IJWR

Analysis and Optimization of Customer Lifetime Value Prediction using Machine Learning and Deep Learning Models by RFM Techniques

Leila Taherkhani*[a], Amir Daneshvar[b], Hossein Amoozad Khalili[c], MohammadReza Sanaei[d]

[a] Department of Information Technology Management, Science and Research Branch, Islamic Azad University, Tehran, Iran; leila.taherkhani@srbiau.ac.ir

[b] Department of Industrial Management, Science and Research Branch, Islamic Azad University, Tehran, Iran; a_daneshvar@iauec.ac.ir

[c] Department of Industrial Engineering, Sari Branch, Islamic Azad University, Sari, Iran; amoozad92@yahoo.com

[d] Department of Information Technology Management, Qazvin Branch, Islamic Azad University, Qazvin, Iran, mohamadrezasanaei@gmail.com

ABSTRACT

In today's data-driven hospitality sector, customer interactions increasingly occur through digital platforms, generating extensive behavioral and transactional information. This study analyse the prediction of Customer Lifetime Value (CLV) using machine learning models—Linear Regression, Random Forest, and LightGBM—trained on features derived from hotel website interactions and booking records. After comprehensive data preprocessing, the models were evaluated using MAE, RMSE, and R^2 metrics. LightGBM achieved the highest predictive performance (R^2 = 0.504), followed by Random Forest (R^2 = 0.497), while Linear Regression underperformed (R^2 = 0.386), highlighting the advantages of non-linear models in modeling intricate customer patterns. Residual analyses confirmed LightGBM's stability and low bias across diverse customer profiles. Apart from prediction, the study applies Recency-Frequency-Monetary (RFM) analysis to segment customers into distinct value-based groups. These segments form the basis for tailored marketing strategies, allowing hotels to allocate resources more efficiently, enhance customer retention, and develop targeted campaigns aligned with customer potential. By integrating web-derived behavioral data with advanced modeling and segmentation, this research offers hotel managers practical tools for strategic planning in customer relationship management.

Keywords*— *Customer Lifetime Value (CLV), Machine Learning, Random Forest, LightGBM, RFM.

1. Introduction

Traditional metrics such as occupancy rates and Average Daily Rate (ADR) offer a snapshot of a hotel's financial performance but fall short in capturing the long-term value of customer relationships. Customer Lifetime Value (CLV) addresses this limitation by estimating the total revenue a guest is expected to generate throughout their relationship with the hotel [1, 2]. This metric enables a strategic shift from transactional thinking to long-term customer engagement.

By analyzing guest history, purchasing behavior, and loyalty program engagement, hotels can identify high-value customers and tailor marketing strategies accordingly [3, 4]. These insights drive personalized guest experiences, targeted promotions, and enhanced customer satisfaction, all of which contribute to greater guest loyalty, increased repeat visits, and long-term revenue growth [5, 6]. Moreover, CLV insights allow for dynamic pricing strategies that adjust room rates based on individual guest profiles and market trends, maximizing

http://dx.doi.org/10.22133/ijwr.2025.508322.1272

Citation L. Taherkhani, A. Daneshvar, H. Amoozad Khalili and M. Sanaeid, "Analysis and Optimization of Customer Lifetime Value Prediction using Machine Learning and Deep Learning Models by RFM Techniques ", *International Journal of Web Research*, vol.8, no.2,pp.79-92, 2025, doi: http://dx.doi.org/10.22133/ijwr.2025.508322.1272.

**Coressponding Author*

Article History: Received: 21 January 2025; Revised: 7 March 2025; Accepted: 25 March 2025.

Copyright © 2025 University of Science and Culture. Published by University of Science and Culture. This work is licensed under a Creative Commons Attribution-Noncommercial 4.0 International license(https://creativecommons.org/licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

profitability while maintaining guest satisfaction [7, 8].

However, current CLV prediction approaches face several challenges. Fragmented guest data across systems such as CRM, PMS, and online reservation platforms limit the construction of unified guest profiles. Additionally, guest behavior is influenced by dynamic factors like seasonality, travel purpose, competition, and economic fluctuations, making CLV estimation more complex. Existing models often fail to provide actionable outputs such as guest segmentation, churn risk, or marketing recommendations, limiting their practical value.

This study addresses these gaps by proposing an innovative machine learning-based framework for CLV prediction tailored specifically for the hotel industry. The innovation of this research lies in the comprehensive integration of structured and unstructured data—including booking history, spending patterns, loyalty program participation, and online reviews—to build accurate, dynamic, and actionable CLV models. Unlike prior studies that rely on static metrics or partial data, this research leverages advanced machine learning algorithms to process diverse data inputs and capture the evolving nature of guest behavior.

In addition to predicting individual guest spending and churn probability, the proposed framework delivers strategic insights for guest segmentation and personalized marketing initiatives. This enables hotels to allocate resources efficiently, increase guest retention, and drive revenue growth through targeted engagement strategies.

By applying machine learning techniques in the context of the hospitality industry, this study contributes to both academic research and practical hotel management. It offers a scalable, data-driven approach to understanding and optimizing customer value, helping hotels remain competitive in an increasingly data-centric market. Thus, this research fills a critical gap in the literature by demonstrating how machine learning can be effectively utilized for actionable CLV prediction in the hotel sector.

2. Literature Review

2.1. Related Concepts

Various approaches have been proposed in the literature for predicting CLV. [1] identified six key modeling techniques: The Recency, Frequency, and Monetary (RFM) model; probability models based on the Pareto/NBD model and Markov chains; customer acquisition; customer retention; and customer margin and expansion models. Similar to probabilistic models like the Pareto/NBD model, econometric lifespan models emphasize forecasting key components such as acquisition, retention, and cross-selling. In contrast, computer science models are rooted in theoretical frameworks (e.g., utility theory) and are relatively easy to interpret [9].

RFM analysis is a data-driven customer segmentation technique used to identify and categorize customers based on their purchasing behavior [10]. RFM represents Recency, Frequency, and Monetary. **Recency (R)** indicates how recently a customer has completed a purchase [11]. Customers who have recently purchased are often considered more engaged and valuable. Frequency (F): measures how many times a customer purchases in a specific period [12]. Customers who make repeat purchases are likely to be more loyal and valuable to the business. Monetary value (M): represents the amount of money a customer has spent on a business in a particular period [13]. Customers with higher monetary value are generally more profitable. By analyzing these three dimensions, customers can be divided into different groups and points assigned to each dimension [14]. RFM analysis allows businesses to identify their most valuable customers as well as those who may be at risk of churn [15]. This information can then be used to adjust marketing strategies, personalize communications, and optimize customer retention efforts [16].

Customer segmentation is an essential aspect of modern marketing strategies, allowing businesses to customize their approaches based on unique customer characteristics and behaviors. Clustering, a key data analysis technique, facilitates this process by categorizing customers into groups with similar traits.

Significance of Customer Segmentation:

- Targeted Marketing – Enables personalized marketing efforts.
- Resource Optimization – Ensures efficient allocation of resources.
- Improved Customer Retention – Helps businesses retain valuable customers.

The K-means clustering algorithm is widely used for segmenting customers into *k* distinct groups based on their purchasing behavior and preferences. It is particularly effective when the number of segments is predetermined. Research indicates that combining K-means with RFM modeling enhances segmentation accuracy, leading to better CLV prediction [17].

Integrating clustering techniques into CLV prediction models offers several benefits, including:

- Identifying High-Value Segments – Recognizing the most profitable customer groups.

- Dynamic Segmentation – Allowing real-time updates to customer segments.
- Behavioral Insights – Understanding customer preferences to predict future purchasing patterns.

For instance, detecting customers who frequently respond to discounts can help refine promotional strategies, ultimately increasing CLV [18].

2.2. Common Methods for CLV

CLV is a crucial metric for businesses, helping them understand the long-term value of their customers. Predicting CLV accurately allows for better resource allocation, targeted marketing strategies, and ultimately, business growth. This review explores the growing importance of machine learning techniques in CLTV forecasting.

Traditionally, businesses relied on simpler methods to estimate CLV. However, the recent surge in machine learning offers powerful tools for more precise predictions. Common machine learning algorithms used for CLV prediction include:

- Clustering models: These group customers with similar characteristics, enabling CLV prediction for each cluster.
- Multi-class classification: This approach categorizes customers based on their potential CLV, allowing for personalized marketing strategies.
- Regression models: These techniques estimate future customer value based on historical data and various predictors.
- Deep neural networks: Recent research shows promise in using deep learning frameworks for CLTV prediction, potentially outperforming traditional models [19].

The effectiveness of a CLV forecasting model is typically evaluated based on three key criteria: predictive accuracy, its influence on strategic business decisions, and its ability to optimize resource allocation and customer segmentation efforts [20][21]. In developing these models, firms commonly incorporate variables such as customer demographics, purchasing behavior, and historical sales transactions to enhance the precision of predictions [22].

Recent advancements in CLV modeling reflect the growing adoption of machine learning and deep learning techniques across various industries:

- A novel framework tailored for B2B SaaS companies illustrates the superiority of machine learning techniques over traditional statistical models, especially in handling high-dimensional and dynamic datasets [23].
- Multi-output deep neural networks have been effectively utilized in predicting CLV for complex, multi-tier e-commerce platforms, allowing simultaneous prediction of multiple customer-related outcomes [24].
- Several studies underscore the critical role of historical consumption behavior and anticipated purchase patterns in refining CLV forecasts, reinforcing the need for behaviorally rich datasets [25][26].
- An innovative system named *perCLTV* has been introduced to forecast personalized CLV in the context of online gaming, employing machine learning algorithms to tailor predictions at the individual user level [19].

These collectively demonstrate the increasing sophistication and applicability of CLV models in diverse domains, moving beyond generic estimation toward highly customized and actionable insights. Also, these studies highlight the efficiency of machine learning in predicting CLV.

[28] provide a comprehensive overview of the key technologies, challenges, and future directions in CLV prediction using machine learning and deep learning techniques. Their work highlights the potential for a shift in recommender systems towards long-term customer value goals.

Beyond prediction accuracy, recent research explores incorporating risk factors into CLV calculations. For example, [29] introduces a risk-adjusted return (RAR) measure in the telecommunications industry to account for customer risk in CLV calculations.

[30] emphasize the importance of identifying key variables that significantly impact CLV predictions. [31] propose a framework combining clustering and regression models for customer segmentation based on predicted CLV. This allows businesses to prioritize high-value customer groups for targeted marketing efforts.

Machine learning offers a powerful toolkit for businesses to gain deeper customer insights and make data-driven decisions. [32] demonstrates the effectiveness of various machine learning algorithms in CLV analysis. [33] highlight the overall value of CLV models in evaluating customer relationships and driving business growth. [33] described the consumers through a variant of the RFM model. They categorized customers into clusters and measured their profitability using the CLV. They used the self-organizing map algorithm for classification. Their results are applicable to

retailers. [34] express machine learning algorithms significantly outperforms traditional CLV estimation methods. [35] used a cohort analysis to investigate CLV for customer cohorts acquired before and during the COVID-19 pandemic. Their research estimates CLV in a continuous-time setting of customer transactions within the online grocery sector. They combined stochastic models with the Gamma-Gamma spending model to predict CLV at individual and aggregate levels. [36] introduced a stacked ensemble learning approach, which integrates multiple machine learning techniques for CLV prediction. This method was evaluated against several widely used predictive models, including deep neural networks, bagging support vector regression, light gradient boosting machine, random forest, and extreme gradient boosting.

According to the review of the research background and the existing gap, this research intends to use machine learning techniques, regression algorithms, as well as the combination of RFM and cohort techniques to predict CLV. The results of this research can help business owners including hotels with valuable insights to maintain their loyal customers by creating successful marketing campaigns.

3. Methodology

3.1. Dataset

Hotel customer dataset with 31 features describing a total of 83,590 items (customers). This dataset contains information from three full years of behavioral data of customers of hotels in Lisbon, Portugal, which was collected by [37]. In addition to personal and behavioral information, the dataset also contains demographic and geographic information [37]. This dataset can be used in particular to build customer segmentation models, including clustering and RFM models, as well as classification and regression problems. This research focused on a subset of features from the dataset. The selection prioritized those most relevant to addressing the specific research problems, rather than utilizing all available features. Table 1 describes the dataset features and their meanings.

3.2. Proposed Method

The steps of conducting this research are as follows. Figure 1 also shows these steps.

Stage 1:

Input data is prepared to calculate RFM variables.

Stage 2:

- New variables R, F, and M are added as new columns.
- CLV is calculated.
- Features that have a correlation with CLV are selected, and grouping is performed.
- K-Means clustering is performed to group customers.

Table 1. Dataset Description

Features	***Description***
ID	Unique identifier for each customer
Nationality	Country of origin of the customer
Age	Age of the customer
DaysSinceCreation	Number of days since the customer profile was created
NameHash	Hashed representation of the customer's name for anonymity
DocIDHash	| Hashed document ID for customer identification
AverageLeadTime	Average number of days between booking and check-in
LodgingRevenue	Revenue generated from lodging services
OtherRevenue	Revenue from additional services beyond lodging
BookingsCanceled	Number of bookings canceled by the customer
BookingsCheckedIn	Number of bookings where the customer successfully checked in
PersonsNights	Total number of nights booked per person
RoomNights	Total number of nights booked per room
DaysSinceLastStay	Number of days since the customer's last stay
DaysSinceFirstStay	Number of days since the customer's first stay
DistributionChannel	Channel through which the booking was made (e.g., corporate, travel agent)
MarketSegment	| Market segment classification of the customer (e.g., corporate, leisure)
SRHighFloor	Special request for a high-floor room
SRLowFloor	Special request for a low-floor room
SRAccessibleRoom	Special request for an accessible room
SRMediumFloor	Special request for a medium-floor room
SRBathtub	Special request for a bathtub in the room
SRShower	Special request for a shower in the room
SRCrib	Special request for a crib in the room
SRKingSizeBed	Special request for a king-size bed
SRTwinBed	Special request for a twin bed
SRNearElevator	Special request for a room near the elevator
SRAwayFromElevator	Special request for a room away from the elevator
SRNoAlcoholInMiniBar	Special request for no alcohol in the minibar
SRQuietRoom	Special request for a quiet room

Stage 3:

- The data is divided into two sets: a training set containing 80% of the data and a testing set containing 20% [38].
- appropriate algorithms for regression modeling are selected and trained.

Stage 4:

The performance of the models is evaluated based on metrics of MSE[1], MAE[2], and R-squared [39].

Stage 5:

The results of the model performance are used to predict CLV for new and existing customers

4. Results

4.1. RFM Indicators

To calculate the RFM indicators, from the feature of the first purchase 'DaysSinceLastStay' for Recency is used. A new column Recency is created and filled with the values from the existing column 'DaysSinceLastStay'. This indicates how recently the customer last stayed. In order to discretize, the data is divided into quartiles based on the distribution of Recency. The labels ['4', '3', '2', '1'] are assigned to the quartiles. Quartile 4: Represents customers with the highest Recency values (i.e., least recent customers). Quartile 1: Represents customers with the lowest Recency values (i.e., most recent customers). Since lower Recency is better, customers with the lowest values get a higher score (4).

'BookingsCanceled', 'BookingsNoShowed', and 'BookingsCheckedIn' features are uesed for Frequency. The Frequency column is calculated by summing three other columns:

- BookingsCanceled: Number of bookings canceled by the customer.
- BookingsNoShowed: Number of bookings where the customer didn't show up.

'BookingsCheckedIn': Number of bookings where the customer actually checked in.

This provides the total number of interactions the customer had, regardless of the outcomes. The data divides the ranked data into quartiles. Labels: ['1', '2', '3', '4'] are assigned to the quartiles. Quartile 4: Represents customers with the highest Frequency values (most frequent interactions). Quartile 1: Represents customers with the lowest Frequency values. Higher Frequency is better, so customers with more interactions get a higher score.

[1] Mean Squared Error
[2] Mean Absolute Error

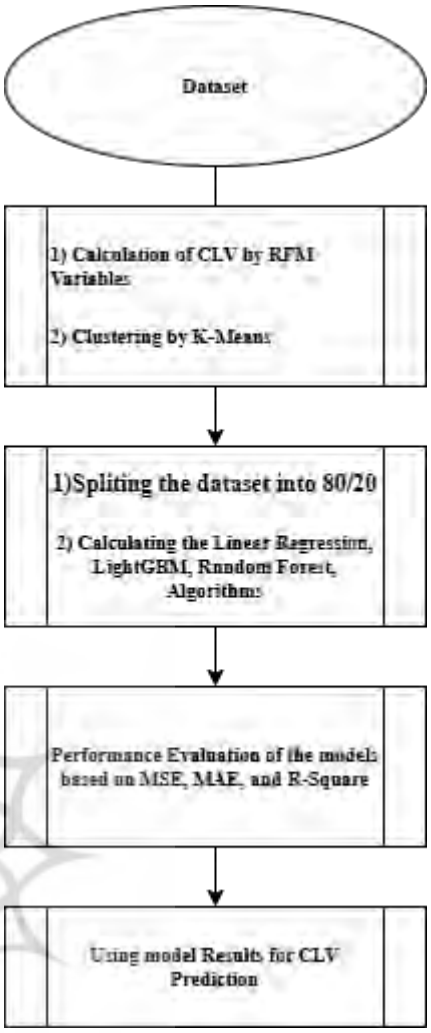


Figure. 1. Research implementation process

The Monetary column is calculated by summing 'LodgingRevenue' that is revenue generated from the customer's lodging bookings and 'OtherRevenue' that is revenue from other sources, such as dining, spa, or additional services. This represents the total monetary value contributed by the customer. The data Divides the Monetary values into quartiles. Labels: ['1', '2', '3', '4'] are assigned to the quartiles. Quartile 4: Represents customers with the highest Monetary values (largest contributions). Quartile 1: Represents customers with the lowest Monetary values. Higher Monetary value is better, so customers who spent more get a higher score.

4.2. Customer Grouping

Final Output will have three new columns:

- R_{Score}: Indicates the recency quartile score.
- F_{Score}: Indicates the frequency quartile score.
- M_{Score}: Indicates the monetary quartile score.

These scores can then be combined to calculate an RFM_{score} as Equation (1), often used for customer segmentateon:

$$RFM_{Score}=R_{Score} + F_{Score} + M_{Score} \quad (1)$$

The RFM scoring system helps identify customer segments such as:

- Champions: High RFM scores.
- At Risk: Low Recency, but moderate Frequency and Monetary.
- Loyal Customers: High Frequency and Monetary but moderate Recency.

Weighted RFM Score Calculation:

Weights:

- w_R = 0.5
- w_F = 0.3
- w_M = 0.2

The weights (w_R, w_F, w_M) represent the proportional significance of Recency, Frequency, and Monetary scores, respectively:

- Recency (R_Score) is assigned the highest importance (50% weight) since retaining recent customers is often critical.
- Frequency (F_Score) has moderate importance (30% weight), indicating the significance of repeat interactions.
- Monetary (M_Score) has the lowest importance (20% weight) in this case.
- Conversion to Integers: R_Score, F_Score, and M_Score are stored as strings (since they were created with labels). They are converted to integers to perform arithmetic.
- Weighted Sum: The final score is calculated as Equation (2):

$$RFM_{Score}=(R_{Score}*w_R) + (F_{Score}*W_F) + (M_{Score}*W_M) \quad (2)$$

The Equation (2) gives higher importance to Recency, followed by Frequency, and then Monetary values. Then a new column RFM_Score is added to the DataFrame, containing the weighted RFM score for each customer. The score represents a composite measure of customer value, where:

- Higher RFM_Score: Indicates a more valuable customer.
- Lower RFM_Score: Indicates a less engaged or less valuable customer.

The weighted RFM score allows for fine-tuned customer segmentation by reflecting the relative importance of each component. This is especially useful in tailoring marketing strategies or prioritizing customer groups based on their overall value.

4.3. Customer Segmentation

- Customer Segmentation takes an individual RFM_score as input and assigns a segment based on predefined thresholds:
- Best Customers: Customers with RFM scores of 4 or higher (most engaged and valuable).
- Loyal Customers: Customers with RFM scores between 3 (inclusive) and 4 (exclusive) who engage consistently.
- Potential Customers: Customers with RFM scores between 2 (inclusive) and 3 (exclusive) who show promise but are not yet fully engaged.
- At Risk: Customers with RFM scores below 2, indicating disengagement or low value.

The Segment column categorizes customers into one of the four groups. This segmentation provides a clearer understanding of customer behavior and value, allowing for targeted strategies for each group. Table 2 and Figure 2 show the count of customers in each segment based on their RFM_Score.

Table 2. Customer Segmentation

Segment	*Count*	*Percentage (%)*
At Risk	19,970	34.3%
Loyal Customers	19,504	33.5%
Potential Customers	19,394	33.3%
Best Customers	4,802	8.3%

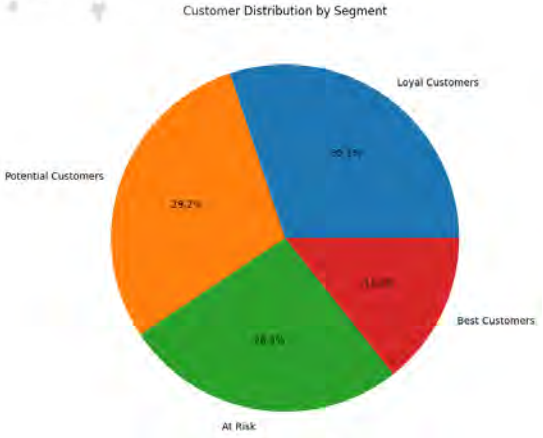


Figure. 2. Customer Segmentation

4.4. Calculation of CLV

To calculate CLV, first based on formula (3) a variable as named 'CustomerLifetimeYears' calculated. Then it was used for calculation of CLV as Equation (4).

$$CustomerLifetimeYears = DaysSinceFirstStay / 365 \quad (3)$$

$$CLV = (Monetary / Frequency) * Frequency * CustomerLifetimeYears \quad (4)$$

- Monetary: The total amount of money the customer has spent.
- Frequency: The total count of purchases made by the customer.
- 'CustomerLifetimeYears': The duration (in years) of the customer's relationship with the business.

Since the Frequency cancels out in the numerator and denominator, the formula simplifies to Equation (5).

$$CLV = Monetary * CustomerLifetimeYears \quad (5)$$

Thus, this formula essentially calculates CLV based on the monetary value and the customer's lifetime duration.

After calculating the two variables (CLV, and 'CustomerLifetimeYears'), they will be added to the corresponding dataset as two new features.

4.5. Pre-Processing

Data Preparation:

Irrelevant or extreme rows are removed based on CLV and 'CustomerLifetimeYears' columns. Sorting ensures that data is cleaned and ordered appropriately.

Feature Normalization:

Scaling the CLV, RFM_Score, and 'CustomerLifetimeYears' ensures that all three features are in the same range (0 to 1). This is crucial when features are used in models sensitive to scale (e.g., distance-based algorithms like KNN or clustering).

Optimization:

Removing rows and normalizing data reduces the effect of outliers and ensures better performance during analysis or model training.

4.6. Correlation Analysis

The correlation analysis identifies relationships between variables such as CLV, RFM_Score, 'CustomerLifetimeYears', Frequency, and Monetary, and helps assess which variables are strongly or weakly correlated, which can guide feature selection or interpretation of relationships. Figure 3 provides an intuitive way to observe correlations, with colours enhancing interpretability (e.g., red for strong positive correlations, blue for strong negative correlations). This heatmap shows how strongly different numerical features relate to CLV. 'LodgingRevenue' and 'OtherRevenue' have very high positive correlations with CLV, especially 'LodgingRevenue' (r = 0.97)—which suggests they might be closely tied to the target, possibly even revealing data leakage. To keep our analysis honest, we decided to remove these features in later tests. The heatmap also indicates that most other features don't have strong relationships with CLV, hinting that to capture the complex patterns, we might need to use more advanced, non-linear modeling techniques.

4.7. Clustering based on CLV using K-Means

K-Means initializes 4 centroids randomly (or based on initialization settings).

Each data point is assigned to the nearest centroid based on Euclidean distance. The centroids are updated iteratively until convergence (when the centroids no longer change significantly) [40].

Groups customers into 4 clusters based on their CLV. Each cluster represents a group of customers with similar CLV values. Figure 4 represent customer segmentation based on 'RFM_Score'. Customer segmentation is useful for analyzing and categorizing customers. High CLV customers (cluster with high centroid values) can be targeted for loyalty programs. Low CLV customers may require marketing strategies to improve engagement. The Cluster column serves as a categorical feature for further analysis or visualization. Figure 5 shows the clusters base on 'CustomerLifttimeYears'.

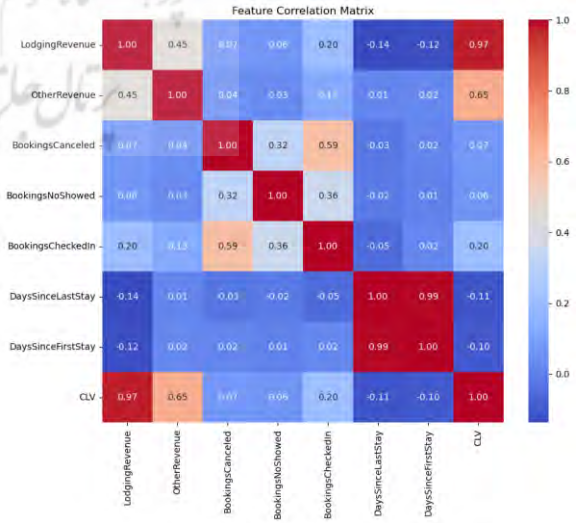


Figure. 3. Correlation Matrix

4.8. Modeling

Linear Regression

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable and one or more independent variables. The primary goal is to find a linear equation that best predicts the dependent variable based on the given independent variables.

In Multiple Linear Regression type an extends simple linear regression to include multiple independent variables. As an Equation (6):

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \quad (6)$$

Here, x_1, x_2, …, x_n are the independent variables, and each has its corresponding coefficient. Figure 6 reveals that the errors are systematically concentrated around specific values, exhibiting a discernible pattern within the scatter plot. This pattern indicates the presence of nonlinear relationships that the linear model fails to adequately represent. Consequently, this suggests that the model demonstrates limited generalization capability when applied to real-world data.

Random Forest

Random Forest is a powerful ensemble learning method primarily used for classification and regression tasks. It operates by constructing multiple decision trees during training and outputting the mode of the classes (for classification) or mean prediction (for regression) of the individual trees [41]. This approach enhances predictive accuracy and controls overfitting. The algorithm is also effective in predicting customer behaviour, such as hotel booking cancellations, allowing businesses to refine their strategies based on predictive insights [42]. Figure 7 reveals the residuals in this plot are more evenly dispersed around zero, exhibiting no apparent pattern, which suggests an improved fit and reduced bias in the predictions. Nonetheless, the presence of some minor deviations indicates that there remains potential for further refinement of the model.

LightGBM (Light Gradient Boosting Machine)

LightGBM is a gradient boosting system that employs tree-based learning methods. It is optimized for distributed and efficient training, making it especially suitable for handling large datasets and intricate prediction challenges. Advantages of LightGBM is high efficiency, scalability, flexibility, and robustness. LightGBM has been successfully applied in various domains, demonstrating its versatility and effectiveness [43]. Figure 8 reveals the residuals are symmetrically distributed around the zero line, exhibiting no discernible pattern, which implies that the model demonstrates robust generalization capabilities. The uniformity and absence of systematic bias further substantiate LightGBM's superior performance in this CLV prediction task.

4.9. Distribution of Residuals

Residual analysis was conducted to evaluate the quality and generalization behavior of each model. The residuals represent the difference between predicted and actual CLV values. Figure 9 depicts

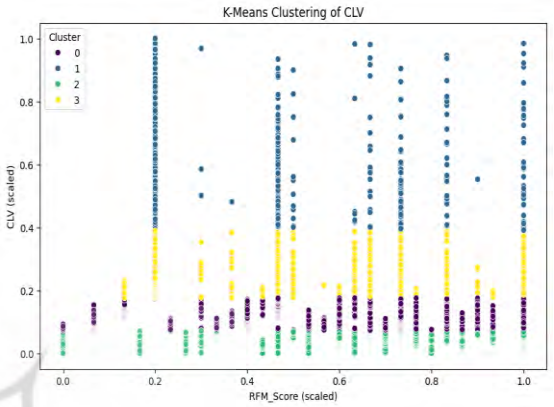


Figure. 4. K-Means Clustering based on CLV

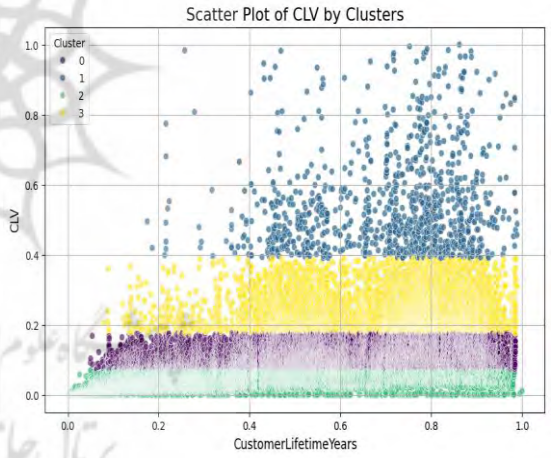


Figure. 5. Customer Segmentation based on 'CustomerLifetimeYears'

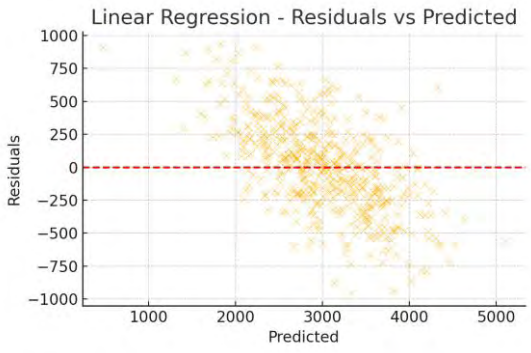


Figure. 6. Linear Regression Algorithm Results

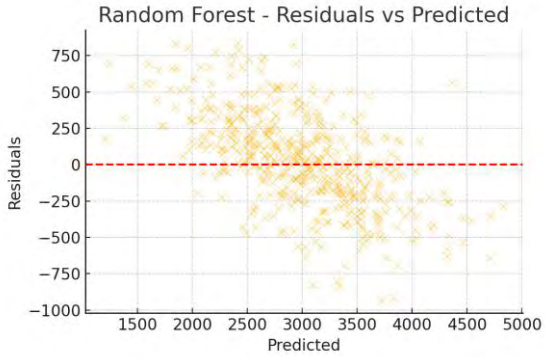


Figure. 7. The Random Forest Algorithm Result

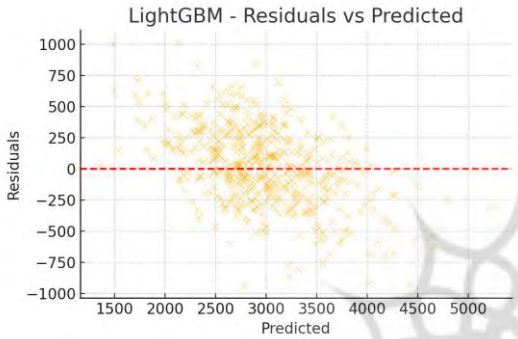


Figure. 8. The LightGBM Algorithm Result

the residual distribution derived from the linear regression model. The residuals exhibit substantial variability and markedly deviate from a normal distribution, indicating that the linear model was unable to adequately capture the nonlinear structures inherent in the data. Additionally, the extensive range of errors suggests the presence of underfitting within the model. The residuals from the Random Forest model in Figure 10, exhibit greater symmetry and are more closely centered around zero in comparison to those from the linear model. This suggests that the model achieved a more accurate prediction of CLV. However, some residual dispersion persists, potentially attributable to outliers or underlying nonlinear interactions not fully captured by the model. The LightGBM model in Figure 11, exhibits the most concentrated and symmetrically distributed residuals among all evaluated models. The predominance of errors gravitating close to zero indicates a high level of predictive accuracy. This histogram further corroborates LightGBM's superior performance.

However, neither model shows a perfectly normal distribution of residuals, especially Random Forest which has slight skewness, highlighting some model bias. LightGBM presents the most balanced residual spread, reinforcing its superior R^2 performance. These residual patterns confirm that

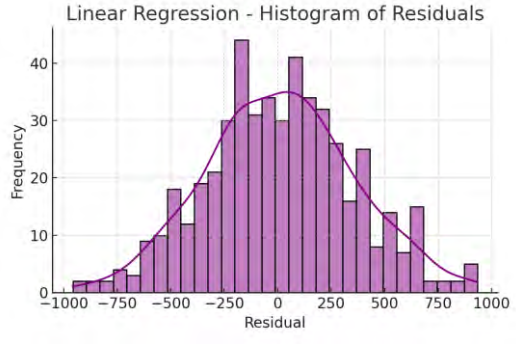


Figure. 9. Residual Distribution of LR

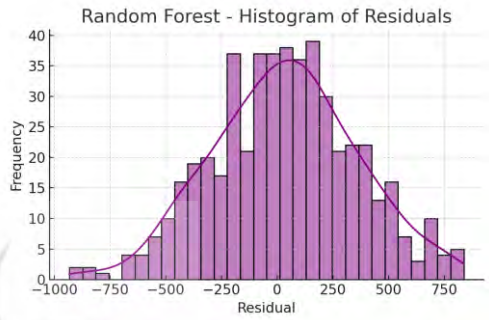


Figure. 10. Residual Distribution of RF

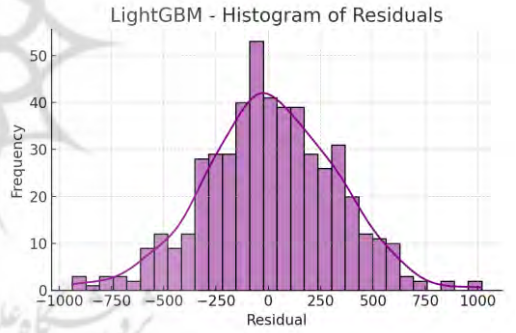


Figure. 11. Residual Distribution of LightGBM

non-linear models are more suitable for CLV prediction in this context.

4.10. Performance Evaluation

MSE, MAE and R-Squared criteria were used to evaluate the performance of the models used in the research. MSE is a common metric used to measure the squared difference between the values predicted by a model and the actual values. A lower MSE indicates a better fit. MAE is another common metric used to measure the absolute difference between the values predicted by a model and the actual values. A lower MAE indicates a better fit. R-squared quantifies the proportion of variation in the dependent variable that is accounted for by the independent variables. In this context, it represents the proportion of variance in the target variable that can be explained by the characteristics or predictors.

An R-squared value of 1 indicates a perfect fit. Table 3 presents the performance evaluation outcomes of the models and provides a comparative analysis.

Linear Regression:

- High MAE (179.11) and low R^2 (0.386) indicate that this model failed to model the complex relationships between features and CLV well.
- Also, the large RMSE (355.19) indicates that there are large errors in some samples.
- Conclusion: This model is underfitting and is not suitable for this problem.

Random Forest:

- It reduces errors compared to linear regression (MAE = 150.45) and provides more explanation of the CLV variance (R^2 = 0.497).
- However, the RMSE is still high, indicating that the model has high errors in samples.
- Conclusion: The model is acceptable and better than the linear model, but it can still be improved.

LightGBM:

- It has the best R^2 (0.504) and the lowest RMSE (319.38).
- Although its MAE is slightly higher than RF, its predictions are generally more stable and accurate.
- Residual analysis also showed that LightGBM has less dispersion and bias.
- Conclusion: LightGBM is the best model among the three models examined for CLV prediction.

4.11. Summary and Disucussion

The results of this research underscore the varying degrees of efficacy exhibited by different machine learning techniques in forecasting CLV. Among the models evaluated, LightGBM demonstrated the most balanced performance, achieving the highest R^2 score (0.504) along with the lowest RMSE, thereby evidencing its superior capacity to generalize to unseen data. Although Random Forest produced a marginally lower MAE, residual analyses indicated minor biases, rendering it somewhat less robust than LightGBM. Conversely, Linear Regression substantially underperformed, emphasizing the critical importance of employing non-linear models for accurate CLV prediction.

Residual analysis corroborated these findings, revealing that LightGBM's errors were most symmetrically distributed and narrowly dispersed. In contrast, the other models, particularly Linear Regression, showed greater variance and signs of underfitting. These observations suggest that CLV prediction, especially when utilizing real-world behavioral data, substantially benefits from advanced tree-based models capable of capturing complex, non-linear relationships.

Table 3. The Performance Evaluation

Model	*MAE*	*RMSE*	*R^2*
Linear Regression	*179.11*	*355.19*	*0.386*
Random Forest	*150.45*	*321.59*	*0.497*
LightGBM	*154.05*	*319.38*	*0.504*

In the customer segmentation component, RFM-based clustering produced meaningful groups. Assigning greater weight to recency within the RFM scores—especially—facilitated the identification of high-value segments, enabling effective targeted marketing strategies.

Overall, this study emphasizes that non-linear, tree-based models outperform both linear approaches and deep learning techniques such as LSTM in this application.

5. Scientific Contribution

The scientific contributions of this research encompass several critical dimensions:

- Methodological Integration: This study introduces a hybrid framework that combines Recency-Frequency-Monetary (RFM) analysis, customer segmentation, and advanced machine learning algorithms to improve the accuracy of CLV prediction within the hospitality industry. This integrated approach facilitates both behavioral insights and precise value estimation, providing a novel perspective on customer analytics.
- Model Evaluation and Benchmarking: By evaluating the performance of Linear Regression, Random Forest, and LightGBM models, the research identifies LightGBM as the most effective for CLV prediction, based on metrics such as R^2 and RMSE. The assessment methodology and residual analyses demonstrate that non-linear, tree-based models outperform linear models in capturing customer value dynamics, offering pragmatic guidance for selecting robust predictive models in hotel customer analytics.

- Empirical Validation and Interpretability: Utilizing real-world hotel customer data, the study not only forecasts CLV but also assesses residuals to examine model bias and generalizability. These findings furnish hotel managers with tangible tools for customer segmentation, marketing strategies, and resource optimization—effectively bridging predictive analytics with strategic decision-making.
- Practical Significance: The results provide actionable recommendations for hospitality managers to identify and target high-value customer segments, enhance retention efforts, and optimize marketing expenditures. Additionally, by addressing methodological challenges such as data leakage and residual bias, the research enhances analytical rigor and establishes a framework applicable to future CLV studies employing web-derived behavioral data within digital hospitality platforms.

6. Conclusion and Discussion

This study investigated the predictive performance of several machine learning models—Linear Regression, Random Forest, LightGBM for estimating CLV in the hotel industry. The models were evaluated using RMSE, MAE, and R-squared metrics. Among these, the LightGBM model demonstrated superior overall performance, evidenced by its highest R^2 and lowest RMSE, thereby indicating its robust capacity to model nonlinear and complex relationships. The Random Forest model performed comparably well, with a marginally higher MAE but somewhat reduced generalization capability. In contrast, Linear Regression markedly underperformed, indicating its inadequacy in capturing the nonlinear patterns inherent in CLV modeling.

This study distinguishes itself by conducting a comprehensive comparative analysis between classical and tree-based machine learning algorithms using a real-world hotel dataset encompassing behavioral, geographic, and demographic variables. Unlike prior research predominantly centered on e-commerce or retail sectors, this work extends CLV modeling to the hospitality industry. Notably, residual analysis was employed to examine model bias and error distribution, thereby providing a level of depth often absent in previous CLV investigations.

The findings furnish hotel management with data-driven tools to more effectively identify and target high-value customers. Algorithms such as LightGBM and Random Forest facilitate enhanced segmentation, personalized marketing strategies, and optimized promotional efforts. These predictive insights have direct applicability to loyalty programs, pricing strategies, and customer retention initiatives, thereby bridging the gap between theoretical development and practical strategic implementation.

Although the results are promising, several limitations should be acknowledged:

Data Scope:

The dataset was derived solely from customers in Lisbon, Portugal, potentially limiting the generalizability of the findings across different regions or industries. Additionally, external economic variables—such as inflation rates, promotional activities, and trends in foreign tourism—were not incorporated into the analysis.

Methodological Considerations:

The evaluation was limited to three models—Linear Regression, Random Forest, and Light Gradient Boosting Machine. The application of more advanced techniques, including AutoML frameworks or Transformer-based architectures, could yield further improvements. Furthermore, the presence of asymmetric residuals indicates the potential benefit of applying data transformations, such as logarithmic or Box-Cox transformations.

Temporal Dynamics:

CLV is influenced by evolving market trends. Static models trained on historical data risk becoming obsolete over time. Implementing dynamic updating mechanisms is essential to maintain model accuracy, particularly in a post-pandemic context.

Suggestions for future research

Recommendations for Future Research:

- Utilize comprehensive datasets spanning multiple industries.
- Investigate the application of more sophisticated architectures, such as AutoML, XGBoost, and Transformer models.
- Incorporate external economic and marketing indicators into CLV modeling.
- Adopt dynamic model retraining strategies to capture and reflect evolving customer behaviors.

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions

LT: Study design, interpretation of the results, analysis, drafting the manuscript, and revision of the manuscript;

AD, HAKH, and MRS: Study design, interpretation of the results, revision of the manuscript.

Conflict of interest

The authors declare that no conflicts of interest exist.

Acknowledgements

The authors would like to express their sincere gratitude to Engineer Mohammad Mehdi Ahmadi Babadi for their valuable support in the implementation and the extraction of results. Their technical expertise and dedication played an important role in the successful completion of this research.

Data Availability

DOI: 10.17632/j83f5fsh6c.1

References

[1] S. Gupta, D. Hanssens, B. Hardie, W. Kahn, V. Kumar, N. Lin, and S. Sriram, "Modeling customer lifetime value," *Journal of Service Research*, vol. 9, no. 2, pp. 139–155, Nov. 2006, https://doi.org/10.1177/1094670506293810.

[2] D. Jain and S. S. Singh, "Customer lifetime value research in marketing: A review and future directions," *J. Interact. Mark.*, vol. 16, no. 2, pp. 34–46, 2002, https://doi.org/10.1002/dir.10032.

[3] P. D. Berger and N. I. Nasr, "Customer lifetime value: Marketing models and applications" *J. Interact. Mark.*, vol. 12, no. 1, pp. 17–30, 1998, https://doi.org/10.1002/(SICI)1520-6653(199824)12:1<17::AID-DIR3>3.0.CO;2-K.

[4] N. Glady, B. Baesens, and C. Croux, "Modeling churn using customer lifetime value," *Eur. J. Oper. Res.*, vol. 197, no. 1, pp. 402–411, 2009, https://doi.org/10.1016/j.ejor.2008.06.027.

[5] D. R. Mani, J. Drew, A. Betz, and P. Datta, "Statistics and data mining techniques for lifetime value modeling," in *Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 1999, pp. 94–103, https://dl.acm.org/doi/pdf/10.1145/312129.312205.

[6] Z. Pollak, "Predicting Customer Lifetime Values—ecommerce use case," *arXiv preprint arXiv:2102.05771*, 2021, https://doi.org/10.48550/arXiv.2102.05771.

[7] B. P. Chamberlain, A. Cardoso, C. B. Liu, R. Pagliari, and M. P. Deisenroth, "Customer lifetime value prediction using embeddings," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2017, pp. 1753–1762, https://doi.org/10.1145/3097983.3098123.

[8] W. D. Dahana, Y. Miwa, and M. Morisada, "Linking lifestyle to customer lifetime value: An exploratory study in an online fashion retail market," *J. Bus. Res.*, vol. 99, pp. 319–331, 2019, https://doi.org/10.1016/j.jbusres.2019.02.049.

[9] P. Čermák, "Customer profitability analysis and customer lifetime value models: Portfolio analysis," *Procedia Econ. Finance*, vol. 25, pp. 14–25, 2015, https://doi.org/10.1016/S2212-5671(15)00708-X.

[10] A. Szilagyi, L. I. Cioca, L. Bacali, E. S. Lakatos, and A. L. Birgovan, "Consumers in the circular economy: A path analysis of the underlying factors of purchasing behavior," *Int. J. Environ. Res. Public Health*, vol. 19, no. 18, p. 11333, 2022, https://doi.org/10.3390/ijerph191811333.

[11] A. Petrov and C. Macdonald, "RSS: Effective and Efficient Training for Sequential Recommendation using Recency Sampling," *ACM Transactions on Recommender Systems*, vol. 3, no. 1, 1-32, 2022, https://doi.org/10.1145/3604436.

[12] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019, https://doi.org/10.1109/TASLP.2019.2915167.

[13] H. Rey, "Dilemma not trilemma: the global financial cycle and monetary policy independence," *Natl. Bur. Econ. Res.*, Working Paper No. w21162, 2015, https://www.nber.org/papers/w21162.

[14] A. H. L. Chen and S. Gunawan, "Enhancing Retail Transactions: A Data-Driven Recommendation Using Modified RFM Analysis and Association Rules Mining," *Appl. Sci.*, vol. 13, no. 18, p. 10057, 2023, https://doi.org/10.3390/app131810057.

[15] L. Saha, H. K. Tripathy, T. Gaber, H. El-Gohary, and E. S. M. El-kenawy, "Deep churn prediction method for telecommunication industry," *Sustainability*, vol. 15, no. 5, p. 4543, 2023, https://doi.org/10.3390/su15054543.

[16] A. M. A. Serwah, K. W. Khaw, C. S. P. Yeng, and A. Alnoor, "Customer analytics for online retailers using weighted k-means and RFM analysis," *Data Analytics and Applied Mathematics (DAAM)*, pp. 1–6, 2023, https://doi.org/10.15282/daam.v4i1.9171

[17] C. C. Aggarwal, *Data mining: the textbook*, Cham: Springer, 2015.

[18] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Syst.*, vol. 47, no. 4, pp. 547–553, 2009, https://doi.org/10.1016/j.dss.2009.05.016.

[19] G. Y. Benk, B. Badur, and S. Mardikyan, "A new 360 framework to predict customer lifetime value for multi-category e-commerce companies using a multi-output deep neural network and explainable artificial intelligence," *Information*, vol. 13, no. 8, p. 373, 2022, https://doi.org/10.3390/info13080373

[20] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2010,

[21] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, Cambridge, MA: MIT Press, 2016.

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[23] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012, https://doi.org/10.1145/2347736.2347755.

[24] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 2, 1995, pp. 1137–1143, https://dl.acm.org/doi/10.5555/1643031.1643047.

[25] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830,

2011, https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

[26] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, New York: Springer, 2013.

[27] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Univ. Waikato, Hamilton, New Zealand, 1999.

[28] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 161–168, https://doi.org/10.1145/1143844.1143865.

[29] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, https://doi.org/10.1023/A:1010933404324.

[30] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[31] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006, https://doi.org/10.1162/neco.2006.18.7.1527

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, https://doi.org/10.1162/neco.1997.9.8.1735.

[33] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2014, pp. 1746–1751, https://doi.org/10.3115/v1/D14-1181.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014, https://doi.org/10.48550/arXiv.1412.6980.

[35] J. Brownlee, *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python*, Machine Learning Mastery, 2018.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014, https://doi.org/10.48550/arXiv.1409.1556.

[37] N. Antonio, A. de Almeida, and L. Nunes, "A hotel's customers personal, behavioral, demographic, and geographic dataset from Lisbon, Portugal (2015–2018)," *Data in Brief*, vol. 33, p. 106583, 2020. https://doi.org/10.1016/j.dib.2020.106583

[38] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with statistical models," *J. Artif. Intell. Res.*, vol. 4, pp. 129–145, 1996, https://doi.org/10.1613/jair.295.

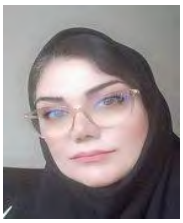
[39] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, https://doi.org/10.1007/BF00994018.

[40] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. 10th Eur. Conf. Mach. Learn.*, 1998, pp. 137–142, https://doi.org/10.1007/BFb0026683.

[41] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989, https://doi.org/10.1109/5.18626.

[42] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York: Springer, 2006.

[43] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne: OTexts, 2021.



Leila Taherkhani is a faculty member at Islamic Azad University, Karaj Branch, specializing in Information Technology Management. She earned her Ph.D. in IT Management-Business Intelligence from the Islamic Azad University, Science and Research Branch, Tehran (2019–2024), following an M.Sc. in IT Management-Information Resource Management from the same university's Electronic Campus (2016–2018), and a B.Sc. in Applied Mathematics in Computer Science from the Central Tehran Branch (1991–1995. She's research interests include artificial intelligence, data analytics, and business intelligence. She has authored five books, notably Metaverse: Concepts and Applications (2023), and Knowledge Management in the Age of Artificial Intelligence (2023). Her scholarly contributions encompass four journal articles, four conference papers, and two international publications, such as Intelligent Decision Support System Using Nested Ensemble Approach for Customer Churn in the Hotel Industry (2023). She has also presented at prominent conferences, including the 14th National Conference of Management and Humanities Researches in Iran (2023). In addition to her academic endeavors, she has served as an instructor at the University of Science and Culture (2024) and has been a lecturer at Islamic Azad University, Karaj Branch, since 2022.

Amir Daneshvar holds a Ph.D. in Systems Management from the Islamic Azad University, Science and Research Branch, Tehran (2014). He earned his M.Sc. in IT Management – Advanced Information Systems from the same university in 2007, and a B.Sc. in Industrial Engineering – Systems Analysis from Iran University of Science and Technology in 2004. He completed his high school education in Mathematics and Physics at Alborz High School in 1997. He ranked 1st nationwide in both the Ph.D. entrance exam for Industrial Management (2007) and the Master's entrance exam in IT Management (2004) at Islamic Azad University. He specializes in systems analysis, IT strategy, and project management, with extensive experience in designing and leading complex organizational initiatives. At present he is an assistant professor of industrial management department in Science and Research Branch of Islamic Azad University, Tehran, Iran.

Hossein Amoozad Khalili is an Associate Professor of Industrial Engineering at Islamic Azad University, Central Tehran Branch. He has authored and co-authored over 100 papers in national and international journals and conferences in the fields of industrial engineering and management. He serves as a member of the National Committee for Industrial Engineering at Islamic Azad University and is an interviewer for the university's Ph.D. entrance examinations. He is also a reviewer for several reputable domestic and international journals in industrial engineering and management. He has chaired the First International Conference on Industrial Engineering, Management, and Accounting and has been a scientific committee member in numerous national and international conferences. He has supervised more than 100 master's theses and 20 doctoral dissertations. In addition to his academic activities, he has authored several textbooks, including *Production and Operations Management, Plant Layout Design, Production and Inventory Planning and Control, Technical and Economic Project Evaluation, and Work and Time Study*.

Mohammad Reza Sanaei is an Assistant Professor of Information Technology Management at the Qazvin branch of Islamic Azad University, a position he has held since 2016. He earned his Ph.D. in Information Technology Management from the Institute for Cultural and Social Studies, affiliated with the Ministry of Science, Research, and Technology, in 2017. He has authored over 16 peer-reviewed articles in national and international journals and conferences, focusing on areas such as smart cities, digital marketing, artificial intelligence applications, and IT governance. Notable works include studies on gamification in data science education, blockchain applications in banking, and intelligent models for stock price prediction. He serves as the editor-in-chief of the Journal of Strategic Data Management Studies and is actively involved in academic publishing and peer review. His research interests encompass knowledge management, digital transformation, and the development of IT-based platform services.