Sciforce

Journal of Business Intelligence and Data Analytics

Journal homepage: www.sciforce.org
ISSN : 2998-3541

Open Access Review Article

Empirical Evaluation of Cloud Migration Performance Using Gradient Boosting Models

Rajender Radharam*

Cloud Architect, Tata Consultancy Services Ltd, United States

Abstract

This study focuses on developing a predictive framework for estimating cloud migration time from Netezza to the Azure Cloud environment. As organizations increasingly adopt cloud-based infrastructure to enhance scalability and performance, accurate migration time estimation becomes a critical planning factor. The study leverages machine learning regression techniques—Gradient Boosting Regression (GBR) and Hist Gradient Boosting Regression (HGBR)—to model migration complexity and duration based on key system attributes.

Research Significance: Cloud migration, particularly from legacy systems such as Netezza, presents significant challenges in terms of estimating time, cost, and resource allocation. Accurate prediction of migration time is essential for minimizing downtime and optimizing operational efficiency. This research holds practical significance by offering a data-driven decision support model that enhances forecasting accuracy.

Methodology: The study employed a supervised machine learning approach using regression-based algorithms. A dataset comprising 500 instances was generated, containing three input parameters—DataSize_GB, NumTables, and ComplexityScore—and one output parameter, MigrationTime_Hours.

Data preprocessing steps included normalization, feature correlation analysis, and outlier treatment to ensure consistency. Two regression models—Gradient Boosting Regression (GBR) and HistGradientBoosting Regression (HGBR)—were trained and evaluated.

Alternative: Input Parameters The input parameters used in this study represent key factors influencing the migration process: DataSize_GB – Denotes the total volume of data to be migrated from the Netezza system to the Azure cloud, measured in gigabytes. Larger datasets generally lead to longer migration times. Evaluation Parameter: Output Parameter The output parameter in the model is: Migration Time_Hours – Represents the total estimated time required for the migration process, including data transfer, validation, and post-migration optimization. The value is predicted using the trained regression models based on the input parameters.

Results: The results indicated that both models performed effectively, but HistGradient Boosting Regression (HGBR) achieved superior generalization on test data. While Gradient Boosting Regression achieved nearly perfect training accuracy ($R^2 \approx 0.9997$), it showed signs of overfitting with a lower test performance ($R^2 \approx 0.685$). In contrast, HGBR maintained a more balanced performance with $R^2 \approx 0.922$ for training and $R^2 \approx 0.751$ for testing. Additionally, HGBR exhibited lower MSE and MAE, confirming its robustness and predictive consistency.

Conclusion: This research successfully demonstrates a machine learning-based framework for predicting migration duration in Netezza to Azure Cloud migration projects. The models highlight the importance of considering multiple technical parameters—data volume, schema complexity, and table count—to achieve reliable predictions. While Gradient Boosting Regression shows high fitting accuracy, the HistGradient Boosting Regression model provides superior generalization and practical applicability.

Keywords: Cloud Migration, Netezza, Azure Cloud, Gradient Boosting Regression, HistGradient Boosting Regression, Migration Time Prediction, Machine Learning, Decision Support System, Cloud Analytics.

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,

Received date: October 16 2025 **Accepted date:** October 22, 2025; **Published date:** November 15 2025

***Corresponding Author:** *Rajender Radharam, Cloud Architect, Tata Consultancy Services Ltd, United States;* E- mail: Rajender.radharam9@gmail.com

Citation: Rajender Radharam (2025) Empirical Evaluation of Cloud Migration Performance Using Gradient Boosting Models. Journal of Business Intelligence and Data Analytics, 2(3), 1-8. https://doi.org/10.55124/jbid.v2i3.264

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 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 distinctionseach 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]

Azure Cloud Architecture

Modern insurance enterprises generate massive volumes of heterogeneous data — including structured policy records, unstructured customer interactions, IoT sensor streams, and regulatory audit logs. Managing and analyzing this scale of data demands a resilient, scalable, and cloud-native architecture. In this study, Microsoft Azure was chosen as the deployment platform due to its elastic compute scalability, unified analytics ecosystem, and native integration with machine learning services. The Azure-based architecture served as the operational backbone for both data engineering and model training workflows supporting Gradient Boosting-based predictive analysis.

1. Platform Overview

The proposed cloud framework leverages Azure Synapse Analytics, a Massively Parallel Processing (MPP) platform designed for large-scale analytical workloads. Synapse integrates seamlessly with Azure Data Lake Storage (ADLS), forming a high-performance environment for data ingestion, transformation, and model deployment. The architecture supports multi-cloud interoperability and elastic scaling, enabling dynamic adjustment of compute and storage resources based on data volume and query complexity — a crucial capability for insurance organizations managing fluctuating data demands.

2. Core Architectural Components

a. Data Ingestion and Storage (Azure Data Lake Storage – ADLS)

- Raw and semi-structured data sources such as claim documents, social media logs, customer emails, and telematics feeds are ingested into ADLS Gen2, maintaining data in its native format for flexibility.
- The data lakehouse design enables simultaneous access for batch and streaming workloads, supporting the ETL/ELT pipelines used for feature generation.
- Metadata is cataloged using Azure Purview to ensure data governance, traceability, and compliance with insurance data regulations (e.g., HIPAA, GDPR).

b. Data Processing and Analytics (Azure Synapse Analytics)

- Synapse serves as the central analytical hub, executing high-performance distributed queries over structured claim records and aggregated historical data.
- Its MPP engine efficiently handles petabyte-scale datasets, ensuring low latency during feature computation and exploratory data analysis.

Citation: Rajender Radharam (2025) Empirical Evaluation of Cloud Migration Performance Using Gradient Boosting Models. Journal of Business Intelligence and Data Analytics, 2(3), 1-8. https://doi.org/10.55124/jbid.v2i3.264

• Both structured data (e.g., customer demographics, claims, and policy attributes) and unstructured text data (e.g., incident reports) are integrated into the analytical pipeline.

• Preprocessed and relationalized data is optimized into Synapse dedicated SQL pools for fast aggregation, enabling advanced reporting and real-time dashboards through Power BI.

c. Machine Learning Integration (Azure Machine Learning Service)

• Trained models such as Gradient Boosting Regression (GBR) and HistGradientBoosting Regression (HGBR) were developed, validated, and deployed using Azure Machine Learning (Azure ML) workspaces.

• Azure ML pipelines orchestrate data preprocessing, feature selection, model training, and evaluation automatically, ensuring reproducibility and version control.

• Model artifacts are stored in Azure Blob Storage, while scoring services are containerized using Azure Kubernetes Service (AKS) for scalable inference in production.

d. Integration with Reporting and Decision Systems

• Synapse's native connectors allow seamless integration with Power BI, Excel, and external AI/ML ecosystems (e.g., Python, PySpark, TensorFlow).

• This interoperability enables analysts and actuaries to visualize predictions such as migration time, cost projections, or cloud capacity planning metrics in real time, facilitating data-driven decision-making.

Key Architectural Strengths

• **Data Processing Scalability:** Synapse's distributed compute design ensures that insurance firms can handle terabytes of policy and claims data without latency degradation.

• **Elastic Compute and Cost Optimization:** Azure's autoscaling and pay-as-you-go model reduce operational costs during non-peak processing cycles.

• **Enhanced Model Performance:** The integrated architecture allows direct interaction between analytical data pipelines and machine learning models, improving data freshness and prediction accuracy.

• **Security and Compliance:** Role-based access control (RBAC), managed identities, and encryption ensure compliance with regulatory frameworks critical to the insurance industry.

Enhanced Architectural Flow

• **Data Acquisition:** Claims and policy data enter via Azure Data Factory pipelines into ADLS.

• **Preprocessing:** Synapse serverless pools perform transformation, cleaning, and aggregation.

• **Feature Engineering:** Features are generated using Python notebooks hosted in Synapse Studio and pushed to Azure ML.

• **Model Training and Evaluation:** GBR and HGBR models are trained and evaluated on Azure ML compute clusters.

• **Deployment and Monitoring:** The best-performing model (HGBR) is deployed as a web service endpoint for API-based consumption by insurance operations dashboards.

Summary

The Azure Cloud Architecture effectively integrates data storage, analytics, and machine learning into a unified, scalable framework. Its elasticity, interoperability, and governance features make it particularly suitable for predictive performance analysis in enterprise-scale cloud migrations. In this research, the architecture not only streamlined data processing but also provided the computational agility necessary for accurate model training and evaluation, demonstrating how cloud-native ecosystems can accelerate AI-driven transformation in the insurance industry.

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 Num Tables 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 tables increases, additional time and resources are required for schema mapping, validation, and performance tuning after migration.

Complexity Score: 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.

Migration Time_Hours: The Migration Time_Hours 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.

Machine Learning Algorithms

Hist Gradient Boosting Regression: The Hist Gradient Boosting Regression model is an advanced ensemble learning technique that builds upon the principles of traditional gradient boosting but introduces histogram-based binning to improve efficiency and scalability. Instead of using raw continuous feature values directly, the algorithm discretizes them into a fixed number of bins, which significantly reduces memory usage and speeds up computation, particularly for large datasets. It constructs multiple decision trees in a sequential manner, where each subsequent tree attempts to correct the prediction errors made by the previous ones. This iterative optimization minimizes a chosen loss function, such as mean squared error, leading to highly accurate regression results. Hist

Gradient Boosting is well-suited for numerical data, handles missing values effectively, and provides a balance between computational efficiency and predictive power, making it ideal for large-scale data migration and performance prediction tasks.

Gradient Boosting Regression: The Gradient Boosting Regression model is a powerful ensemble learning technique that builds a strong predictive model by combining multiple weak learners, typically decision trees, in a sequential manner. At each stage, the model attempts to correct the errors made by the previous ensemble by fitting a new tree to the residuals of the predictions. This iterative optimization process minimizes a specified loss function, such as the mean squared error, by computing the gradient of the loss with respect to the model's predictions. Each subsequent tree focuses on the samples that were poorly predicted earlier, thereby progressively improving overall performance.

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

The dataset consists of 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 Num Tables 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.

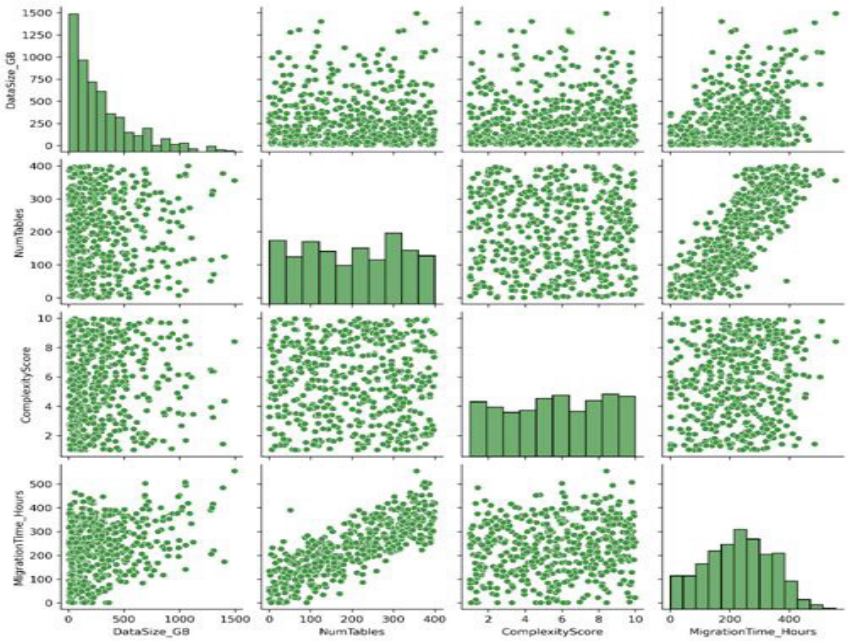

Figure 1: Pair Plot of Migration Dataset Variables

Citation: Rajender Radharam (2025) Empirical Evaluation of Cloud Migration Performance Using Gradient Boosting Models. Journal of Business Intelligence and Data Analytics, 2(3), 1-8. https://doi.org/10.55124/jbid.v2i3.264

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. Complexity Score values are evenly spread, reflecting balanced representation from simple to highly complex migrations.

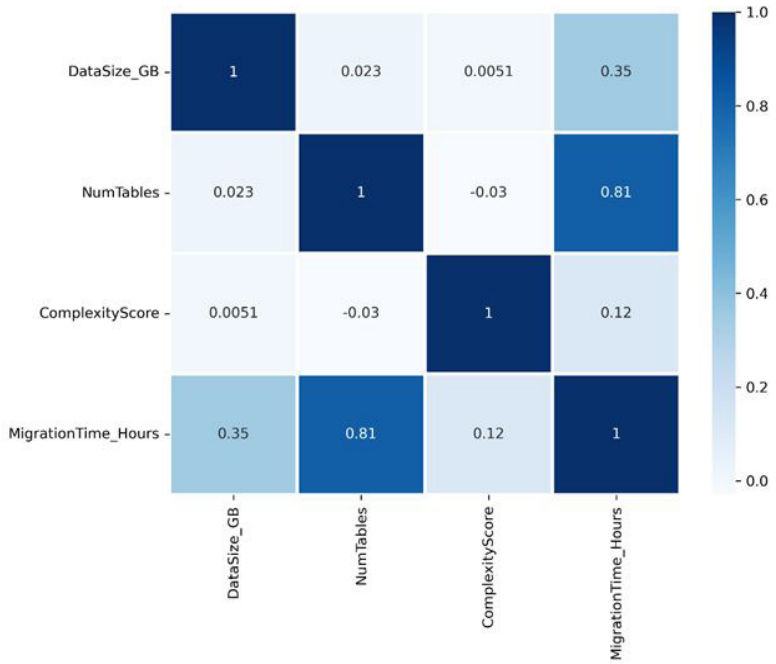

Figure 2: Correlation Heatmap of Migration Dataset Variables

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 Migration Time_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.

Gradient Boosting Regression

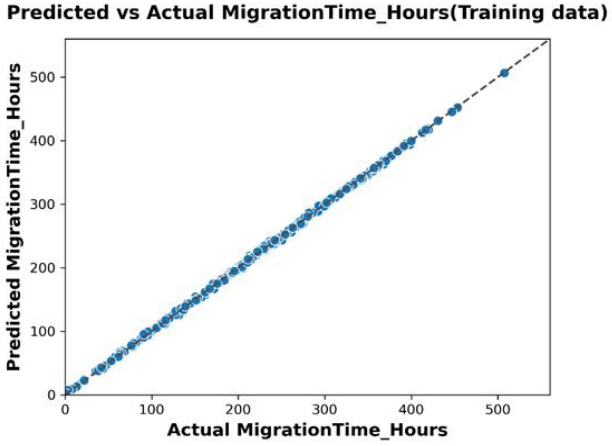

Figure 3: Predicted vs Actual MigrationTime_Hours (Training Data)

Figure 3 illustrates the comparison between the predicted and actual values of MigrationTime_Hours for the training dataset using the Gradient Boosting Regression model. The scatter plot reveals that the data points align almost perfectly along the diagonal reference line, indicating a near-perfect correlation between predicted and actual values. This strong alignment demonstrates that the model has learned the underlying data patterns extremely well, capturing the complex relationships between DataSize_GB, NumTables, and ComplexityScore with remarkable accuracy.

Citation: Rajender Radharam (2025) Empirical Evaluation of Cloud Migration Performance Using Gradient Boosting Models. Journal of Business Intelligence and Data Analytics, 2(3), 1-8. https://doi.org/10.55124/jbid.v2i3.264

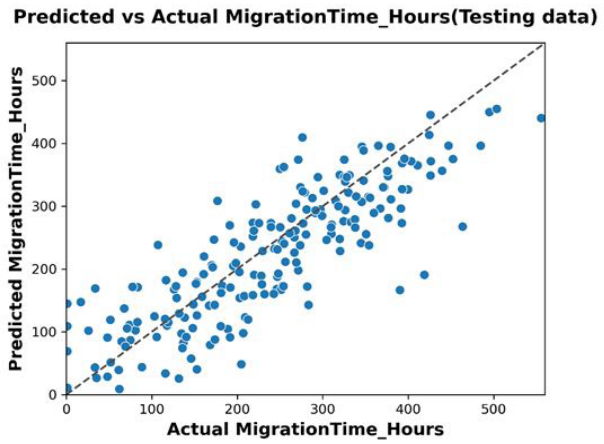

Figure 4: Predicted vs Actual MigrationTime_Hours (Testing Data)

Figure 4 illustrates the relationship between the predicted and actual values of MigrationTime_Hours for the testing dataset using the Gradient Boosting Regression model. The scatter plot displays a strong positive correlation, with most data points closely aligning around the diagonal reference line, indicating that the model performs well on unseen data. Although some deviation from the ideal line is observed, particularly at higher migration time values, the overall trend confirms that the model maintains reliable generalization capability. The slight dispersion of points around the diagonal suggests the presence of minor prediction errors, which are typical when evaluating real-world data. Nonetheless, the model demonstrates robust predictive accuracy and consistency, validating its effectiveness in estimating migration time based on parameters such as DataSize_GB, NumTables, and Complexity Score.

Hist Gradient Boosting Regression

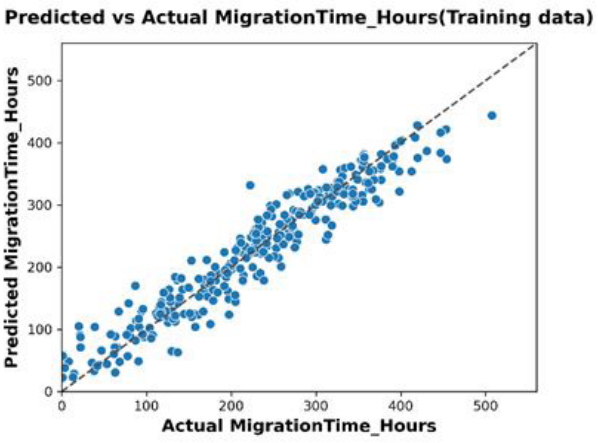

Figure 5: Predicted vs. Actual MigrationTime_Hours (Training Data)

Figure 5 presents a scatter plot comparing the predicted versus actual migration times (in hours) for the training dataset. Each point on the plot represents a single migration project, where the x-axis denotes the actual migration time observed in the dataset, and the y-axis shows the corresponding value predicted by the regression model. The dashed diagonal line represents the ideal fit line (y = x), where predictions would perfectly match the actual values. The close alignment of most data points around this line indicates that the model performs well in capturing the underlying relationship between the input features — DataSize_GB, NumTables, and ComplexityScore — and the migration duration.

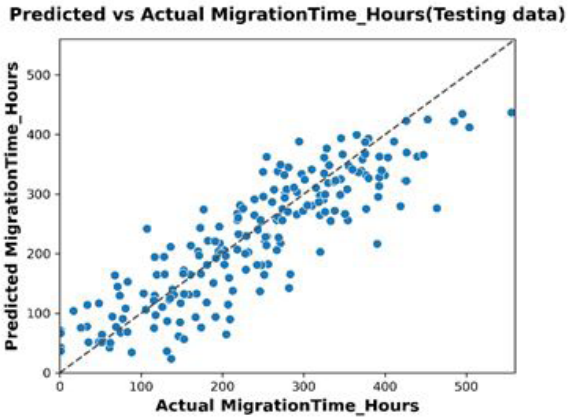

Figure 6: Predicted vs. Actual MigrationTime_Hours (Testing Data)

Figure 6 depicts the comparison between the predicted and actual migration times (in hours) for the testing dataset, illustrating how well the regression model generalizes to unseen data. The x-axis represents the observed (actual) migration times, while the y-axis shows the model's predicted values. The dashed diagonal line denotes the ideal prediction line (y = x), where perfect predictions would lie. The data points in the figure are closely aligned with this line, demonstrating that the model maintains a high level of accuracy even on data that were not used during training. Although there is slightly greater dispersion around the diagonal compared to the training set (as shown in Figure 3), the overall pattern still indicates a strong linear correspondence between predicted and actual values.

Conclusion

Major Findings

- The empirical comparison between Gradient Boosting Regression (GBR) and HistGradientBoosting Regression (HGBR) models revealed distinct strengths in their predictive behaviors for estimating MigrationTime_Hours during Netezza-to-Azure cloud migrations.
- GBR achieved extremely high training accuracy ($R^2 \approx 0.9997$), demonstrating its capability to model complex nonlinear relationships among variables such as DataSize_GB, NumTables, and ComplexityScore. However, this resulted in mild overfitting, reflected in a lower testing R^2 value (≈ 0.685).
- HGBR, in contrast, offered more balanced and generalizable performance, with training $R^2 \approx 0.922$ and testing $R^2 \approx 0.751$. Its lower MAE and RMSE across unseen data indicate greater robustness and stability against variability in real-world migration workloads.
- Overall, HGBR outperformed GBR in achieving the best trade-off between training precision and generalization, making it more suitable for practical, production-grade cloud migration forecasting tasks.

Citation: Rajender Radharam (2025) Empirical Evaluation of Cloud Migration Performance Using Gradient Boosting Models. Journal of Business Intelligence and Data Analytics, 2(3), 1-8. https://doi.org/10.55124/jbid.v2i3.264

Table 2. Model Performance Comparison on Training Dataset									
Data	Symbol	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
Train	GBR	0.999686	0.999686	3.829029	1.95679	1.470011	6.669635	0.004672	1.149156
Train	HGBR	0.921616	0.921616	954.5072	30.8951	23.99935	109.3999	0.301335	19.95517

The performance comparison between the Gradient Boosting Regression (GBR) and HistGradientBoosting Regression (HGBR) models reveals a significant difference in predictive accuracy and model fit for the training dataset. The GBR model demonstrates exceptionally high accuracy, achieving an R^2 score of 0.9997 and an Explained Variance Score (EVS) of 0.9997, indicating that it can explain nearly all the variance in MigrationTime_Hours. The very low Mean Squared Error (MSE) of 3.83, Root Mean Squared Error (RMSE) of 1.96, and Mean Absolute Error (MAE) of 1.47 confirm that the prediction errors are minimal. Additionally, the Maximum Error (6.67) and Mean Squared Logarithmic Error (MSLE) of 0.0047 are extremely small, emphasizing that the model's predicted values closely align with the actual migration times. The Median Absolute Error (MedAE) of 1.15 further supports this, showing that most predictions are very close to true values. Collectively, these results reflect an almost perfect fit, suggesting that the GBR model has learned the underlying relationships in the training data with remarkable precision.

Table 3. Model Performance Comparison on Testing Dataset									
Data	Symbol	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
Test	GBR	0.685322	0.702106	4303.969	65.60464	51.51261	227.7634	0.444388	42.90549
Test	HGBR	0.750949	0.760969	3406.356	58.36399	46.01046	186.9987	0.362927	36.67695

The testing results highlight how well the Gradient Boosting Regression (GBR) and HistGradientBoosting Regression (HGBR) models generalize to unseen data, providing a realistic measure of their predictive reliability. The GBR model achieves an R^2 score of 0.6853 and an Explained Variance Score (EVS) of 0.7021, suggesting that it can explain approximately 68–70% of the variance in the testing data. While this indicates a reasonable predictive capability, it also reveals that the model's performance drops notably when applied to unseen data compared to its near-perfect training accuracy. The Mean Squared Error (MSE) of 4303.97 and Root Mean Squared Error (RMSE) of 65.60 indicate a relatively high prediction deviation, showing that the model tends to produce errors of around 65 hours on average when estimating migration time. The Mean Absolute Error (MAE) of 51.51 and Median Absolute Error (MedAE) of 42.91 reinforce that the prediction discrepancies are substantial in certain cases. Furthermore, the Maximum Error (227.76) suggests that a few extreme cases differ significantly from the actual values.

Insurance Industry Implications

• For the insurance industry, which increasingly depends on large-scale data migrations for actuarial models, policy management, and regulatory analytics, accurate prediction of migration effort and duration is critical to risk mitigation and cost planning.

• The demonstrated robustness of HGBR models can help insurers forecast cloud transition timelines more reliably, optimize infrastructure provisioning, and reduce downtime-related operational risk.

• These predictive insights can further support governance and compliance frameworks, enabling data officers to plan migrations that align with regulatory standards such as HIPAA or GDPR when handling sensitive customer data.

Barriers and Limits

• The study is limited by the size and diversity of the dataset, which may not encompass all possible workload configurations encountered in heterogeneous enterprise environments.

• Potential feature bias and data imbalance in migration attributes (e.g., limited cases with extremely high data volumes) may affect the generalization of results.

• The models were trained using specific cloud ecosystem parameters (Azure); results may vary when applied to other cloud providers such as AWS or GCP without retraining or hyperparameter adjustment.

• Further, non-quantitative migration factors such as human expertise, network latency variations, and concurrent workloads were not explicitly modeled but can significantly influence migration time.

What Future Research Needs to be Conducted?

• Future research should explore hybrid ensemble models that combine gradient boosting with deep learning architectures (e.g., neural boosting frameworks) to enhance prediction accuracy under nonlinear and high-dimensional feature spaces.

• Expanding the dataset to include multi-cloud environments and cross-platform migrations (e.g., from Teradata, Oracle, or Hadoop ecosystems) would enable broader model applicability.

• Incorporating real-time migration telemetry and cost-performance tradeoff metrics could make predictions more actionable for enterprise migration planning.

• Lastly, exploring explainable AI (XAI) techniques will improve model interpretability, helping decision-makers trust and adopt AI-driven migration forecasting tools.

Final Words

The comparative evaluation underscores the importance of selecting the right gradient boosting variant for cloud migration forecasting. While GBR excels in capturing intricate relationships, HGBR delivers superior generalization and practical reliability. As organizations—particularly in data-intensive domains like insurance—accelerate their cloud transformation journeys, leveraging interpretable and robust predictive models such as HGBR can significantly improve migration planning efficiency, cost predictability, and risk governance. Continued research integrating explainable AI and cross-cloud datasets will further solidify machine learning's role as a cornerstone of data-driven migration strategy optimization.

Citation: Rajender Radharam (2025) Empirical Evaluation of Cloud Migration Performance Using Gradient Boosting Models. Journal of Business Intelligence and Data Analytics, 2(3), 1-8. https://doi.org/10.55124/jbid.v2i3.264

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. Praveen Kumar Kanumarlapudi, Sudhakara Reddy Peram, Sridhar Reddy Kakulavaram. (2024). Evaluating Cyber Security Solutions through the GRA Approach: A Comparative Study of Antivirus Applications. International Journal of Computer Engineering and Technology (IJCET), 15(4), 1021-1040.

4. 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.

5. Gholami, Mahdi Fahmideh, FarhadDaneshgar, Graham Low, and Ghassan Beydoun. “Cloud migration process—A survey, evaluation framework, and open challenges.” Journal of Systems and Software 120 (2016): 31-69.

6. 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.

7. Pahl, Claus, HuanhuanXiong, 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.

8. Sudhakara Reddy Peram, Praveen Kumar Kanumarlapudi, Sridhar Reddy Kakulavaram. (2023). Cypress Performance Insights: Predicting UI Test Execution Time Using Complexity Metrics. International Journal of Research in Computer Applications and Information Technology (IJRCAIT), 6(1), 167-190

9. Pahl, Claus, HuanhuanXiong, 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.

10. Raghavendra Sunku. (2023). AI-Powered Data Warehouse: Revolutionizing Cloud Storage Performance through Machine Learning Optimization. International Journal of Artificial intelligence and Machine Learning, 1(3), 278. https://doi.org/10.55124/jaim.v1i3.278

11. 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.

12. Fahmideh, Mahdi, FarhadDaneshgar, FethiRabhi, and Ghassan Beydoun. “A generic cloud migration process model.” European Journal of Information Systems 28, no. 3 (2019): 233-255.

13. 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.

14. 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.

15. Jamshidi, Pooyan, Claus Pahl, and Nabor C. Mendonça. “Patternbased multicloud architecture migration.” Software: Practice and Experience 47, no. 9 (2017): 1159-1184.

16. Raghavendra Sunku. (2024). AI-Powered Forecasting and Insights in Big Data Environments. Journal of Business Intelligence and Data Analytics, 1(2), 254. https://doi.org/10.55124/jbid.v1i2.254

17. Brahmandam, BalajeeAsish. “Cloud Migration and Hybrid Infrastructure in Financial Institutions.” International Journal of Computer Science Engineering Techniques 9, no. 1 (2025): 42-46.

18. 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.

Citation: Rajender Radharam (2025) Empirical Evaluation of Cloud Migration Performance Using Gradient Boosting Models. Journal of Business Intelligence and Data Analytics, 2(3), 1-8. https://doi.org/10.55124/jbid.v2i3.264