
**Electronic Journal of Applied Statistical Analysis**
**EJASA, Electron. J. App. Stat. Anal.**
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v17n3p621

# Predicting Domestic Tourists' Length of Stay in Italy leveraging Regression Decision Tree Algorithms

By Antolini, Cesarini

15 December 2024

*Electronic Journal of Applied Statistical Analysis*
Vol. 17, Issue 03, December 2024, 621-635
DOI: 10.1285/i20705948v17n3p621

# Predicting Domestic Tourists' Length of Stay in Italy leveraging Regression Decision Tree Algorithms

Fabrizio Antolini*[a] and Samuele Cesarini[a]

[a]*Department of Business Communication, University of Teramo, Campus Aurelio Saliceti, Via Renato Balzarini, 1- Teramo, Italy*

15 December 2024

This study innovates in predicting domestic tourists' Length of Stay (LoS) in Italy by using decision tree models, addressing the gap in understanding LoS's determinants, and improving upon inconsistent results from traditional parametric analyses. Utilizing the 2019 "Viaggi e Vacanze" survey by the Italian National Institute of Statistics and categorizing variables into sociodemographic, economic, travel-related, and psychological factors, the research applies one-hot encoding to analyse 48,410,000 trips. Through evaluating random forest and gradient boosting models, the study highlights their superiority in identifying complex data patterns, offering actionable insights for tourism policymakers. These models enable precise LoS estimation, facilitating enhanced strategic planning for extending stays, optimizing services, and improving promotional efforts to maximize tourism's economic impact. This approach offers a comprehensive tool for developing policies that boost visitor engagement and economic benefits, showcasing a significant advancement in tourism management practices.

**keywords:** Microdata, Length of stay, Machine-learning models, Decision trees,Tourism sector.

## 1 Introduction

Tourism is a vital economic sector that contributes significantly to the global economy (WTO, 2021). Accurately predicting the length of stay (LoS) of domestic tourists plays

*Corresponding author: fantolini@unite.it


ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

a pivotal role in tourism planning, resource allocation and destination management (Li et al., 2018; Gössling et al., 2018). Understanding the factors that influence tourists' LoS enables policymakers and industry stakeholders to make informed decisions regarding infrastructure development, marketing strategies and service provision (Wang et al., 2018). By accurately forecasting tourists' LoS , destinations can optimize their tourism offerings, enhance visitor experiences, and maximize economic benefits. Various scholars have emphasized the significance of LoS in terms of the income generated by tourists at a destination (Marrocu et al., 2015; Aguiló et al., 2017; Park et al., 2020; Antolini et al., 2024). Indeed, it should be noted that LoS is not only a measure of tourists' engagement but also directly impacts the economic benefits derived from their expenditures in the local economy. As highlighted by Gössling et al. (2018), shorter stays can in fact stimulate heightened demand for transport infrastructure, necessitating things such as increased airport capacity. Moreover, shorter stays tend to restrict tourists to popular attractions, often overshadowing lesser-known regions and attractions, leading to an uneven distribution of tourist flow. This concentration of visitors in specific areas can exacerbate the issue of overtourism, characterized by overcrowding at certain destinations and stagnation at others (Oklevik et al., 2021). In contrast, tourists who opt for longer stays can explore a wider array of smaller businesses in peripheral locations. Their extended LoS allows for a more comprehensive exploration of the destination, fostering a deeper understanding and appreciation of its unique characteristics.

While these conclusions are generally shared by the scientific community, the practical implications lead many researchers to employ different statistical models and econometric methodologies to explore various factors and their impact on the predicted LoS of tourists. To date, a substantial body of literature has focused on parametric modelling of the relationships between tourists' LoS and the identified determinant variables. However, the review of existing studies indicates that the precise relationships between the determinants and the prediction of LoS have not yet been fully elucidated. Moreover, it appears there is a certain lack of consensus regarding the results obtained, which often show positive or negative associations according to the methodology employed. This inconsistency highlights the limitations of traditional parametric models in capturing the complex patterns and nonlinear relationships inherent in LoS prediction. Therefore, this study aims to bridge this research gap by proposing a novel predictive approach based on the implementation of decision tree models to estimate the LoS of domestic tourists in Italy using microdata obtained from the 2019 "Viaggi e Vacanze" (VV) survey conducted by the Italian National Institute of Statistics (ISTAT, 2022). The main objective is to develop predictive models based on decision tree methodology, incorporating the most relevant predictive variables identified in the scientific literature. By harnessing the power of machine learning, which excels in uncovering complex patterns, this study aims to overcome the limitations of traditional models and deliver an accurate and reliable predictive model.

The remainder of this paper is organized as follows. Section 2 provides a summary of the main studies in the literature, with specific reference to the models employed. This section also discusses the key determinants used to reconstruct the set of variables to be included in decision tree models. Section 3 provides a general overview of the microdata

used and the variables included in the models. This is followed by an overview of the methodology. Section 4 presents the results, with a particular focus on the root mean square error (RSME) and the predictive accuracy of the models. Section 5 concludes by presenting the main findings and discussing the practical and theoretical implications for tourism policymakers.

## 2 Literature Review and Theoretical Framework

### 2.1 The parametric modelling and determinants of LoS in the literature

Accurate prediction of tourists' Length of Stay (LoS) is a crucial aspect of both tourism research and the tourism industry. In the last 20 years, numerous studies have employed parametric models to predict tourists' LoS, including linear regression, survival analysis, and count data regression models combined with survey data. Linear regression models have been widely used in this context to establish relationships between predictor variables and LoS, assuming a linear association between them. However, it is important to note that these models often assume linearity and require strict assumptions about the underlying data distribution. Survival analysis models, such as Cox proportional hazards models, have also been used and are particularly useful when dealing with censored data and time-to-event outcomes. For these reasons, these models provide insight into the duration of tourist stays, accounting for the possibility that some stays may not be fully observed. Count data regression models, such as Poisson and negative binomial models, have been employed when LoS was measured in terms of discrete counts. In fact, these models are designed to handle data that follows a count distribution and can provide, in this case, valuable insights regarding the factors influencing LoS.

Examining the literature on this subject from the 2000s, it is worth noting that initial studies predominantly employed survival analysis techniques to examine the LoS. Gokovali et al. (2007) utilized the Cox and Weibull survival models to examine the duration of tourists' stays in Bodrum, Turkey. They identified significant associations between the LoS and factors such as education, income, experience, familiarity, and daily spending. Martinez-Garcia and Raya (2008) studied low-cost tourism in Spain and found that variables such as nationality, age, education level, accommodation type, season, and geographic area influenced the LoS. De Menezes et al. (2008) analysed the LoS of tourists in the Azores Islands, considering sociodemographic profiles, trip attributes, sustainability practices, and destination image attributes using a Cox proportional hazard model. Barros et al. (2008) and Barros et al. (2010) investigated the LoS of Portuguese tourists in South America and the Algarve, Portugal, respectively. They used survival models (the Cox model, the Weibull model, and the Weibull model with heterogeneity) and found that the LoS depended on multiple determinants specific to each destination. Raya (2012) focused on the LoS of participants at the International Triathlon Challenge Barcelona–Maresme and identified factors such as satisfaction, foreign participant status, and expenditure as influential. Peypoch et al. (2012) employed a multivariate fractional polynomial duration model in studying tourists' LoS in Madagascar and found

that higher income, older age, male gender, and education level were associated with longer stays. Thrane (2012) investigated the LoS of international summer visitors in Norway, highlighting the impact of nationality, age, spending patterns, and other trip-related characteristics. Thrane and Farstad (2012) explored the relationship between LoS and variables such as previous visits, places visited, satisfaction, and expenditures per day, noting a positive association with the former and a negative association with the average expenditures. Interestingly, Thrane's study suggested that ordinary least squares (OLS) regression models provided an effective description of the impact of independent variables on LoS, comparable to survival models. In fact, Thrane argued that OLS regression models had an advantage over survival models because they allowed for negative impacts of independent variables on the dependent variable, which survival models did not accommodate. Consequently, Thrane recommended that future studies on tourists' LoS should move away from survival models, in line with the principle of parsimony. From this point onwards, numerous studies employing count data regression models have appeared in the literature, even though Thrane (2015) demonstrated in his study that this methodology did not lead to superior results compared to OLS and survival models, thus raising doubts about their future utility. Alén et al. (2014) conducted a study on senior tourists in Spain and identified various factors influencing the LoS, including age, travel purpose, climate, accommodation type, group size, trip type, and activities at the destination. They utilized the negative binomial model for the analysis. Kruger and Saayman (2014) investigated the determinants of LoS at Kruger National Park in South Africa, employing a Poisson regression model. The authors found that sociodemographic characteristics, behavioural variables, and geographical factors influenced LoS. Prebensen et al. (2015) used truncated negative binomial regression to examine the relationship between LoS and sociodemographic and travel behaviour variables in northern Norway. The results indicated that only gender significantly influenced LoS, with females staying for shorter durations. Rodriguez et al. (2018) conducted an extensive study in Santiago de Compostela, Spain utilizing probit, truncated regression, and Heckman models. They confirmed the impact of personal characteristics, travel attributes, and destination factors on LoS. Soler et al. (2020) applied zero-truncated negative binomial and Poisson-inverse Gaussian regression models their study in Malaga, Spain and found that tourists' mode of transport, income, age, and climate of origin significantly influenced LoS. Bavik et al. (2021) measured LoS in Macau, China using a Poisson regression model and observed that larger travel groups and lower spending were associated with longer stays. Atsíz et al. (2022) analysed LoS in Istanbul by employing classical binary logit for group membership and a zero-truncated Poisson model for visitors who stay longer, thus revealing the drivers of LoS.

## 2.2 A new approach to modelling LoS

As previously observed, the use of parametric models to predict the influence of determinants on the predicted LoS is widely accepted. However, while parametric models offer valuable insights and allow for quantitative predictions and statistical inference, they have certain limitations. One main limitation is their assumption of linearity, which

may not capture complex nonlinear relationships and interactions among variables. In situations where the relationship between predictors and LoS is nonlinear, parametric models may not accurately capture the underlying patterns. Additionally, parametric models often require strict assumptions about the data distribution, which may not hold in real-world scenarios. These assumptions can limit the applicability of the models and introduce uncertainty in the predictions. Recently, there has been a shift in the approach to tourism research with the emergence of studies employing innovative methodologies, namely classification and regression trees (CARTs) (Breiman, 2017). While traditional regression methods have commonly been used for tourism market segmentation, Díaz-Pérez et al. (2021), Díaz-Pérez et al. (2020), and Díaz-Pérez and Bethencourt-Cejas (2016) have demonstrated the significant advantages of this technique. Although not systematically employed, this methodological approach has been used to predict tourists' LoS. In their study, Lee and Kim (2021) employed a decision tree machine-learning algorithm to examine the relationship between geographical distance and travellers' hotel stay duration. Jackman and Naitram (2023) employed regression tree models to investigate the impact of tourists' sociodemographic profiles, trip-related characteristics, distance, and economic conditions on predicting LoS. Their analyses revealed significant heterogeneity that would typically remain undisclosed when using simplistic parametric approaches, such as OLS, commonly employed to model LoS. These are just two seminal examples of the potential of using this methodology. Therefore, this study aims to demonstrate the superior predictive capacity of these types of models when compared to parametric techniques. Machine-learning algorithms enable researchers to formulate predictive models and identify target group members that share similar characteristics. This involves stratifying or segmenting the predictor space into several simple regions. In the case under study, this methodology allows estimating tourists' LoS using the mean of the training observations in the region to which it belongs. As the set of rules employed to segment the predictor space can be depicted as a tree, this approach is commonly referred to as the decision tree method. In a visual representation, internal nodes signify divisions of the original observations, while the leaf nodes signify predictions of the target variable—namely, the LoS. Thus, this methodology enables a nuanced prediction, influenced by specific key variables (Antolini et al., 2024). In our study, the regression tree is constructed using a set of variables derived from the VV survey. Operationally, as described by James et al. (2020), they segment the predictor space into distinct regions to minimize the residual sum of squares (RSS) associated with predictions. This process involves iterating over all predictors and their possible split points to find the combination that offers the least RSS. The algorithm splits the feature space into two regions at each step, aiming to isolate subsets of data that are as homogeneous as possible in terms of the response variable. Each split is chosen to best separate the data into groups that reduce the overall prediction error, with the final model presenting a tree-like structure where each terminal node, or leaf, represents a prediction based on the mean outcome of the observations falling into that segment. This iterative process continues until a stopping criterion is met, such as reaching the maximum depth of the tree or encountering a minimum number of observations in a leaf node (Lewis, 2000).

# 3 Data and Method

## 3.1 Dataset and predictor variables

For our analysis, we leveraged a comprehensive dataset of microdata compiled from the 2019 VV survey conducted by the Italian National Institute of Statistics (ISTAT). This dataset encompasses a wide range of variables, including socio-demographic, economic, and travel-related information, providing a rich foundation for exploring domestic tourists' behaviours and preferences in Italy (Marcussen, 2011; Lin et al., 2020). The dataset utilized for the analysis included 3,001 observations from domestic tourist trips by residents. By applying expansion coefficients to each record, we estimate the total population of Italian domestic tourists. This resulted in a dataset representing 48,367,000 tourist trips, lasting up to 35 days. The following categories and variables were employed (Table 1):

Table 1: Categories and variables influencing LoS

| **FACTORS** | **VARIABLES** |
|---|---|
| Sociodemographic | Region of origin |
| | Marital status |
| | Age |
| | Sex |
| | Nationality |
| Economic | Education |
| | Occupation |
| Travel-related | Principal activities |
| | Group of travel party |
| | Accommodation organization |
| | Transport organization |
| | Transport |
| | Accommodation |
| | Month of travel |
| | Region of destination |
| Psychological | Travel motivation |

*Source: Authors' elaboration*

To provide a more analytical interpretation of the determinants of tourists' LoS, it is possible to recode the entire dataset by defining dummy variables based on each level determined by the categorical variables used. This can be done using techniques such as one-hot encoding, converting each category into a binary vector representation where

each element indicates the presence or absence of that category. In other words, for each categorical variable, several columns were created, each of which takes a value of 1 if the associated condition is met and 0 otherwise. For example, if the considered variable is the gender of the tourist and the recorded values are 'male' and 'female,' two columns were created, one for each gender, which takes a value of 1 for the specific gender and 0 otherwise. This technique allowed for a more detailed analysis of the impact that each categorical variable had on the tourists' LoS. After encoding the dataset using the aforementioned method, the number of rows remained the same, while the number of columns (representing variables) increased from 16 (original variables) to 128 (dummy variables). By converting categorical variables into numerical representations, machine-learning techniques can effectively analyse intricate relationships and patterns within categorical inputs.

### 3.2 Decision tree algorithms: Random Forest and Gradient Boosting

The presence of nonlinearity between the LoS and independent variables coupled with the intricate nature of their relationships advocates against making a priori assumptions based on parametric models. Instead, employing a CART methodology is more suitable even if they are not without their challenges. One of the most notable issues associated with CART models is their susceptibility to overfitting (James et al., 2020). To address this concern, ensemble models have been introduced as a standard solution in the realm of machine learning. This approach combines multiple learning algorithms to obtain better predictive performance by averaging out biases, reducing variance, and generally generating a more comprehensive model that captures the underlying patterns in the data without overfitting. Among the most used ensemble models are Random Forest and Gradient Boosting, both of which extend the decision tree methodology to create a more accurate, and reliable predictive tool.

In our study, Random Forest (RF) operates by constructing a multitude of decision trees at training time and outputting the mean prediction of the individual trees. RF correct for decision trees' issue of overfitting to their training set by introducing randomness in two ways: by selecting a random subset of the training data to build each tree (bootstrap aggregating or bagging) and by choosing a random subset of features to consider at each split in the learning process. This randomness helps to make the model more robust and less prone to overfitting, ensuring that the biases of individual trees are minimized through averaging.

Gradient Boosting, on the other hand, is a sequential technique where each subsequent model attempts to correct the errors of the previous models. It builds one tree at a time, where each new tree helps to correct errors made by previously trained trees. This method involves optimizing a loss function, an aspect absent in RF, where the focus is on reducing error by averaging the outcomes of multiple trees. Gradient Boosting (GB) models often use shallow trees as base learners, which are computationally efficient and reduce the risk of overfitting. However, they require careful tuning of parameters such as the learning rate and the number of trees to balance the bias-variance trade-offs effectively. For further insights into the methodology, it is recommended to refer to

specialized texts on the subject (Breiman, 1996, 2001; James et al., 2020).

To assess the predictive performance of the three models, we constructed two randomized samples, the training set and the test set, containing 75% and 25% of the total observations, respectively. The test set was employed to make predictions for unseen data instances. As suggested in Lantz (2019), as this is a numerical prediction problem rather than a classification problem, it is not possible to use a confusion matrix to evaluate the accuracy of the model. Instead, it is necessary to measure the correlation between the model's predictions on the test set and the actual recorded values. This provides an indication of the strength of the linear association between the two variables. Therefore, the predictive capacity of the models was evaluated by visually examining a scatter plot depicting the predicted values against the corresponding actual values. Finally, a comparison between the models was conducted by assessing the standard deviation of the generated residuals.

## 4 Results

Decision tree models may lead to better results in a case like this, where the relationships between LoS and individual variables do not appear to be linear. To address the problem of overfitting, some constraints were implemented in the RF and GB algorithms, as highlighted by Probst et al. (2019). The RF model consisted of 500 regression trees, where the maximum number of nodes in each tree was limited to 300. Moreover, each leaf of the tree had to contain a minimum of 100 observations. These constraints were put in place to create a moderately complex model that avoided capturing small subgroups that could be influenced by noise. In addition to the complexity constraints, a random selection process was utilized to determine the variables used for splitting observations in each tree. Of the 128 explanatory variables, only 43 were randomly chosen for each tree. This approach significantly improved the model's predictive capability and reduced the impact of irrelevant or uninformative variables, as it ensured that only a subset of variables was considered for each tree. In contrast, the GB algorithm aimed to train the regression tree 500 times. The objective was to progressively decrease the RMSE using a learning rate of 0.05. The tree depth, number of usable variables and number of observations remained consistent with the RF algorithm. These modifications were implemented to enhance the performance and generalization ability of the models, striking a balance between complexity and accuracy. By avoiding excessive complexity, the models were able to effectively capture the underlying patterns in the data and make accurate predictions. Below (Table 2), the RMSE and the correlation between the original and predicted values from the two previously described machine-learning decision tree models are presented.

As already highlighted, to ensure the reliability of these models, it is necessary to evaluate their accuracy on unseen data to avoid overfitting to the training data. This practice enables optimal generalization of the results to the entire population under study. The RF and GB models exhibited strong positive linear relationships between predicted and actual values, as indicated by the high correlation coefficients (0.93 and

Table 2: Decision tree model results

| **Model** | **Root mean square error** | $\rho$ |
|---|---|---|
| Random Forest | 2.05 | 0.93 |
| Gradient Boosting | 1.09 | 0.98 |

*Source: Authors' elaboration*

0.98). This demonstrates the models' effectiveness in capturing relevant features and patterns related to LoS in the domestic Italian tourism context. Additionally, the relatively low RMSE values (2.08 and 1.09) suggest that the models' predictions closely align with the actual LoS, indicating accurate representation of underlying data patterns and trends. In this regard, Fig. 1 and Fig. 2 depict, through a scatter plot, the relationship between the actual observed values in 25% of the test set and those estimated by the RF and GB models.

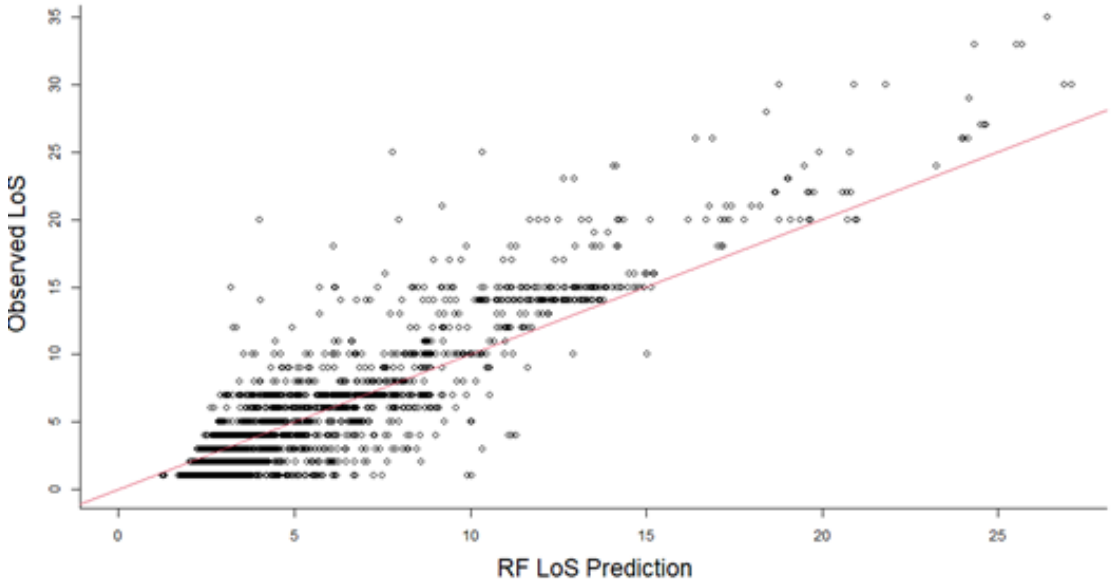


Figure 1: Scatter plot predicted LoS by RF against real values

*Source: Authors' elaboration*

By observing the arrangement of the points around the line, it can be inferred that the model is capable of adequately predicting unseen values and therefore can be used to provide forecasts regarding tourists' LoS. Both models allow the extraction of important metrics for each variable's contribution to the model's predictions. The predictive interpretability of the variables is evaluated by analysing the mean squared error (MSE)

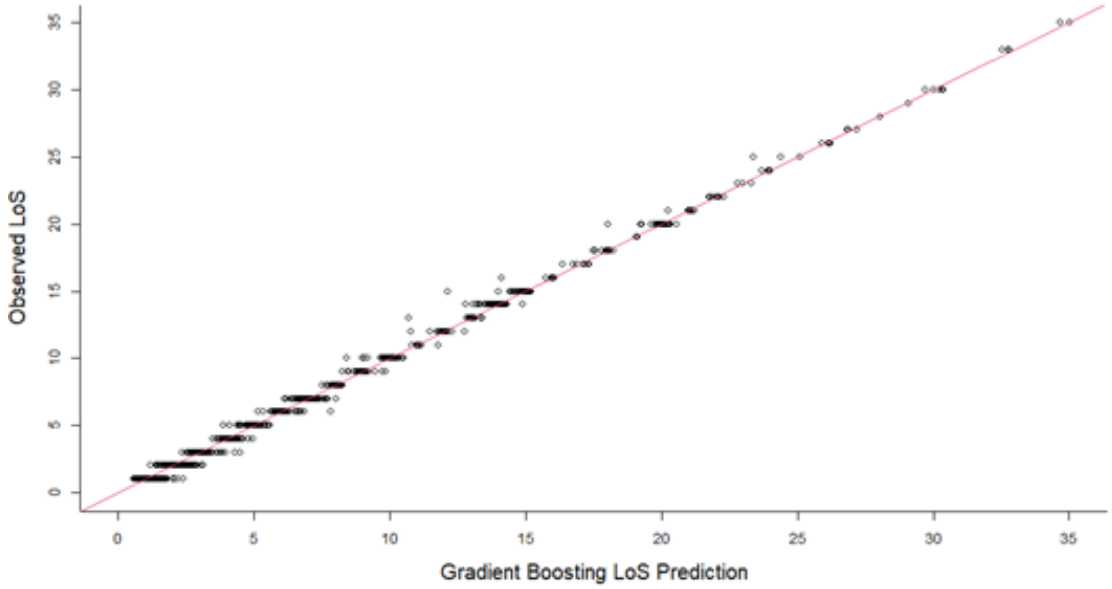


Figure 2: Scatter plot predicted LoS by GB against real values

*Source: Authors' elaboration*

of the predictions obtained when a particular variable is excluded from the model. A graphical interpretation is shown in Fig. 3 and Fig. 4.

Following the variable importance in predicting the LoS calculated by the RF model, we can understand which input variables have the most significant impact on the model's predictions in terms of loss of accuracy (increase in MSEs) if the variable is not used. Months of travel (August, July and June) have the highest importance score (136.33, 54.29 and 41.23), indicating that trips made in summer significantly influence the model's predictions. The number of family members (49.97) and travellers aged over 75 (44.43) both have substantial importance scores, indicating that having three family members on the trip and the age of travellers are significant factors affecting the LoS. The region of residence, Lombardy (43.26) or the chosen destination region—Calabria (41.94), Lazio (41.30) or Apulia (40.19)—all have notable importance values. These scores emphasize that both travellers' region of origin and their selected destination region significantly influence the model's predictions regarding the LoS.

Figure 4 shows the gain calculated by the GB, a metric used to measure the contribution of a particular feature (variable) to the model's predictive performance. The algorithm evaluates how much it can reduce the overall MSE by making a split based on that feature for all possible splits and then aggregates and normalizes the gains for each feature across all possible splits by dividing them by the number of observations in the dataset. Once the gains are calculated and normalized for all features, they can be ranked in descending order. Features with higher gain values are considered more

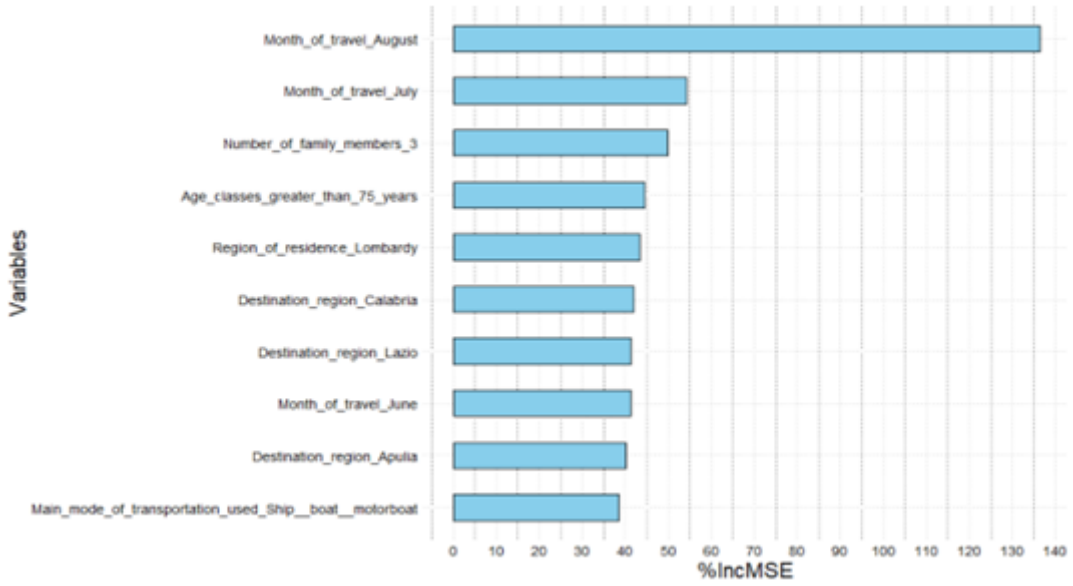


Figure 3: RF estimates of variable importance

*Source: Authors' elaboration*

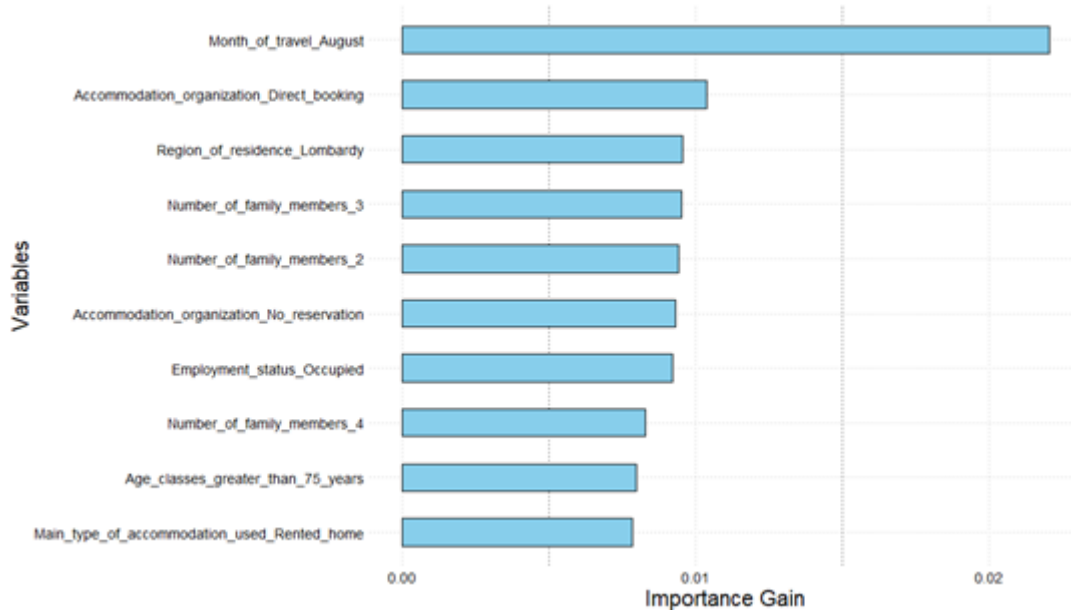


Figure 4: Gradient boosting estimates of variable importance

*Source: Authors' elaboration*

important because they contribute more to reducing the model's loss. August still has the highest score (0.0221). The accommodation organization (direct booking or without a reservation) and the main type of accommodation (rented home) have higher scores (0.0104, 0.0078 and 0.0093) signifying that these factors contribute to the model's predictive accuracy. Similar to the RF calculation, the number of family members and age over 75 emerged as important in the GB importance calculation.

## 5 Conclusion

Accurately predicting domestic tourists' LoS is of utmost importance for various aspects of tourism management, such as planning, resource allocation and destination development. To make informed decisions about infrastructure, marketing strategies and service provision, policymakers and industry stakeholders must understand the factors that influence LoS. While numerous researchers have employed statistical models and econometric methodologies to examine these factors, the existing literature lacks a comprehensive understanding of the precise relationships and predictions relating to LoS. Moreover, different studies often yield inconsistent results due to the limitations of traditional parametric models in capturing the complex and nonlinear nature of LoS.

To address this research gap, this study introduced a novel predictive approach using decision tree models to estimate the duration of domestic tourists' stays in Italy. We utilized microdata from the 2019 VV survey. Based on a comprehensive review of previous studies, we classified a set of explanatory variables into four main categories: sociodemographic factors, economic factors, travel-related factors, and psychological factors. The entire dataset was then recoded using one-hot encoding to provide a more analytical interpretation of the determinants of tourists' LoS. This resulted in the creation of 128 dummy variables for 48,410,000 trips.

The results indicated that the decision tree models—the RF and GB models—revealed strong positive linear relationships between predicted and actual values. This was evident from the high correlation coefficients (0.93 and 0.98), demonstrating the effectiveness of these models in capturing relevant features and patterns related to LoS in domestic Italian tourism. Additionally, the relatively low RMSE values (2.05 and 1.09) suggested that the predictions generated by the models closely aligned with the actual LoS, indicating an accurate representation of the underlying data patterns and trends.

These findings have practical implications for policymakers in the tourism industry. The proposed predictive models utilizing 128 dummy variables enable the estimation of tourists' LoS by inputting binary values based on the characteristics of the target tourist population. Specifically, by entering 1 where the conditions of the tourist whose LoS is to be predicted are met and 0 where they are not, these models provide a prediction regarding their stay. In contrast to traditional linear models that often yield unsatisfactory results, the accurate prediction of tourists' LoS achieved using machine-learning decision tree models can significantly inform and guide

## References


Aguiló, E., Rosselló, J., and Vila, M. (2017). Length of stay and daily tourist expenditure: A joint analysis. *Tourism Management Perspectives*, 21:10–17.

Alén, E., Nicolau, J. L., Losada, N., and Domínguez, T. (2014). Determinant factors of senior tourists' length of stay. *Annals of Tourism Research*, 49:19–32.

Antolini, F., Cesarini, S., and Simonetti, B. (2024). Factors determining italian tourists' expenses: A machine learning approach. *Quality & Quantity*.

Atsíz, O., Leoni, V., and Akova, O. (2022). Determinants of tourists' length of stay in cultural destination: One-night vs longer stays. *Journal of Hospitality and Tourism Insights*, 5(1):62–78.

Barros, C. P., Butler, R., and Correia, A. (2010). The length of stay of golf tourism: A survival analysis. *Tourism Management*, 31(1):13–21.

Barros, C. P., Correia, A., and Crouch, G. (2008). Determinants of the length of stay in latin american tourism destinations. *Tourism Analysis*, 13(4):329–340.

Bavik, A., Correia, A., and Kozak, M. (2021). What makes our stay longer or shorter? a study on macau. *Journal of China Tourism Research*, 17(2):192–209.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

De Menezes, A. G., Moniz, A., and Vieira, J. C. (2008). The determinants of length of stay of tourists in the azores. *Tourism Economics*, 14(1):205–222.

Díaz-Pérez, F. M. and Bethencourt-Cejas, M. (2016). Chaid algorithm as an appropriate analytical method for tourism market segmentation. *Journal of Destination Marketing & Management*, 5(3):275–282.

Díaz-Pérez, F. M., Fyall, A., Fu, X., García-González, C. G., and Deel, G. (2021). Florida state parks: A chaid approach to market segmentation. *Anatolia*, 32(2):246–261.

Díaz-Pérez, F. M., García-González, C. G., and Fyall, A. (2020). The use of the chaid algorithm for determining tourism segmentation: A purposeful outcome. *Heliyon*, 6(7):1–11.

Gokovali, U., Bahar, O., and Kozak, M. (2007). Determinants of length of stay: A practical use of survival analysis. *Tourism Management*, 28(3):736–746.

Gössling, S., Scott, D., and Hall, C. M. (2018). Global trends in length of stay: Implications for destination management and climate change. *Journal of Sustainable Tourism*, 26(12):2087–2101.

ISTAT (2022). Viaggi e vacanze: File ad uso pubblico. https://www.istat.it/it/archivio/178695.

Jackman, M. and Naitram, S. (2023). Segmenting tourists by length of stay using regression tree models. *Journal of Hospitality and Tourism Insights*, 6(1):18–35.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2020). *An introduction to statistical learning*. Springer.

Kruger, M. and Saayman, M. (2014). The determinants of visitor length of stay at the kruger national park. *Koedoe*, 56(2):1–11.

Lantz, B. (2019). *Machine learning with R: Expert techniques for predictive modeling*. Packt Publishing Ltd.

Lee, Y. and Kim, D. Y. (2021). The decision tree for longer-stay hotel guest: The relationship between hotel booking determinants and geographical distance. *International Journal of Contemporary Hospitality Management*, 33(6):2264–2282.

Lewis, R. J. (2000). An introduction to classification and regression tree (cart) analysis. In *Annual meeting of the society for academic emergency medicine in San Francisco, California*, volume 14. Department of Emergency Medicine Harbor-UCLA Medical Center Torrance.

Li, K. X., Jin, M., and Shi, W. (2018). Tourism as an important impetus to promoting economic growth: A critical review. *Tourism Management Perspectives*, 26:135–142.

Lin, V. S., Qin, Y., Li, G., and Wu, J. (2020). Determinants of chinese households' tourism consumption: Evidence from china family panel studies. *International Journal of Tourism Research*, 23(4):542–554.

Marcussen, C. H. (2011). Determinants of tourist spending in cross-sectional studies and at danish destinations. *Tourism Economics*, 17(4):833–855.

Marrocu, E., Paci, R., and Zara, A. (2015). Micro-economic determinants of tourist expenditure: A quantile regression approach. *Tourism Management*, 50:13–30.

Martinez-Garcia, E. and Raya, J. M. (2008). Length of stay for low-cost tourism. *Tourism Management*, 29(6):1064–1075.

Oklevik, O., Kwiatkowski, G., Malchrowicz-Mośko, E., Ossowska, L., and Janiszewska, D. (2021). Determinants of tourists' length of stay. *PloS One*, 16(12):1–17.

Park, S., Woo, M., and Nicolau, J. L. (2020). Determinant factors of tourist expenses. *Journal of Travel Research*, 59(2):267–280.

Peypoch, N., Randriamboarison, R., Rasoamananjara, F., and Solonandrasana, B. (2012). The length of stay of tourists in madagascar. *Tourism Management*, 33(5):1230–1235.

Prebensen, N. K., Altin, M., and Uysal, M. (2015). Length of stay: A case of northern norway. *Scandinavian Journal of Hospitality and Tourism*, 15(sup1):28–47.

Probst, P., Wright, M. N., and Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1301.

Raya, J. M. (2012). Length of stay for triathlon participants in the challenge maresme–barcelona: A survival approach. *Journal of Sport and Social Issues*, 36(1):89–105.

Rodriguez, X. A., Martinez-Roget, F., and Gonzalez-Murias, P. (2018). Length of stay: Evidence from santiago de compostela. *Annals of Tourism Research*, 68:9–19.

Soler, I. P., Gemar, G., and Correia, M. B. (2020). The climate index-length of stay nexus. *Journal of Sustainable Tourism*, 28(9):1272–1289.

Thrane, C. (2012). Analysing tourists' length of stay at destinations with survival models: A constructive critique based on a case study. *Tourism Management*, 33:126–132.

Thrane, C. (2015). Research note: The determinants of tourists' length of stay: Some further modelling issues. *Tourism Economics*, 21(5):1087–1093.

Thrane, C. and Farstad, E. (2012). Tourists' length of stay: The case of international summer visitors to norway. *Tourism Economics*, 18:1069–1082.

Wang, L., Fong, D. K. C., Law, R., and Fang, B. (2018). Length of stay: Its determinants and outcomes. *Journal of Travel Research*, 57(4):472–482.

WTO (2021). The economic contribution of tourism and the impact of covid-19, preliminary version.