Research on Hotel Reservation Customer Churn Based on Deep Neural Networks

Haoran Sun[a]
Mathematics and Applied Mathematics, Qingdao University of Science and Technology,
99 Songling Road, Laoshan District, Qingdao, China

Keywords: Machine Learning, Ensemble Learning, Hotel Reservation Dataset Forecast.

Abstract: In recent years, the Internet's rapid development has led to the increasing popularity of various online booking methods. Online hotel booking has emerged as a highly convenient option for individuals to plan their accommodations in advance. However, it is not uncommon to encounter cancellations following reservation confirmations. Hence, predicting the probability of hotel booking cancellations offers significant convenience for both customers and hotel operators, in line with the forecast. To enhance prediction accuracy, this study leverages a range of machine learning techniques and deep learning models, including Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), Gradient Boosting Decision Trees (GBDT), and Deep Neural Networks (DNN). Prior to the experimentation phase, there was an expectation that the DNN model would deliver superior outcomes. The final results validated the exceptional efficiency and effectiveness of the DNN model compared to all other models, achieving an AUC of 0.88 and an accuracy rate of 0.82, positioning it as the leading model among those assessed.

1 INTRODUCTION

For hotel operators, the occupancy rate is the most important indicator and a crucial means of generating profits. With the increasing popularity of advance hotel bookings, there has also been a rise in "no-show" or cancellation cases, which leads to significant wastage of accommodation resources. Timely prediction of customer attrition and understanding the reasons behind it can provide valuable insights for hotels and platforms to devise relevant strategies (Bai, 2021). By segmenting customers based on their consumption habits and implementing targeted retention measures, hotel operators can reduce customer attrition and capture a larger market share (Chen,2024). Moreover, whether it is online or offline bookings, a substantial amount of visitor and reservation data is generated. Mining critical customer information from these datasets can serve as a reference and support for formulating customer management plans by business operators (Fu,2020; Gordini, 2023; Hwang,2023).

In recent times, more and more researchers have applied machine learning models to customer attrition prediction, including the prediction of hotel customer attrition. Many researchers have combined customer big data collected by enterprises with machine learning models to establish customer attrition prediction models. Various models such as logistic regression, decision tree, support vector machine (SVM), etc., have been used to analyze different datasets and derive numerous findings and conclusions.

The dataset analyzed in this study is derived from two hotels in Portugal and includes features such as arrival dates, parking requirements, and room types. In comparison to other similar research reports, this study employs efficient deep learning methods, specifically neural network models, and compares them with other advanced machine learning models, including ensemble learning techniques like random forest and XGBoost, as well as traditional machine learning models such as logistic regression.

The structure of this paper is outlined as follows: Section 2 introduces previous relevant works, showcasing the analysis and research conducted by different scholars on various datasets. Section 3 provides a detailed description of the research

[a] https://orcid.org/0009-0000-7168-3919

Sun, H.
Research on Hotel Reservation Customer Churn Based on Deep Neural Networks.
DOI: 10.5220/0012918000004508

In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 175-182
ISBN: 978-989-758-713-9

methodology, including the rationale behind selecting these models and the theoretical framework. Subsequently, in Section 4, the experimental results are presented and analyzed. Finally, Section 5 concludes the study and includes the list of references cited in this paper.

2 RELATED WORK

Hotel booking cancellation prediction is influenced by multiple factors such as room type, number of occupants, and parking requirements. As more factors are considered, the research process becomes more straightforward. Previous work on hotel booking cancellation prediction has encompassed various solutions (Wu, 2021). In the realm of machine learning, Neslin and Gupta analyzed this issue using logistic regression and decision tree models, achieving satisfactory results. Verbeke al. conducted a comprehensive analysis of various models and found that the SVM-POLY algorithm performed the best (Verbeke, 2020). In a similar field, Coussement et al. applied the support vector machine model to newspaper subscription customer churn and achieved favorable outcomes. Additionally, Chinese scholar Bai compared logistic regression, decision tree, random forest, and XGBoost models, highlighting XGBoost's superior performance across various metrics such as AUC, ACC, Precision (Bai, 2021; Mishra, 2021; Ullah, 2019).

The establishment of customer churn models currently focuses primarily on finance and telecommunications industries, with a major emphasis on the role of machine learning models. This study, however, concentrates on the establishment of deep neural networks in hotel customer churn models and after comparing them with various machine learning models, discovers their excellent performance.

3 METHODOLOGIES

In this study, the dataset was first analyzed, and the input data was preprocessed. Subsequently, several machine learning models were constructed and trained, including linear regression (LR), random forest (RF), deep neural network (DNN), decision tree (DT), and gradient boosting decision tree (GBDT) models, to obtain results for further analysis. Figure 1 illustrates the workflow of this paper.

3.1 Data Analysis

The dataset was analyzed to understand the distribution of the data, the meanings of each feature column, and the correlations between variables.

3.2 Data Preprocessing

Before constructing and training the hotel customer churn model, data preprocessing was performed in this study. Since the used dataset had no missing values, there was no need for data cleaning. However, due to the large number of feature columns, Furthermore, after conducting calculations, it was found that some feature columns have low correlation.This study used the Pearson correlation coefficient method to remove columns with coefficients below 0.06.

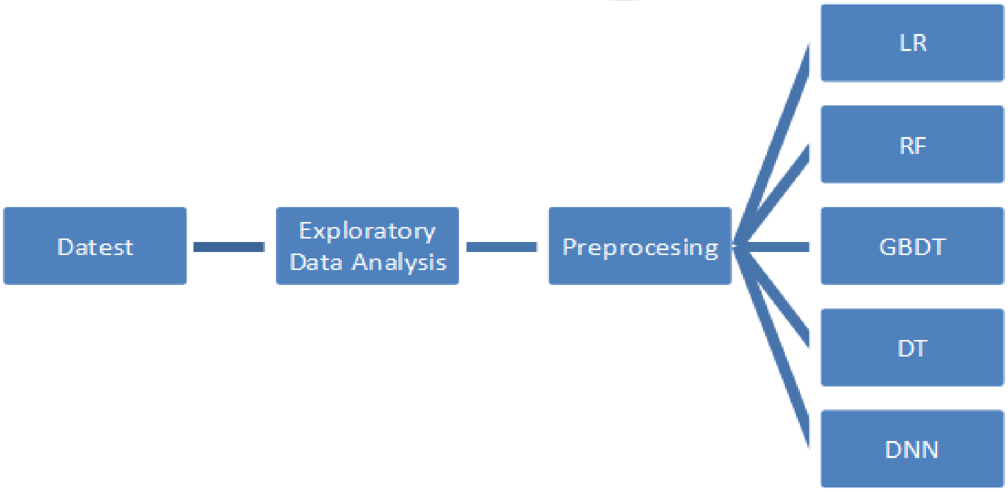

Figure 1: Research Workflow (Photo/Picture credit :Original).

Additionally, since the dataset contained various data types such as int and str, this experiment converted them all to int type. The specific descriptions of each feature column in the dataset will indicate the corresponding meanings of the int type data.

Considering the imbalanced nature of the dataset, this study employed three methods to balance the data: random undersampling, SMOTE oversampling, and random oversampling.

Apart from data scaling, the original dataset in this study was split into a training set (80%) and a test set (20%).

3.3 Model Selection and Construction

This study selected and implemented various ensemble learning models and machine learning models to predict customer churn in hotel bookings. The specific model selection and operational principles are as follows:

3.3.1 LR

LR is a commonly used statistical model for predicting which category a certain entity belongs to. In this model, features are associated with probabilities, where the probability represents the likelihood of a certain entity belonging to a specific category.

The specific operational steps involve assigning a weight to each feature, which represents the impact of that feature on the outcome. During the final calculation, each feature is multiplied by its weight, and then the results are summed. Subsequently, the result is input into the sigmoid function to obtain the probability value.

The sigmoid function (also known as the logistic function) is used to calculate the final probability. The output of the sigmoid function ranges from 0 to 1, representing the probability of a certain entity belonging to a specific category.

$$S(t) = \frac{1}{1+e^{-t}} \tag{1}$$

The sigmoid function, similar to the normal distribution, is the most typical representation among all probability distributions. However, the normal distribution comes with significant computational costs. Therefore, the sigmoid function, which is akin to the normal distribution, becomes the preferred choice.

3.3.2 DT

In the DT model, relevant questions suitable for learning are automatically proposed based on the features of the data. These questions are then structured into a tree-like format, where progression from one node to the next is only possible through the previous node, ultimately leading to the identification of the optimal classification.

One of the key metrics in the DT model is the Gini index . This index is utilized to measure the level of disorder within the system, enabling the model to formulate appropriate questions and carry out classification tasks.

$$gini(T) = 1 - \sum p_i^2 \tag{2}$$

3.3.3 RF

The RF model can be seen as a collaboration of numerous decision trees, each offering different solutions. These decision trees work together to provide diverse outcomes. Subsequently, random subsets of data and features are independently selected from each decision tree, introducing randomness into the process. When reaching the final decision, the ultimate result is determined through various methods such as averaging.

Formula for calculating the average value:

$$Imp(X_j) = \frac{1}{M}\sum_{m=1}^{M}\sum_{t\epsilon\varphi_m} 1(j_t = j)[p(t)\Delta i(s_t,t)] \tag{3}$$

3.3.4 GBDT

The GBDT model continuously enhances its efficiency and accuracy through iterative learning. Initially, the model generates a decision tree based on the given problem. Subsequently, it constructs the next tree with improved analytical capabilities based on the performance of the previous tree, iteratively learning and refining to ultimately produce the most efficient and capable decision tree.

Formula for calculating the parameter space of the model:

$$Wt = Wt - 1 - \rho t\nabla wL|_{w=wt-1} \tag{4}$$

This approach of boosting the performance of decision trees in the GBDT model results in a powerful ensemble learner that excels in predictive tasks across various domains. By sequentially building trees to correct the errors of the preceding ones, GBDT effectively combines the strengths of multiple weak learners to create a strong predictive model with high efficiency and accuracy.

3.3.5 DNN

The DNN model is inspired by the structure of the human brain, consisting of multiple layers, each containing numerous neurons. Broadly categorized into input layer, hidden layers, and output layer, the input layer receives raw data, which is then subjected to a series of nonlinear transformations by the hidden layers through activation functions, culminating in the output layer providing the model's prediction. Graph 3 and Graph 4 represent two commonly used activation functions, ReLU and Sigmoid. To enhance accuracy, there may be multiple hidden layers, as illustrated in Figure 2 with a DNN model having two hidden layers. During training, the DNN continuously adjusts the weights of connections between neurons through extensive data analysis to improve prediction accuracy.

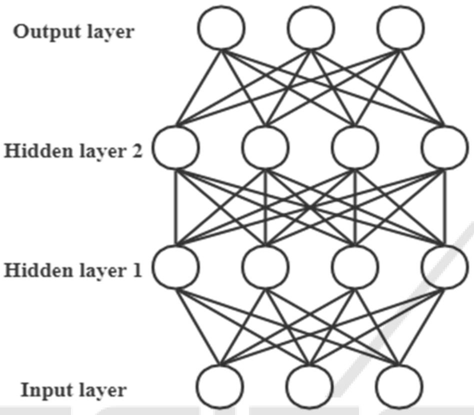

Figure 2: The construction of DNN (Photo/Picture credit :Original).

3.4 Evaluation Metrics

The metrics used for model evaluation are precision, recall, AUC, accuracy, f1-score.

Area Under the ROC Curve(AUC)

$$AUC = \frac{\sum_{i\epsilon positiveclass}^{n} rank_i - \frac{M(1+M)}{w}}{M \times N} \tag{5}$$

The AUC value ranges from 0 to 1, with a higher value indicating better performance of the model.

Classification Accuracy (accuracy)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

Accuracy ranges from 0 to 1, representing the probability of making correct predictions.

precision

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

Precision is a key metric for evaluating the accuracy of the model, focusing primarily on the model's precision. Since this study focuses on predicting hotel cancellations, author will only focus on precision1.

Recall

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

Recall is an important metric for evaluating the accuracy of the model, focusing primarily on the model's ability to find actual positive cases, i.e., the model's recall rate. Since this study focuses on predicting hotel cancellations, author will only focus on recall1.

f1-score

$$f1 - score = \frac{2 \times (accuracy \times AUC)}{accuracy + AUC} \tag{9}$$

This metric combines both recall and precision to provide a comprehensive analysis of performance. Since this study focuses on predicting hotel cancellations, author will only focus on f1-score1.

4 EXPERIMENTAL SETUP AND RESULTS

4.1 Dataset Overview

This study utilized the Hotel Reservations Dataset from the Kaggle website, which includes data from 36,275 customers who made online reservations at two hotels in Portugal, along with their actual check-in status and nineteen numerical attributes. After data preprocessing, six feature columns were removed.

The dataset provides comprehensive information on customer booking behavior and check-in outcomes for analysis and modeling. It serves as a valuable resource for understanding and predicting customer churn in the hotel industry.

In addition, this study also explored the correlations between all the feature columns to facilitate further training and selection of the hotel customer churn model. According to the correlation matrix (Figure 3 and table 1, no significant linear correlations were found among the features. Using the Pearson correlation coefficient method, feature columns with correlations below 0.06 were removed).

This analysis ensures that only relevant and highly correlated features are included in the model, improving its effectiveness in predicting customer churn. The findings from this analysis provide valuable insights for developing an accurate and robust hotel customer churn model.

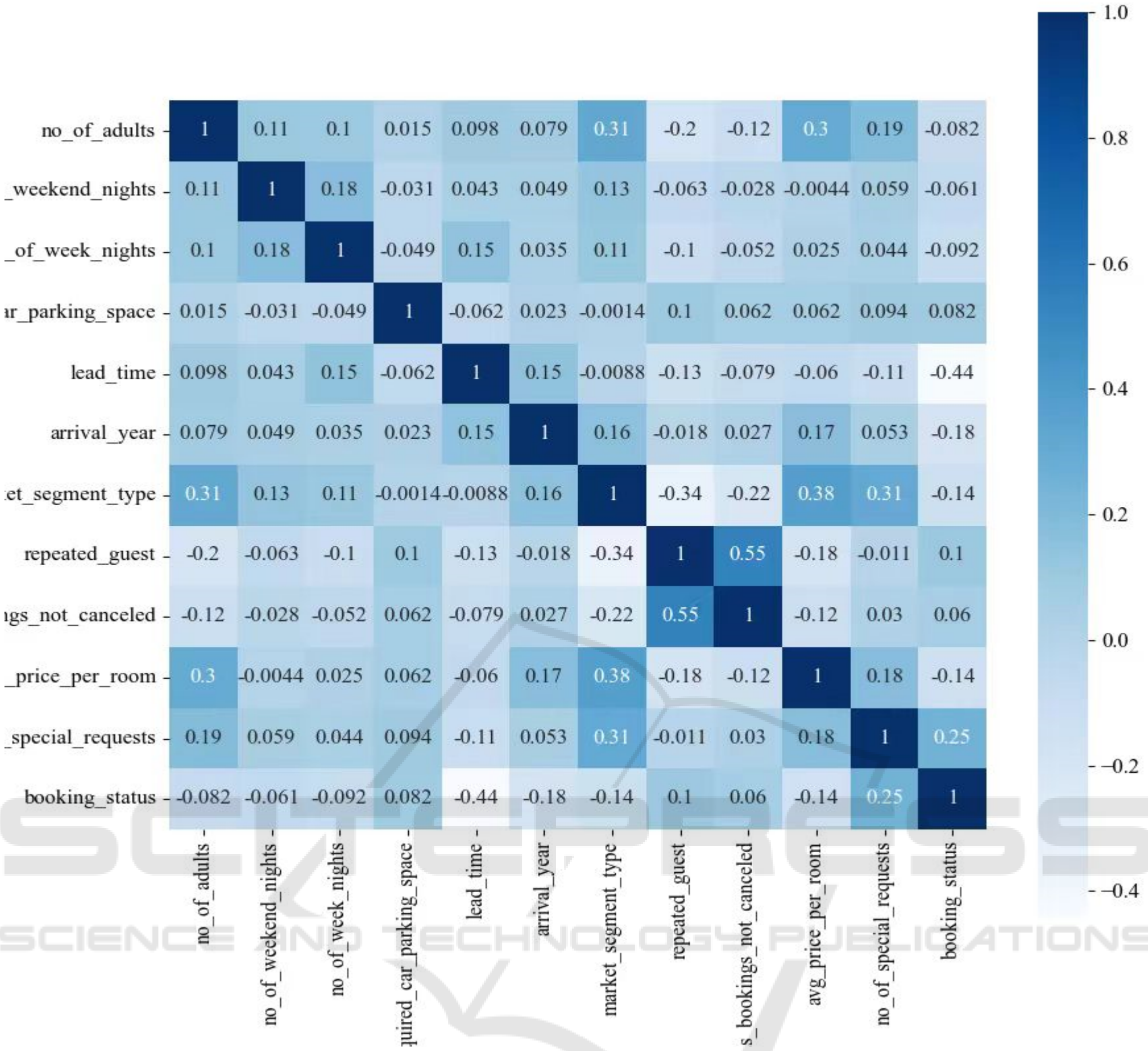

Figure 3: The correlation matrix (Photo/Picture credit :Original).

Table 1: Dataset attribute.

Attribute	Description
Booking_ID	unique identifier of each booking
no_of_adults	The number of adults specified by the customer during the hotel booking
no_of_weekend_nights	The number of weekends included in the booked hotel stay duration (including Saturday and Sunday)
no_of_week_nights	The number of weekdays selected by the customer for the hotel booking stay duration (including Monday to Friday)
required_car_parking_space	The customer's parking space requirement specified during the hotel booking? (0 - No, 1- Yes)
lead_time	Number of days between the date of booking and the arrival date
arrival_year	Year of arrival date
market_segment_type	Market segment designation (0 - Aviation,1 -Complementary,2-Corporate,3-Offline,4-Online）

repeated_guest	Whether the booked customers for this hotel are repeat guests? (0 - No, 1- Yes)
no_of_previous_bookings_not_canceled	The number of previous bookings used and checked in without cancellation
avg_price_per_room	The average price per room (in euros)
no_of_special_requests	The number of special requests from customers
booking_status	Whether to Check-in After Hotel Reservation(0 - No, 1- Yes)

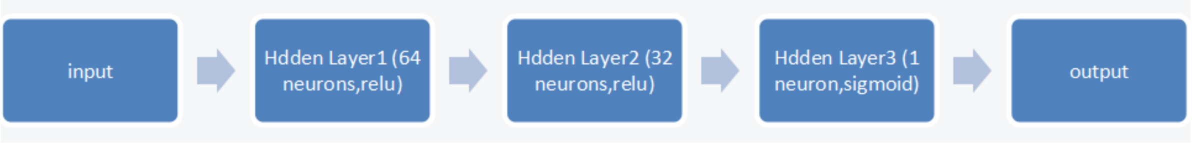
input
Hdden Layer1 (64 neurons,relu)
Hdden Layer2 (32 neurons,relu)
Hdden Layer3 (1 neuron,sigmoid)
output

Figure 4: The operational principle (Photo/Picture credit :Original).

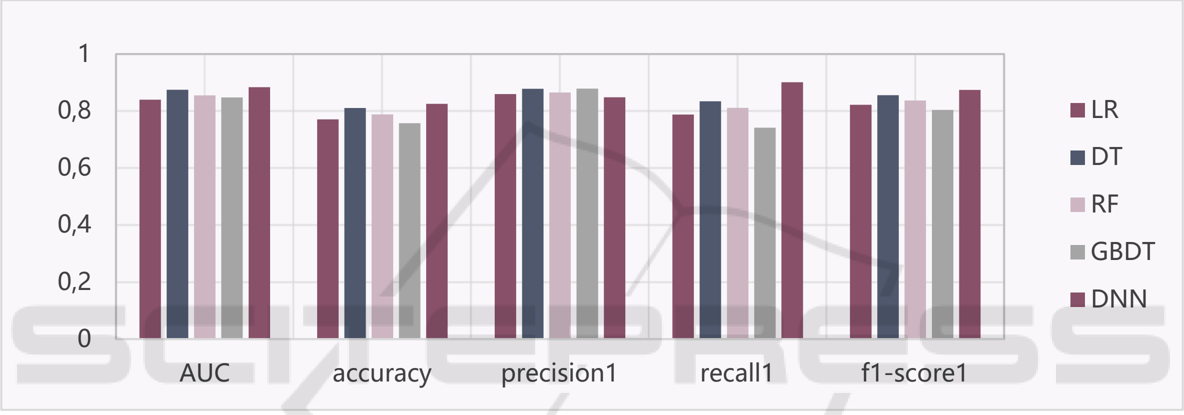

Figure 5: The result from different methods (Photo/Picture credit :Original).

4.2 Experimental Settings

In this experiment, all models were run in the Python IDE called PyCharm 2021.3, utilizing packages such as Pandas, Scikit-Learn, TensorFlow, and imbalanced-learn (imblearn).The specific parameter settings used are as follows:

- LR

Feature selection based on penalty terms was employed.

- DT

A total of 57 trees were generated

Three hidden layers were established, utilizing relu, relu, and sigmoid as activation functions, with 64 neurons, 32 neurons, and 1 neuron set for each layer respectively. The figure 4 depicts the operational principle.

4.3 Model Evaluation

These 5 models are evaluated using accuracy, AUC, recall (section 1), precision (section 1), f1-score(section 1). The results are demonstrated in Figure 5.

Among all the models, the DNN performs the best in the four evaluation metrics: AUC, Accuracy, recall1, and f1-score1. Although its precision1 value ranks last among the five models, the difference from the top-performing model is only 0.03. On the other hand, GBDT, which excels in precision1, lags behind in the other four metrics, possibly due to its low efficiency in parallel computation.

4.4 Feature Importance Exploration

For this study, DT and RF also demonstrate superiority in the four evaluation metrics, and they provide relative importance assessments for each feature in the dataset during the training process. Therefore, the list is presented here for reference purposes (Figure 6 and figure 7).Although DT and RF provide different results regarding the importance of each feature, the two models agree on the most important features measured, which are: lead_time, market_segment_type, avg_price_per_room, and no_of_special_requests. The reasons are as follows: Lead_time: Differences in arrival times may lead to customer loss in hotel bookings. For example, early arrivals with no available rooms may result in customers canceling their reservations and seeking alternative accommodations. Market_segment_type: Variances in market segmentation can also contribute to customer booking cancellations. Online bookings, for instance, are more prone to cancellation, whereas offline bookings, due to their complexity, may lead customers to cancel after weighing the pros and cons. Avg_price_per_room: Considerations regarding room prices at hotels may also be a significant factor in customer loss, as customers assess the value for money of the hotel rooms. No_of_special_requests: Fulfilling special requests from hotel guests upon arrival or before arrival may also lead to customer loss, as the satisfaction of such requests could influence customer decisions to cancel.

The above importance analysis was derived using LR and RF models.

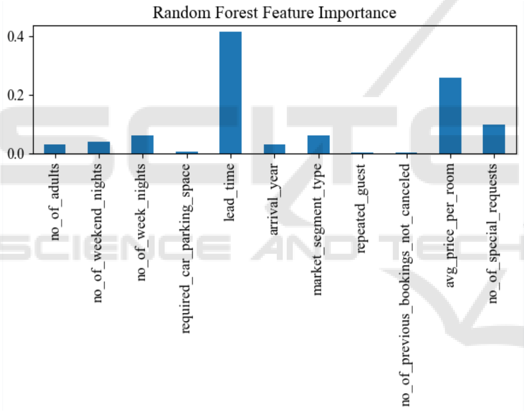

Figure 6: RR feature importance (Photo/Picture credit: Original).

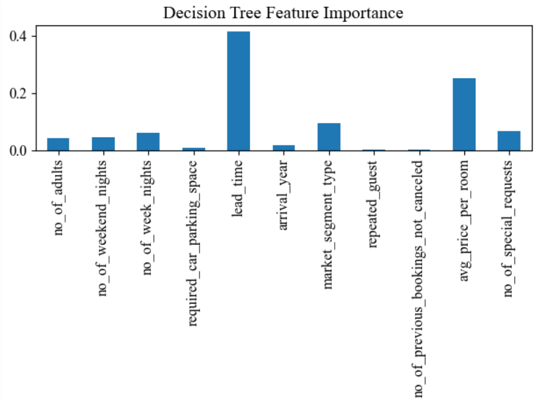

Figure 7: DT feature importance (Photo/Picture credit : Original).

5 CONCLUSION

In summary, this study combines machine learning and deep neural network methods to predict house prices in Miami. The models utilized include LR, RF, GBDT, DNN, and DT. Among all these models, the neural network performs the best, while the two machine learning methods - RF and DT - also excel in predicting customer churn in hotel bookings. In this research, the DNN deep neural network is configured with three hidden layers, each with varying numbers of neurons to enhance efficiency and accuracy. Ultimately, impressive results are achieved with an AUC of 0.883, an accuracy of 0.825, a precision1 of 0.848, a recall1 of 0.901, and an f1-score of 0.873. The strong performance of RF and DT also aids in providing insights into relative importance, revealing that lead_time, market_segment_type, special requests, and room price are the four most critical features influencing customer churn in hotel bookings. Additionally, the study explains how each factor impacts reality.

REFERENCES

Bai, R. 2021, Customer Churn Prediction and Retention Strategies in Bai Ruirui Hotel Booking Platform (Thesis). *Zhengzhou University*.

Chen, Y., Trivedi, M., Kalaida, N., & Sinha, A. P. 2024, Predicting customer churn in subscription-based business using gradient boosting. *Journal of Business Research*, 92, 400-407.

Fu, X., Gu, X., & Li, Y. 2020, Customer churn prediction in telecommunications based on gradient boosting tree algorithm. *Telecommunication Systems*, 68(1), 91-106.

Gordini, N., & Veglio, V. 2023, Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. *Industrial Marketing Management*.

Hwang, H., & Lee, D. 2020, Customer churn prediction using deep learning models in the online game industry. *Expert Systems with Applications*, 105, 159-173.

Li, C., & Liu, H. 2018, A churn prediction model in e-commerce industry based on deep learning. In 2018 *IEEE International Conference on Big Data* (Big Data) 1645-1650.

Mishra, A., Dash, M., & Mishra, D.2020. Churn prediction and customer segment identification using deep learning for telecommunication industry. *Journal of King Saud University-Computer and Information Sciences,* 31(4), 424-441.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. u., & Kim, S. W. 2019. A Churn Prediction Model Using Random Forest: Analysis of Machine Learning

Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access*.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. 2020. Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications.*

Wu, X., & Yu, L. 2021. Ensemble learning based churn prediction model in telecommunication industry. *Information Systems Frontiers*, 18(1), 163-175.