding RMSE in the literature. Geoscientific model development, 7(3), 1247-1250.


APPENDIX


Fig. 24: Linear regression predicted Vs actual plot of time related risks

Fig. 25: Interaction linear predicted Vs actual plot of time related risks
**Fig. 26:** Robust linear predicted Vs actual plot of time related risks

**Fig. 27:** Stepwise linear predicted Vs actual plot of time related risks

**Fig. 28:** Fine tree predicted Vs actual plot of time related risks
Fig. 29: Medium tree predicted Vs actual plot of time related risks

Fig. 30: Coarse tree predicted vs actual plot of time related risks

Fig. 31: Linear regression predicted
Fig. 32: Interaction Linear Predicted vs Actual Plot of Cost Related Risks

Fig. 33: Robust Linear Predicted vs Actual Plot of Cost Related Risks

Fig. 34: Stepwise Linear Predicted vs Actual Plot of Cost Related Risks
Fig. 35: Fine tree predicted vs actual plot of cost related risks

Fig. 36: Medium tree predicted vs actual plot of cost related risks

Fig. 37: Coarse tree predicted Vs actual plot of cost related risks
Fig. 38: Linear Regression Predicted vs Actual Plot of Resource Related Risks

Fig. 39: Interaction Linear Predicted vs Actual Plot of Resource Related Risks

Fig. 40: Robust linear predicted vs actual plot of resource related risks
Fig. 41: Stepwise linear predicted vs actual plot of resource related risks

Fig. 42: Fine tree predicted Vs actual plot of resource related risks

Fig. 43: Medium tree predicted Vs actual plot of resource related risks
Fig. 44: Coarse tree predicted vs actual plot of resource related risks