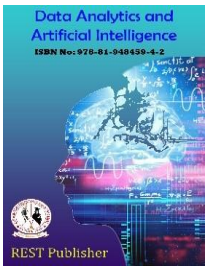
Data Analytics and Artificial Intelligence

Vol: 5(2), 2025

REST Publisher; ISBN: 978-81-948459-4-2

Website: http://restpublisher.com/book-series/daai/

DOI: https://doi.org/10.46632/daai/5/2/8



An Empirical Study on Migration Time Estimation for Netezza-to-Azure Cloud Migration

***Venkata Ramachandra Karthik Chundi**

Staff Software Engineer, GE Vernova

*Corresponding author email: venkatakarthikchundi@gmail.com

Abstract: *Predicting data migration time accurately plays an important role in organizing and executing large-scale database migration projects. This study examines the effectiveness of machine learning regression approaches in forecasting migration duration using three essential factors: data size, table count, and database complexity. Linear Regression and Random Forest Regression techniques are developed and evaluated using statistical performance indicators. The research aims to improve estimation accuracy, minimize operational delays, and support efficient resource management. Historical migration records are collected and preprocessed to ensure data quality and consistency. The selected input features include DataSize_GB, Notables, and Complexity Score, while MigrationTime_Hours is considered the target variable. Both algorithms are trained and tested using evaluation measures such as R^2, RMSE, MAE, and MSE to determine their predictive capabilities. The experimental outcomes indicate that Random Forest Regression achieves excellent performance on training data by identifying complex nonlinear patterns. However, Linear Regression provides slightly better results on testing data, suggesting stronger generalization to unfamiliar observations. Overall, the findings establish that machine learning techniques can support dependable migration time estimation, while highlighting the need to balance predictive accuracy, model complexity, and generalization for practical implementation.*

Keywords: *Data Migration, Migration Time Prediction, Linear Regression, Random Forest Regression, Machine Learning, Database Systems.*

1. INTRODUCTION

This analysis indicates that research on cloud migration is still in its infancy; however, the evidence gathered indicates that its maturity level is steadily advancing. The findings also highlight the need to establish a standardized framework to effectively guide the migration process. Their study identifies the fundamental principles governing cloud migration processes and highlights the differences between various cloud deployment models. In addition, they introduce a process-oriented framework to support and streamline the migration approach. [1] In addition to addressing the increasing cloud migration demand, it is an equally important need to develop and study a research framework that ensures secure and reliable cloud migration practices. This research primarily focuses on cloud migration, not including service-oriented architecture (SOA) migration. While recent studies have focused heavily on SOA migration, Research that focuses on cloud migration in particular is very lacking. Each participant had a thorough understanding of cloud computing, its different service models, associated technologies, and useful experience working within cloud environments. In their professional roles, they were actively involved in teams responsible for migrating a variety of applications to the cloud, thus giving them direct, hands-on exposure to real-world migration processes. [2] A total of a characterization approach was used to examine twenty-one papers on cloud migration that included factors such as type of contribution, evaluation approach, migration method, migration type, specific migration tasks, migration objectives, tool support, and associated constraints. Insufficient attention has been paid to aspects such as domain independence, validation, and generalizability. Therefore, this study aims to establish well-defined analytical criteria for inclusion in an evaluation framework, which enables a more systematic and comprehensive evaluation of cloud migration approaches. [3] The Benefits and Risks spreadsheet serves as an initial reference for risk assessment, detailing the advantages and possible drawbacks of implementing IaaS cloud solutions from an enterprise standpoint, including organizational, legal, security, technical, and financial aspects. Two case studies—one involving a technological system run by a small team and the other representing a large-scale enterprise—were used to assess the tools' efficacy organization. Cloud migration methodologies outline a structured set of activities designed to plan, execute, and evaluate the migration process. Existing approaches have been developed to adapt migration strategies based on these specific context requirements to address the contextual

Copyright@ REST Publisher

environment of applications—such as their security, performance, and availability requirements. [4] Our research has focused on identifying the fundamental processes involved in cloud migration. Through this investigation, we found significant IaaS, PaaS, and SaaS cloud deployment models differ from one another. Model-specific migration procedures serve as a representation of these distinctions each derived from a list of common migration activities that serve as a unified foundation for comparison and implementation. [5] The complexity of migrating a web server to the cloud can be effectively reduced by using a decision support system (DSS). Such a system improves the quality of cloud infrastructure service and virtual machine (VM) image selections, ensuring optimal migration outcomes. These selection processes can be modeled as multi-criteria decision-making problems, where multiple alternatives are systematically evaluated to determine the most suitable options for a successful cloud migration. [6] During this phase, key stakeholders should be involved, including the infrastructure team, cloud security team, developers, contractors, and all employees directly involved in the cloud migration project. Additionally, a critical task during this phase is to develop a comprehensive migration plan, outlining responsibilities, timelines, security considerations, and resource requirements, to ensure a smooth and well-coordinated transition to the cloud environment. [7] Accordingly, the main goal of this article is to create and assess a single metamodel that incorporates and unifies the common process components of cloud migration. This metamodel is designed to facilitate the creation, standardization, and sharing of context-specific cloud migration models. Through an extensive literature review, key common concepts and activities were identified and integrated into a unified process metamodel, which was then evaluated and refined using real-world industry cloud migration examples to ensure its practicality and applicability. [8] A conceptual-level model that focuses on identifying core domain concepts and their relationships can serve as a critical foundation for building, representing, and maintaining customized cloud migration models. Such a model enables a comprehensive understanding of the migration domain, facilitating better organization, consistency, and adaptability in managing diverse cloud migration scenarios. [9] The clear advantages of cloud computing, including flexibility, scalability, cost effectiveness, and improved accessibility, have made cloud adoption and migration increasingly attractive to organizations looking to improve their operational efficiency and technological capabilities. [10] A systematic approach in the form of a cloud migration framework was put forth by Nussbaumer and Xiaodong with the goal of analyzing and promoting cloud adoption for small and medium-sized enterprises (SMEs). The framework provides a structured approach to guide SMEs through the decision-making, planning, and implementation phases of migrating to the cloud. [11] The process begins by gathering and preparing relevant information you in making well-informed decisions about moving to the cloud. At this point, At this stage, any identified barriers or challenges should be carefully analyzed and managed – either by developing appropriate solutions or strategically avoiding them to ensure a smooth and effective migration process. [12] This approach promotes a comprehensive understanding of cloud migration within the financial sector. By combining qualitative data from various sources, this methodology effectively captures both broad industry trends and specific technology insights, enabling a well-rounded assessment of migration strategies and their impacts on financial environments. [13] We propose a framework that helps organizations conduct a structured feasibility study to determine whether migrating to the cloud is a viable and beneficial option. If deemed appropriate, the framework further guides organizations in developing an effective cloud migration strategy, outlining the optimal approach, resources, and processes required for a successful transition [14].

2. MATERIAL AND METHODS

Materials: DataSize_GB: The DataSize_GB parameter represents the total size of the data stored in the Netezza environment that needs to be migrated to the Azure cloud. It is measured in gigabytes (GB) and serves as a critical input variable because the overall data volume directly affects migration time, storage requirements, and transfer strategies. Larger datasets typically require more bandwidth, extended data extraction periods, and complex optimization techniques to minimize downtime during migration. Therefore, this parameter provides an essential quantitative measure of workload magnitude in the migration process. NumTables: The NumTables parameter indicates the total number of individual tables contained within the Netezza database that are subject to migration. This value reflects the logical complexity and data segmentation of the source system. A higher number of tables often implies more schema objects, dependencies, and relationships that must be accurately transformed and validated within the Azure environment. Consequently, as the number of table's increases, additional time and resources are required for schema mapping, validation, and performance tuning after migration. ComplexityScore: The Complexity Score parameter quantifies the overall technical and structural complexity of the migration project. This score can be assigned based on several qualitative and quantitative factors, such as data transformations, inter-table dependencies, procedural logic (e.g., stored procedures, triggers), and custom scripts. It typically ranges from 1.0 (very simple) to 10.0 (extremely complex). Higher complexity scores represent migrations that demand extensive re-engineering, testing, and validation to ensure compatibility with Azure services. Thus, this parameter serves as an important predictor of both effort and duration. MigrationTime_Hours: The MigrationTime_Hour's parameter represents the total estimated or observed time required to complete the migration from Netezza to Azure, expressed in hours. It serves as the output variable in the dataset and depends on multiple factors, including data volume, number of tables, and migration complexity. The value encompasses all key stages of the migration process — data extraction, transformation, transfer, validation, and deployment in the

Azure environment. Analyzing this metric helps project managers and engineers predict timelines, identify performance bottlenecks, and optimize migration strategies for future projects.

3. MACHINE LEARNING ALGORITHMS

Linear Regression: Linear Regression is a fundamental statistical technique used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It assumes a linear relationship, independence of errors, homoscedasticity, and normal distribution of residuals. Due to its simplicity and interpretability, Linear Regression is widely applied in prediction, trend analysis, and explanatory modeling. However, its performance may be limited when dealing with complex, non-linear relationships or high-dimensional datasets.

Random Forest Regression: Random Forest Regression is an ensemble learning method that constructs multiple decision trees during training and produces predictions by averaging the outputs of individual trees. It improves predictive accuracy and controls overfitting by introducing randomness through bootstrap sampling and random feature selection. Random Forest Regression is capable of handling non-linear relationships, large datasets, and interactions among variables effectively. Although it offers high accuracy and robustness, the model is less interpretable compared to traditional regression techniques and may require higher computational resources.

4. RESULT AND DISCUSSION

TABLE 1. Descriptive Statistics

	DataSize_GB	NumTables	ComplexityScore	MigrationTime_Hours
count	500	500	500	500
mean	306.293	198.368	5.6616	229.0664
std	292.2064	117.2229	2.640737	113.481
min	6.52	1	1	1
25%	87.835	97.75	3.4225	140.6
50%	220.95	204	5.76	233.03
75%	428.3275	302	7.96	316.7775
max	1492.05	400	9.99	555.21

Table 1 Shows 500 migration records, each representing a unique Netezza-to-Azure migration scenario characterized by data size, number of tables, migration complexity, and the resulting migration time in hours. The DataSize_GB parameter ranges from 6.52 GB to 1,492.05 GB, with an average size of approximately 306.29 GB. The relatively high standard deviation (292.21) indicates substantial variability in data volumes among projects, reflecting a mix of small, medium, and large-scale migrations. The median value (220.95 GB) suggests that most projects are moderately sized, while the maximum value represents a few outlier projects involving massive data volumes. The NumTables parameter, which denotes the total number of tables migrated, varies between 1 and 400, with a mean of 198.37 tables and a standard deviation of 117.22. This distribution implies that migration projects differ widely in structural complexity, from very simple databases with only a few tables to highly complex systems with hundreds of interlinked tables.

Copyright@ REST Publisher

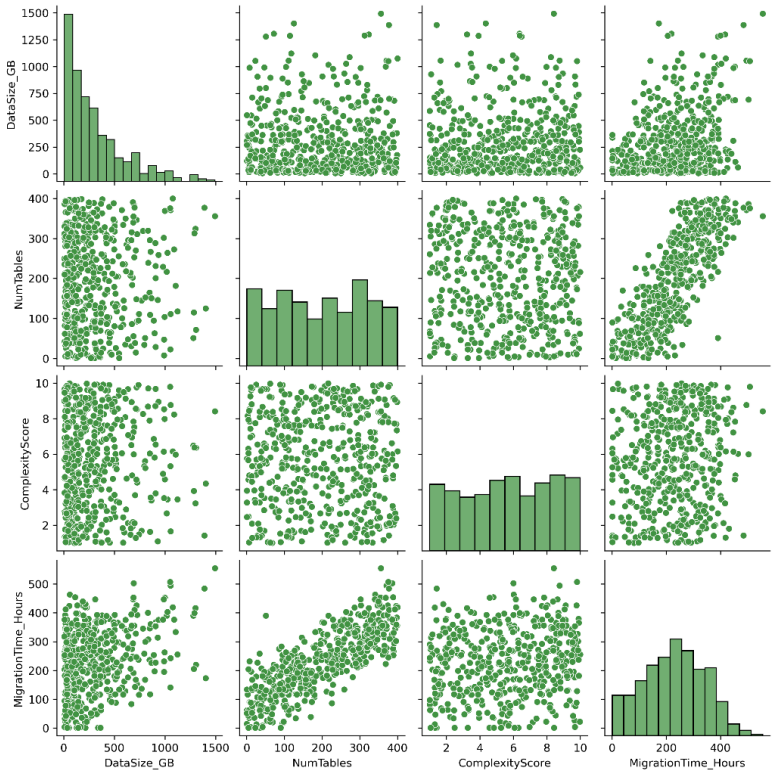


FIGURE 1. Pair Plot of Migration Dataset Variables

Figure 1 illustrates the pairwise relationships among the four key numerical parameters in the Netezza-to-Azure cloud migration dataset: DataSize_GB, NumTables, ComplexityScore, and MigrationTime_Hours. The diagonal plots display the distribution of each variable, while the off-diagonal scatter plots reveal the correlations between them. The distribution of DataSize_GB is notably right-skewed, indicating that most migration projects involve relatively smaller datasets, with a few instances of very large data volumes. The NumTables variable appears uniformly distributed, suggesting a diverse range of database structures across projects. ComplexityScore values are evenly spread, reflecting balanced representation from simple to highly complex migrations.

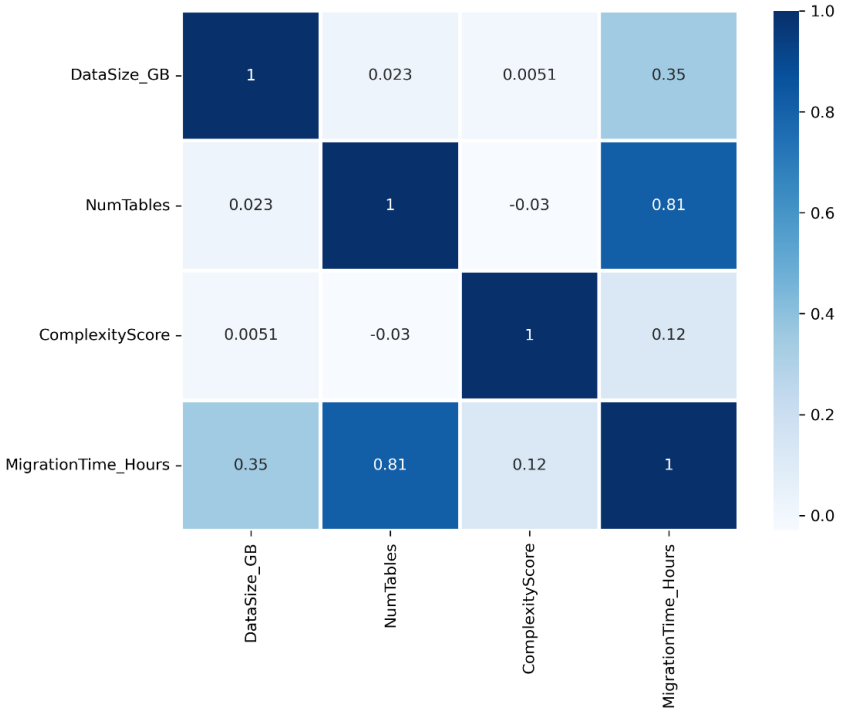


FIGURE 2. Correlation Heatmap of Migration Dataset Variables

Copyright@ REST Publisher

Figure 2 presents a correlation heatmap illustrating the linear relationships among the four key numerical variables: DataSize_GB, NumTables, ComplexityScore, and MigrationTime_Hours. The color intensity represents the strength and direction of the correlation coefficients, with darker shades indicating stronger relationships. From the heatmap, it is evident that MigrationTime_Hours has a strong positive correlation with NumTables (r = 0.81), suggesting that the number of tables migrated is the most influential factor in determining overall migration duration. This indicates that projects involving a higher number of tables tend to require significantly longer migration times, likely due to increased schema complexity, data validation efforts, and dependency handling.

Linear Regression

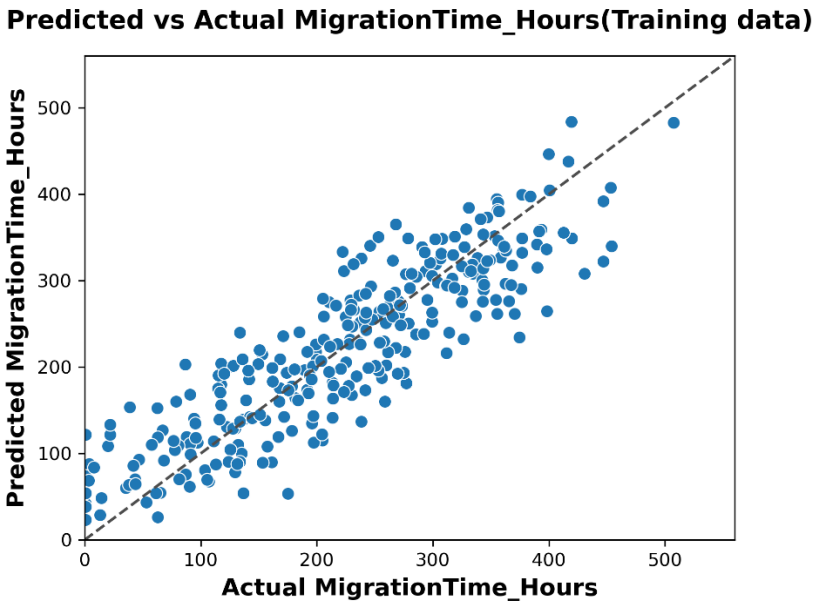


FIGURE 3. Predicted vs. Actual Migration Time (Training Data)

Figure 3 illustrates the relationship between the actual migration time (in hours) and the corresponding values predicted by the regression model using the training dataset. Each point represents an individual observation, while the dashed diagonal line denotes the ideal scenario where predicted values exactly match the actual values. The clustering of data points around this reference line indicates a strong agreement between predictions and observed migration times, demonstrating that the model effectively captures the underlying trend in the training data. Minor deviations from the diagonal suggest the presence of prediction errors, particularly at higher migration times.

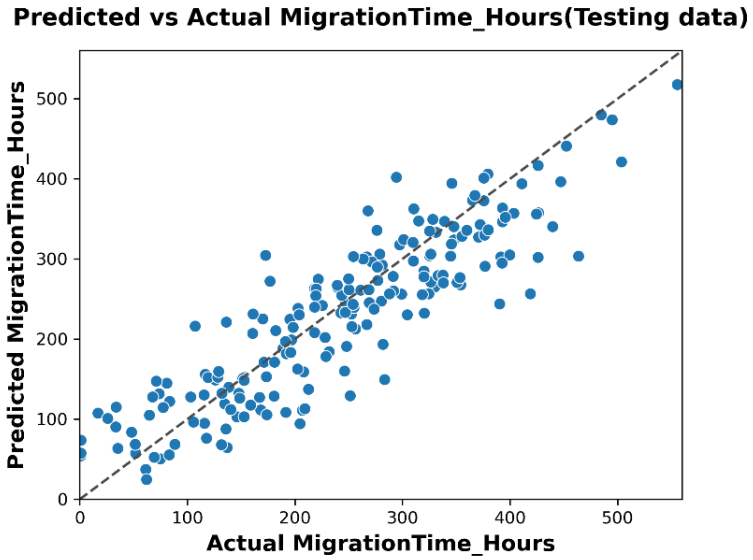


FIGURE 4. Predicted vs. Actual Migration Time (Testing Data)

Copyright@ REST Publisher

Figure 4 presents a comparison between the actual migration time and the values predicted by the regression model on the testing dataset. The dashed diagonal line represents the ideal one-to-one correspondence between predicted and actual values. The dispersion of points around this line indicates that the model generalizes well to unseen data, maintaining a strong predictive relationship beyond the training phase. Although slightly higher variability is observed compared to the training results, particularly at larger migration times, the overall alignment with the reference line demonstrates satisfactory predictive accuracy and robustness of the model on the testing dataset.

Random Forest Regression

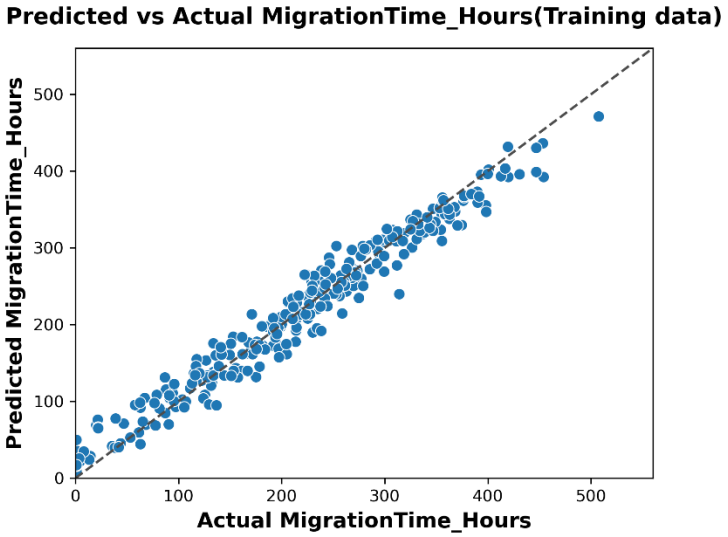


FIGURE 5. Predicted vs. Actual Migration Time (Training Data)

Figure 5 illustrates the comparison between actual migration time and the corresponding values predicted by the regression model for the training dataset. The scatter plot shows a strong linear alignment of data points along the diagonal reference line, indicating a high level of agreement between predicted and observed values. This close clustering suggests that the model effectively learns the underlying relationship within the training data and exhibits minimal prediction error. Minor deviations from the ideal line, particularly at higher migration times, reflect natural variability in the data but do not significantly affect the overall goodness of fit, confirming the robustness and reliability of the model during training.

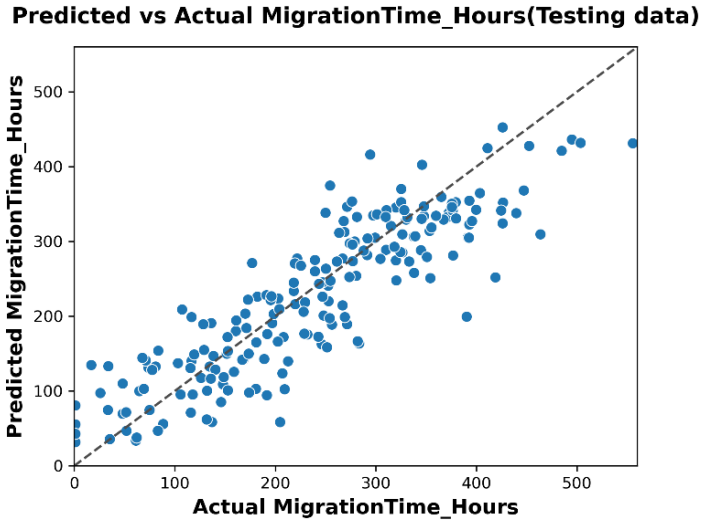


FIGURE 6. Predicted vs. Actual Migration Time (Testing Data)

Figure 6 depicts the relationship between actual migration time and the values predicted by the regression model on the testing dataset. The scatter plot shows that most data points are distributed close to the diagonal reference line, indicating a strong correspondence between predicted and observed migration times. Compared to the training results, a slightly

wider dispersion is evident, particularly at higher migration time values, reflecting the model's exposure to unseen data. Nevertheless, the overall alignment with the ideal line confirms that the model maintains good generalization capability and delivers reliable predictive performance on the testing dataset.

TABLE 2. Model Performance Comparison on Training Dataset

Data	Symbol	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
Train	LR	0.780322	0.780322	2675.08	51.72118	41.6357	139.957	0.453235	34.72351
Train	RFR	0.961356	0.961357	470.582	21.6929	17.13041	73.77045	0.216974	14.05978

The performance metrics for the training dataset indicate a clear difference between Linear Regression (LR) and Random Forest Regression (RFR). The Linear Regression model achieves an and Explained Variance Score (EVS) of 0.7803, suggesting that it explains approximately 78% of the variance in migration time. However, the relatively high error values—MSE (2675.08), RMSE (51.72), MAE (41.64), and Max Error (139.96)—indicate moderate prediction inaccuracies, reflecting the model's limited ability to capture complex patterns in the data. In contrast, the Random Forest Regression model demonstrates substantially superior performance, with an and EVS of 0.9614, explaining over 96% of the variance. This improvement is further supported by significantly lower error metrics, including MSE (470.58), RMSE (21.69), MAE (17.13), and Max Error (73.77). Additionally, lower MSLE and Median Absolute Error (MedAE) values confirm that RFR provides more accurate and stable predictions on the training data, highlighting its effectiveness in modeling non-linear relationships.

TABLE 3. Model Performance Comparison on Testing Dataset

Data	Symbol	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
Test	LR	0.789655	0.799859	2876.971	53.6374	42.74929	162.2449	0.348175	35.91326
Test	RFR	0.765446	0.774386	3208.076	56.63988	45.03803	190.8785	0.339038	35.99655

The evaluation results on the testing dataset show comparable performance between Linear Regression (LR) and Random Forest Regression (RFR), with a slight advantage observed for the Linear Regression model. The LR model attains an value of 0.7897 and an Explained Variance Score (EVS) of 0.7999, indicating that it explains nearly 79–80% of the variance in migration time on unseen data. The associated error metrics—MSE (2876.97), RMSE (53.64), MAE (42.75), and Max Error (162.24)—suggest a moderate level of prediction accuracy, with relatively stable performance across test samples.

5. CONCLUSION

This study evaluated the effectiveness of Linear Regression (LR) and Random Forest Regression (RFR) for predicting migration time using multiple performance metrics. The results from the training phase demonstrated that RFR significantly outperformed LR, achieving higher and EVS values and substantially lower error measures, which confirms its strong capability to capture complex and non-linear relationships in the data. In contrast, LR showed moderate performance, reflecting its limitation in modeling intricate patterns. However, testing results revealed a different trend. While both models maintained reasonably good predictive accuracy on unseen data, Linear Regression slightly outperformed Random Forest Regression in terms of generalization, with lower error values and marginally higher. This indicates that RFR, despite its superior training performance, experienced mild overfitting. Overall, the findings suggest that Random Forest Regression is highly effective when model accuracy on training data is prioritized, whereas Linear Regression offers more stable and consistent performance on unseen data. These insights highlight the importance of balancing model complexity and generalization when selecting predictive models for migration time estimation.

REFERENCES

[1]. Jamshidi, Pooyan, Aakash Ahmad, and Claus Pahl. "Cloud migration research: a systematic review." IEEE transactions on cloud computing 1, no. 2 (2013): 142-157.

[2]. Zhao, Jun-Feng, and Jian-Tao Zhou. "Strategies and methods for cloud migration." international Journal of Automation and Computing 11, no. 2 (2014): 143-152.

[3]. Rai, Rashmi, Gadadhar Sahoo, and Shabana Mehfuz. "Exploring the factors influencing the cloud computing adoption: a systematic study on cloud migration." Springer Plus 4, no. 1 (2015): 197.

[4]. Gholami, Mahdi Fahmideh, Farhad Daneshgar, Graham Low, and Ghassan Beydoun. "Cloud migration process—A survey, evaluation framework, and open challenges." Journal of Systems and Software 120 (2016): 31-69.

[5]. Khajeh-Hosseini, Ali, Ian Sommerville, Jurgen Bogaerts, and Pradeep Teregowda. "Decision support tools for cloud migration in the enterprise." In 2011 IEEE 4th International Conference on Cloud Computing, pp. 541-548. IEEE, 2011.

[6]. Pahl, Claus, Huanhuan Xiong, and Ray Walshe. "A comparison of on-premise to cloud migration approaches." In European Conference on Service-Oriented and Cloud Computing, pp. 212-226. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[7]. Pahl, Claus, Huanhuan Xiong, and Ray Walshe. "A comparison of on-premise to cloud migration approaches." In European Conference on Service-Oriented and Cloud Computing, pp. 212-226. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

Copyright@ REST Publisher

[8]. Alharthi, Dalal N. "Secure cloud migration strategy (SCMS): A safe journey to the cloud." In International Conference on Cyber Warfare and Security, pp. 1-6. Academic Conferences International Limited, 2023.
[9]. Fahmideh, Mahdi, Farhad Daneshgar, Fethi Rabhi, and Ghassan Beydoun. "A generic cloud migration process model." European Journal of Information Systems 28, no. 3 (2019): 233-255.
[10]. Odun-Ayo, Isaac, Frank Agono, and Sanjay Misra. "Cloud migration: Issues and developments." In Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1. 2018.
[11]. Hosseini Shirvani, Mirsaeid, Gholam R. Amin, and Sara Babaeikiadehi. "A decision framework for cloud migration: A hybrid approach." IET software 16, no. 6 (2022): 603-629.
[12]. Jamshidi, Pooyan, Claus Pahl, and Nabor C. Mendonça. "Pattern-based multi-cloud architecture migration." Software: Practice and Experience 47, no. 9 (2017): 1159-1184.
[13]. Brahmandam, Balajee Asish. "Cloud Migration and Hybrid Infrastructure in Financial Institutions." International Journal of Computer Science Engineering Techniques 9, no. 1 (2025): 42-46.
[14]. Ahmed, Monjur, and Navjot Singh. "A framework for strategic cloud migration." In Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence, pp. 160-163. 2019.
[15]. Su, Xiaogang, Xin Yan, and Chih-Ling Tsai. "Linear regression." Wiley Interdisciplinary Reviews: Computational Statistics 4, no. 3 (2012): 275-294.
[16]. Tranmer, Mark, and Mark Elliot. "Multiple linear regression." The Cathie Marsh Centre for Census and Survey Research (CCSR) 5, no. 5 (2008): 1-5.
[17]. Segal, Mark R. "Machine learning benchmarks and random forest regression." (2004).
[18]. Borup, Daniel, Bent Jesper Christensen, Nicolaj Søndergaard Mühlbach, and Mikkel Slot Nielsen. "Targeting predictors in random forest regression." International Journal of Forecasting 39, no. 2 (2023): 841-868.
[19]. Jog, Amod, Aaron Carass, Snehashis Roy, Dzung L. Pham, and Jerry L. Prince. "Random forest regression for magnetic resonance image synthesis." Medical image analysis 35 (2017): 475-488.
[20]. Singh, Balraj, Parveen Sihag, and Karan Singh. "Modelling of impact of water quality on infiltration rate of soil by random forest regression." Modeling Earth Systems and Environment 3 (2017): 999-1004.
[21]. Cootes, Tim F., Mircea C. Ionita, Claudia Lindner, and Patrick Sauer. "Robust and accurate shape model fitting using random forest regression voting." In Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12, pp. 278-291. Springer Berlin Heidelberg, 2012.

Copyright@ REST Publisher