IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

Cloud Cost Optimization and Forecasting Using Machine Learning

Ms. Suganya RM[1], Santhosh A[2], Sharoon Basha S[3]
Assistant Professor, Dept. of Information Technology[1]
Students, Dept. of Information Technology[2,3]
K.L.N. College of Engineering, Sivaganga, India
suganya.april4@gmail.com,santhoshkarthik1222@gmail.com,sharoonbasha69@gmail.com

Abstract: *Cloud computing offers scalable and flexible infrastructure, but inefficient resource utilization often results in excessive and unpredictable operational costs. Organizations frequently over-provision resources or fail to identify idle services, leading to significant financial waste. This project presents a Cloud Cost Optimization and Forecasting System using machine learning techniques to analyze cloud usage data and reduce unnecessary expenditure. The proposed system uses a Random Forest regression model to predict future cloud costs based on CPU utilization, memory usage, storage, and network activity. Time-series forecasting techniques are applied to analyze cost trends, while resources are classified into optimized, underutilized, and idle categories. Actionable optimization recommendations such as downsizing, scheduling, and termination of idle resources are generated. An interactive dashboard visualizes cost insights, optimization potential, and savings. The system improves cost efficiency, supports proactive decision-making, and promotes sustainable cloud resource utilization.*

Keywords: *Cloud Computing, Cost Optimization, Cost Forecasting, Machine Learning, Random Forest, Resource Utilization, Predictive Analytics, Cloud Dashboard*

I. INTRODUCTION

Cloud computing has transformed modern IT infrastructure by enabling organizations to deploy scalable applications without investing in physical hardware. Leading cloud platforms such as AWS, Microsoft Azure, and Google Cloud Platform provide flexible pay-as-you-use pricing models. However, improper resource allocation and lack of continuous monitoring often result in rising cloud costs. Many organizations provision cloud resources based on peak demand assumptions, causing resources to remain underutilized for long durations. Industry studies reveal that nearly 30–35% of cloud expenditure is wasted due to idle and underutilized resources. Existing cloud billing tools mainly provide historical cost reports and lack predictive capabilities. To address these challenges, this project proposes a machine learning-based cloud cost optimization and forecasting system that predicts future costs, identifies inefficiencies, and recommends optimization actions.

II. RELATED WORK

Several approaches have been proposed for cloud cost management. Rule-based systems rely on static thresholds for CPU and memory utilization, making them simple but ineffective for dynamic workloads. Time-series-based workload prediction models capture trends but require large datasets and struggle with sudden changes in demand.

Recent studies have explored machine learning models such as decision trees and regression techniques for cloud cost

prediction. While these methods improve accuracy, they often lack integrated optimization mechanisms and visualization capabilities. Random Forest models have shown superior performance in handling non-linear relationships and noisy cloud data. However, there remains a need for an end-to-end system that combines prediction, optimization, and visualization, which forms the motivation for this work.

DOI: 10.48175/IJARSCT-32139

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

III. METHODOLOGY

A. Dataset Collection and Preparation

The dataset used in this study was generated and curated to represent realistic cloud usage patterns observed in modern cloud environments. The dataset contains historical cloud resource utilization data collected across multiple service types, including Virtual Machines (VMs), Cloud SQL databases, Cloud Storage services, and BigQuery workloads. Each record in the dataset represents daily usage statistics of an individual cloud resource.

The collected dataset includes key cost-influencing parameters such as CPU utilization percentage, memory utilization percentage, storage consumption (GB), inbound and outbound network traffic (GB), service type, resource tier, and geographical region. These features were selected based on their direct impact on cloud pricing models across major cloud service providers.

To ensure data quality, preprocessing steps were applied, including removal of missing values, normalization of numerical features, and encoding of categorical attributes such as service type, tier, and region. The dataset was split into training and testing subsets using an 80:20 ratio to enable effective learning and evaluation of the machine learning models.

Synthetic data generation techniques were also employed to simulate long-term cloud usage trends over multiple years. This approach ensured balanced representation of idle, underutilized, and optimized resources, which is essential for accurate cost prediction and optimization analysis.

B. Feature Engineering and Cost Modelling

Feature engineering plays a critical role in improving the accuracy of cloud cost prediction. In this work, relevant features influencing cloud expenditure were carefully selected and transformed to enhance model performance. Categorical features such as service type, resource tier, and region were encoded using label encoding techniques to enable compatibility with machine learning algorithms.

Derived features such as daily cost trends and aggregated usage metrics were also incorporated to capture temporal cost behaviour. These engineered features help the model understand complex non-linear relationships between resource utilization and cloud pricing.

A simplified cost model was applied to compute daily cloud costs based on usage metrics, reflecting real-world pricing strategies such as compute-hour charges, storage rates, and network egress costs. This cost modeling approach ensures that the dataset closely resembles real cloud billing data.

C. Cost Prediction Using Machine Learning

A Random Forest regression model was selected as the primary prediction algorithm due to its robustness, ability to handle non-linear relationships, and resistance to overfitting. Random Forest combines multiple decision trees to produce accurate and stable predictions, making it suitable for cloud cost forecasting.

The model was trained using historical cloud usage data with features including CPU utilization, memory consumption, storage usage, and network traffic. Hyperparameters such as the number of estimators and tree depth were tuned experimentally to achieve optimal performance.

The trained model predicts daily and monthly cloud costs for individual resources. Performance evaluation was carried out using metrics such as Mean Absolute Error (MAE) and R^2 score to validate prediction accuracy.

D. Resource Classification and Optimization Strategy

Based on predicted costs and utilization metrics, cloud resources were classified into three categories:

- Idle Resources: Resources with extremely low CPU and memory utilization over extended periods
- Underutilized Resources: Resources operating below optimal capacity
- Optimized Resources: Resources efficiently utilized within recommended thresholds

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 8, March 2026 Impact Factor: 8.2

This classification enables effective identification of inefficiencies within the cloud environment. An optimization engine generates actionable recommendations such as terminating idle resources, downsizing underutilized instances, applying scheduling (start/stop) rules, and optimizing BigQuery workloads using partitioning and clustering techniques. Each recommendation is associated with estimated cost savings, allowing cloud administrators to prioritize optimization actions based on financial impact.

E. Dashboard Visualization and Cost Simulation

To enhance usability and decision-making, the system incorporates an interactive web-based dashboard developed using modern visualization frameworks. The dashboard presents predicted costs, optimized costs, service-wise cost breakdowns, and potential savings in an intuitive manner.

A cost simulation module allows users to modify input parameters such as CPU usage, memory usage, storage size, and network traffic to observe real-time changes in predicted cost. This feature enables proactive planning and evaluation of different deployment scenarios before allocating cloud resources.

The integrated visualization and simulation components ensure transparency, support data-driven decisions, and improve overall cloud cost governance.

IV. SYSTEM ARCHITECTURE

The proposed system follows a multi-tier client–server architecture designed for efficient cloud cost analysis and optimization. The architecture consists of three main layers: (1) the web-based dashboard client, which serves as the presentation layer for dataset upload, visualization, and result display; (2) the backend analytics and machine learning server, which integrates cost prediction, forecasting, and optimization logic; and (3) the trained machine learning models and processed datasets, which form the data layer. Communication between the client and server components is carried out using RESTful HTTPS APIs with JSON serialization, ensuring secure and reliable data exchange. The inference workflow is as follows: the user uploads cloud usage data through the dashboard interface. The backend server preprocesses the data and performs cost prediction using a trained Random Forest regression model based on resource usage metrics such as CPU utilization, memory consumption, storage, and network activity. The system then applies forecasting and optimization rules to classify resources into idle, underutilized, and optimized categories. Generated optimization recommendations along with estimated cost savings are returned to the dashboard and rendered using graphical visualizations. The architecture supports seamless migration to production-grade cloud environments. The backend server can be containerized using Docker and deployed on scalable cloud infrastructure, while machine learning models can be periodically retrained using updated usage data. This design ensures scalability, maintainability, and adaptability to dynamic cloud workloads.

V. RESULTS AND DISCUSSION

A. Quantitative Evaluation

The proposed cloud cost optimization system was evaluated using historical cloud usage data generated over a period of three years for 20 cloud resources. The evaluation was performed using an 80:20 train–test split, and standard regression performance metrics, including the R^2 score and Mean Absolute Error (MAE), were used to assess prediction accuracy. Table I summarizes the overall performance of the three machine learning models considered in this study. Among the evaluated models, the Random Forest Regressor achieved the best overall performance, recording an R^2 score of 0.9624 and an MAE of $0.50 per day, indicating highly accurate and reliable cloud cost prediction. The Decision Tree Regressor also demonstrated strong performance with an R^2 score of 0.9587, though slightly lower than that of the Random Forest model. In contrast, the Linear Regression model exhibited comparatively lower accuracy, achieving an R^2 score of 0.7921, primarily due to its limited ability to model non-linear relationships present in complex cloud usage patterns. These results clearly indicate that ensemble-based learning approaches, particularly Random Forest, are more effective for cloud cost prediction tasks involving multi-dimensional and non-

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

linear resource utilization data. The superior performance of the Random Forest model justifies its selection as the primary prediction engine for the proposed cloud cost optimization system.

TABLE I. MODEL PERFORMANCE METRICS ON TEST SET

Metric	Value
MSE	0.3824
R^2 Score	0.9624
MAE ($/day)	0.50
Training Split	80%
Backbone	Random Forest Regressor

B. Resource Classification Analysis

The system classifies cloud resources into optimized, underutilized, and idle categories based on predefined utilization thresholds. This classification enables rapid identification of inefficiencies within the cloud environment. Experimental analysis indicates that a considerable number of resources fall under the idle and underutilized categories, leading to unnecessary cloud expenditure. Idle resources incur costs despite minimal usage, while underutilized resources operate below their optimal capacity. The classification results are visualized through the dashboard (Fig. 1), allowing users to quickly identify inefficient resources and apply corrective actions such as termination or downsizing, thereby improving overall cost efficiency and reducing manual intervention

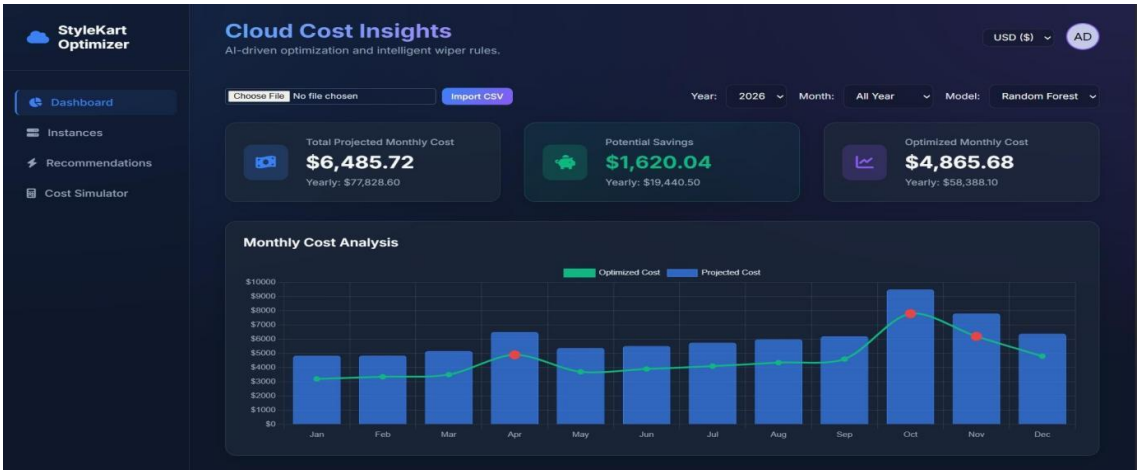

Fig. 1. Cloud Cost Dashboard Showing Projected Cost, Optimized Cost, and Potential Savings

C. Recommendation System Evaluation

The system generates actionable optimization recommendations based on cloud resource utilization patterns. These include terminating idle resources, downsizing underutilized instances, and applying scheduling strategies. The recommendations dashboard (Fig. 2) highlights the total potential savings, along with the count of idle and underutilized resources, and provides visual insights such as savings distribution and top cost-saving opportunities. This enables users to effectively prioritize optimization tasks and significantly improve overall cost efficiency.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

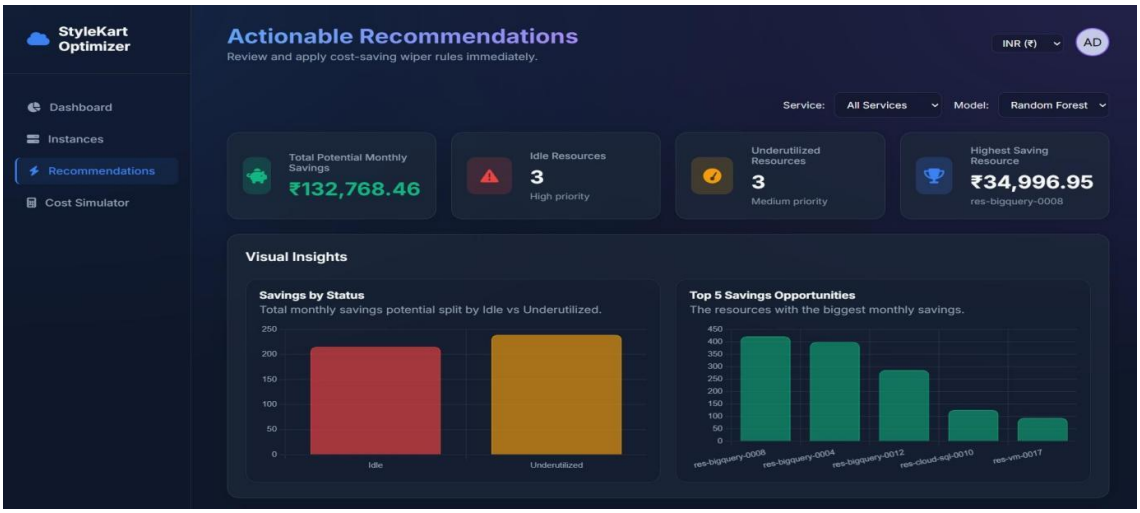

Fig. 2. Actionable Recommendations and Savings Insights Dashboard

D. Optimization Action Center

The Action Center provides detailed, resource-level optimization insights, including issue identification, recommended actions, and estimated cost savings. resources are categorized into high, medium, and low priority, enabling users to address critical optimization actions first. Recommendations such as resource termination, scheduling, and instance downsizing are clearly presented along with their expected savings, ensuring efficient and prioritized cloud resource management.

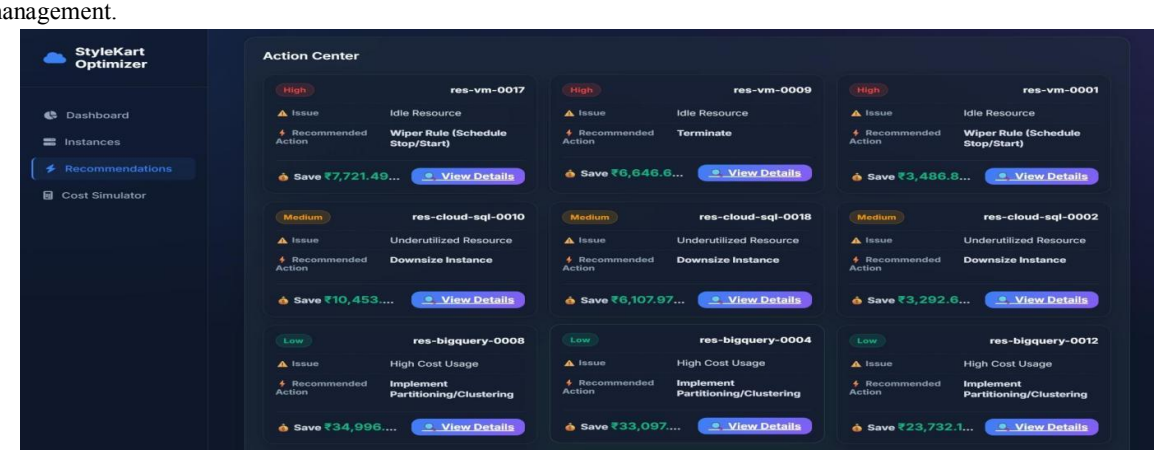

Fig. 3. Resource-Level Optimization Action Center with Priority-Based Recommendations

E. Resource Analysis

The system provides detailed insights into individual cloud resource usage and cost history through a dedicated resource details view. this view presents key information such as service type, current cost, recommended optimization actions, and historical usage patterns, along with graphical representations of predicted cost trends and CPU utilization. This enables users to clearly understand resource behavior and make informed optimization decisions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

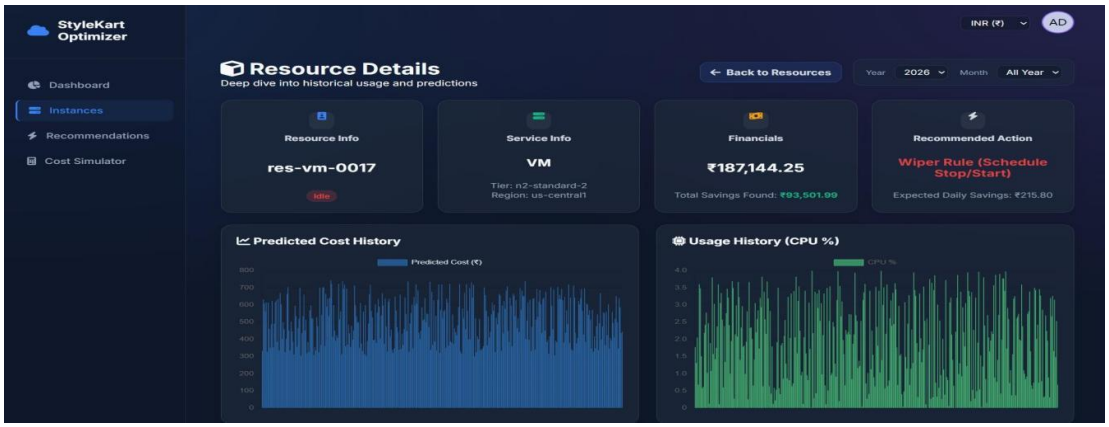

Fig. 4. Resource Details Showing Cost History and CPU Utilization Trends

VI. CONCLUSION

This paper presented an intelligent cloud cost optimization and forecasting system using machine learning techniques. The proposed system effectively analyzes cloud resource utilization and predicts future costs using a Random Forest regression model, achieving high prediction accuracy with an R^2 score of 0.9624. The system successfully classifies cloud resources into optimized, underutilized, and idle categories, enabling efficient identification of resource inefficiencies. Based on this classification, it generates actionable recommendations such as resource termination, instance downsizing, and scheduling, which significantly reduce unnecessary cloud expenditure. The integration of an interactive dashboard enhances usability by providing clear visualizations of cost trends, potential savings, and resource utilization patterns. This enables users to make informed, data-driven decisions and optimize cloud resource management effectively. Overall, the proposed system offers a scalable, automated, and data-driven solution for cloud cost optimization, reducing manual effort while improving overall cost efficiency. In future work, the system can be enhanced by incorporating real-time cloud usage data, adopting advanced deep learning models, and integrating directly with cloud platforms such as AWS, Microsoft Azure, and Google Cloud to enable automated optimization and deployment.

ACKNOWLEDGMENT

The authors would like to thank Ms.R.M.SUGANYA,M.E.,(CSE) , K.L.N. College of Engineering, for her continuous guidance and support throughout this project. The authors also acknowledge the use of Google Colab's GPU resources and the Roboflow platform for dataset curation and annotation management.

REFERENCES

[1] S. S. Manvi and G. K. Shyam, "Cloud Resource Management Techniques," Journal of Network and Computer Applications, vol. 45, pp. 1–13, 2014.
[2] A. Mustafa, M. A. Khan, and S. Iqbal, "Machine Learning-Based Cost Forecasting in Cloud Computing," International Journal of Computer Applications, vol. 120, no. 5, pp. 20–25, 2015.
[3] S. Islam, J. Keung, K. Lee, and A. Liu, "Empirical Prediction Models for Adaptive Resource Provisioning in the Cloud," Future Generation Computer Systems, vol. 28, no. 1, pp. 155–162, 2012.
[4] S. Ratti Halli, S. Hiremath, and P. Patil, "Detection of Idle Virtual Machines in Cloud Data Centers," International Journal of Cloud Computing, vol. 8, no. 2, pp. 123–135, 2019.
[5] Q. Zhang, M. Chen, and L. Li, "Cloud Cost Optimization Using Predictive Analytics," IEEE Transactions on Cloud Computing, vol. 6, no. 2, pp. 450–462, 2018.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

[6] A. Verma and S. Kaushal, "Resource Allocation in Cloud Computing Using Machine Learning," Procedia Computer Science, vol. 125, pp. 623–630, 2017.
[7] Amazon Web Services, "AWS Pricing and Cost Management," [Online]. Available: https://aws.amazon.com/pricing/.
[8] Microsoft Azure, "Azure Pricing Overview," [Online]. Available: https://azure.microsoft.com/pricing/.
[9] Google Cloud, "Google Cloud Pricing," [Online]. Available: https://cloud.google.com/pricing