The First International Conference on
Machine Learning and Knowledge Discovery (MLKD 2024)
Amirkabir University of Technology, December 18-19, 2024

Identifying factors affecting tourist destinations in Iran: Supervised Learning Approach

Mahdi Nezami*

Abstract

Nowadays, numerous features and variables influence tourists' destination choices. With the increase in tourist data and advancements in analytical tools, data analysis in tourism has grown. This article implements three supervised learning algorithms—Random Forest, Support Vector Machine, and Gradient Boosting—on tourist data from Iran for the periods 2015-2018 and 2019-2021. Additionally, it examines the correlation between the number of tourists and various features. The results indicate that history-based sites, such as museums, castles, and historical places, along with caravanserais, are the most effective and popular attractions influencing tourist destinations.

Keywords: Tourism, Data Analysis, Machine Learning, Supervised Learning

1 Introduction

Tourism is crucial in boosting economic growth by generating income and creating job opportunities across various sectors. It stimulates local economies, encouraging investment in infrastructure and services that benefit both visitors and residents. Additionally, tourism fosters cultural exchange and promotes regional development, which can lead to sustainable economic practices. The revenue generated from tourism taxes can be reinvested into community projects, enhancing the quality of life for locals. Overall, a robust tourism sector contributes significantly to a nation's economic stability and resilience [1].

Tourist destinations often showcase a unique blend of natural beauty, cultural heritage, and recreational opportunities that attract travelers from around the world. Popular attractions, such as historical landmarks, scenic landscapes, and vibrant local markets, play a pivotal role in shaping the travel experience. The appeal of these destinations can be enhanced by well-developed infrastructure, hospitality services, and engaging activities [2].

Data is essential in the tourism industry, offering insights that inform decision-making and strategic planning. It helps stakeholders comprehend traveler preferences, behaviors, and trends, facilitating targeted marketing and improved customer experiences. By analyzing data, businesses can identify peak travel seasons, allowing for optimized pricing and resource allocation. Additionally, data analysis aids in assessing tourism's impact on local communities and the environment, supporting effective destination management. Leveraging data enhances operational efficiency and helps organizations adapt to market changes, fostering innovation and the development of new products and services that align with consumer demands [3].

Data analysis holds significant potential to encourage tourism by identifying emerging trends and traveler preferences. By leveraging insights from visitor behavior and feedback, destinations can tailor their offerings to attract specific demographics. Additionally, predictive analytics can forecast demand, allowing businesses to optimize marketing strategies and promotional campaigns . Ultimately, harnessing data analysis fosters targeted initiatives that enhance visitor engagement and drive sustainable growth in the tourism sector [4]. Machine learning significantly enhances data analysis in the tourism industry by processing large datasets to identify patterns and insights. It improves predictive analytics, enabling businesses to forecast traveler behavior and preferences more accurately. This capability allows for personalized marketing strategies and optimized pricing models based on real-time data. Additionally, machine learning can analyze customer feedback and sentiment, offering valuable insights for service enhancement. Ultimately, it automates data-driven decision-making, driving growth and sustainability in tourism [5].

This article is structured as follows: Section 2 presents the literature review, while Section 3 describes the data. In Section 4, three supervised learning algorithms are analyzed using the dataset, and the results are discussed. Section 5 highlights the correlation between variables and the number of tourists. Finally, Section 6 concludes the article.

2 Litreature Review

The increasing integration of technology in tourism research has led to significant advancements in data analysis methodologies. Liu et al. provide a pioneering com-

*M.Sc. student Department of Industrial and Systems Engineering, Isfahan University of Technology, Isfahan 84156-83111, Iran, mahdinz377@gmail.com

prehensive literature review on big data applications in tourism, categorizing data into user-generated content, device data, and transaction data. Their systematic analysis highlights unique characteristics of each data type and addresses specific tourism issues. By examining research focuses, analytic techniques, and challenges, the study offers critical insights into this emerging field, making it a valuable resource for researchers navigating the complexities of big data in tourism [6].

Complementing this, Egger et al. present a comprehensive overview of Machine Learning (ML) and its relevance to the tourism sector. They address the recent surge of interest in ML while noting its limited application in tourism literature. By defining key terms, introducing major ML paradigms, and discussing relevant algorithms, the chapter demystifies the technical complexities of ML. It also highlights typical processes, limitations, and emerging trends like Auto-ML, providing foundational knowledge that supports further discussions on specific algorithms. This work is essential for integrating ML into tourism research and practice [5].

Additionally, Shapoval et al. highlight the application of decision trees in analyzing inbound tourist behavior in Japan, emphasizing a shift in visitor motivation from past experiences to future anticipations. Using a robust dataset of approximately 4,000 observations, their findings reveal that prospective experiences, such as visits to hot springs, significantly influence tourists' return intentions. By employing advanced data mining techniques, the study minimizes researcher bias and uncovers valuable visitor patterns, enhancing destination marketing strategies for governments and organizations. This contribution underscores the transformative potential of technology in understanding and shaping visitor experiences [7].

In examining the economic implications of tourism, one study analyzes causal relationships between tourism spending and economic growth in ten transition countries from 1988 to 2011. Utilizing panel causality analysis, it reveals diverse outcomes, supporting the neutrality hypothesis for Bulgaria, Romania, and Slovenia, while demonstrating growth effects in Cyprus, Latvia, and Slovakia. The findings indicate reverse relationships for the Czech Republic and Poland, with feedback effects observed in Estonia and Hungary. This study offers valuable policy implications, enhancing our understanding of tourism's economic impact in transitional contexts [8].

Further, Dolnicar critically examines the prevalent use of clustering for market segmentation in tourism research, tracing its evolution since Haley's seminal work in 1968. The study systematically reviews recent applications, highlighting fundamental weaknesses that may undermine the validity of segmentation results. By addressing common pitfalls and offering practical recommendations, Dolnicar provides a valuable guide for researchers and practitioners alike, emphasizing the need for cautious interpretation of segmentation findings to enhance their applicability in tourism management [9].

Bhatnagar et al. introduce an innovative framework for opinion mining in the tourism sector, tackling the challenges posed by the vast amount of user-generated reviews on e-commerce platforms. By distinguishing between explicit aspects through frequent nouns and extracting implicit aspects using a supervised machine learning technique (CRF), the study enhances the understanding of user sentiment. The empirical validation of the proposed algorithm demonstrates its effectiveness in identifying both explicit and implicit indicators in tourism-related data, contributing valuable methodologies for optimizing decision-making based on consumer feedback in the tourism industry [10].

In summary, these studies collectively illustrate the transformative impact of advanced data analysis techniques, including big data, machine learning, and opinion mining, on tourism research. They provide essential insights for understanding visitor behavior, economic implications, and effective marketing strategies, paving the way for future advancements in the field.

3 Data Discription

The dataset is found through Kaggle.com. It can be found easily at www.kaggle.com/datasets/saeedrostami1989/teravel-tourism-and-tourist-in-iran. The dataset pertains to Iranian provinces and their tourist attractions, which encompass various categories such as climate types, historical monuments, recreational sports, beaches, and UNESCO-registered sites. The details and explanations of these parameters are provided in the subsequent sections. The variables descrption could be found in Table1.

4 Supervised Learning

Supervised learning is a machine learning approach that trains a model on labeled data, where each training example is associated with a specific output or label. The algorithm learns to map inputs to outputs by minimizing the discrepancy between its predictions and the actual labels. It is commonly used in classification and regression tasks, aiming to predict categorical labels or continuous values, respectively. This article examines the most significant factors influencing tourist numbers in Iran between 2015-2018 and 2019-2021.

Table 1: Tourist Attraction Variables

Variable	Data Type	Description
province	Nominal	Name of the province
Climate	Categorical	Climate type
Location	Nominal	Geographic location
ecotourism	Numerical	The number of ecotourism residences
Caravanserai	Numerical	The number of active and visitable caravanserais
Museum	Numerical	The number of active and visitable museums
Historical mosque	Numerical	Number of historic mosques that can be visited
Church	Numerical	Number of historical and visitable churches
Castle	Numerical	Number of historical castles that can be visited
Lake	Numerical	Number of permanent and important lakes
waterfall	Numerical	Number of important and tourist waterfalls
River	Numerical	Number of permanent and accessible rivers for tourism
dam	Numerical	Number of dams available for tourism

4.1 Random Forrest

Random Forest is an ensemble of decision trees, and its feature importance metrics are based on how much each feature contributes to improving the model's accuracy across the trees. Gini Importance (Mean Decrease in Impurity): When building decision trees, Random Forest splits nodes on certain features. The Gini impurity (or other impurity measures like entropy) is reduced at each split. Feature Importance is computed by summing up the decrease in impurity (Gini impurity or entropy) that each feature brings when used in a split, across all trees in the forest. The more a feature decreases the impurity, the more important it is considered[11]. The formula for feature importance in Random Forest is: Gini Importance Formula in Random Forest

Gini Importance Formula

The Gini importance (also called Mean Decrease in Impurity) for a feature X_i in a Random Forest is calculated as follows:

$$\text{Importance}(X_i) = \sum_{\text{trees in forest}} \sum_{\substack{\text{nodes where} \\ X_i \text{ is used}}} \frac{N_{\text{node}}}{N_{\text{total}}} \Delta G_{\text{node}}$$

Where:

- N_{node} is the number of samples reaching the node where feature X_i is used for the split.
- N_{total} is the total number of samples in the dataset.
- ΔG_{node} is the decrease in Gini impurity at that node, which is given by:

$$\Delta G_{\text{node}} = G_{\text{before}} - G_{\text{after}}$$

- G_{before} is the Gini impurity before the split.
- G_{after} is the Gini impurity after the split.

Thus, the overall importance of feature X_i is the sum of the impurity decreases weighted by the proportion of samples passing through the node where X_i is used, across all trees in the forest.

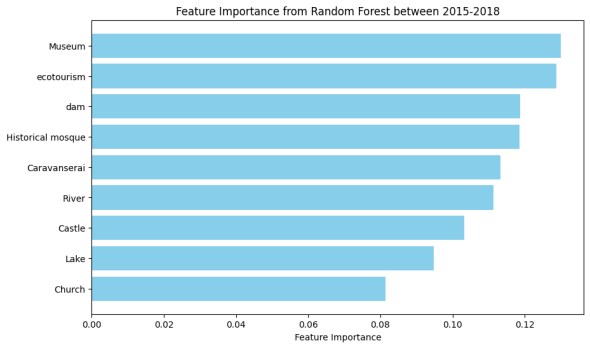

Figure 1: Feature Importance from Random Forest between 2015-2018

As shown in Figure1. and Figure 2, the most effective variable in attracting tourists to a province is the number of museums. After the museums, eco-tourism is the second variable that attracts tourists. Dams and historical places are in positions third and fourth. It

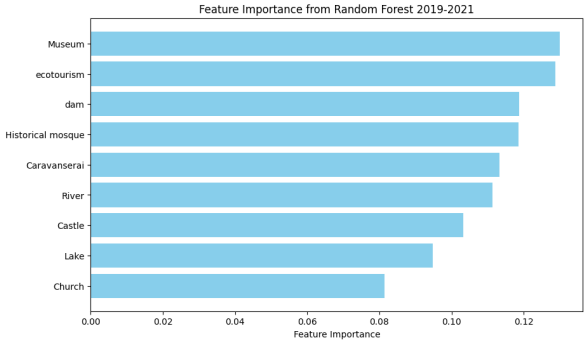

Figure 2: Feature Importance from Random Forest between 2019-2021

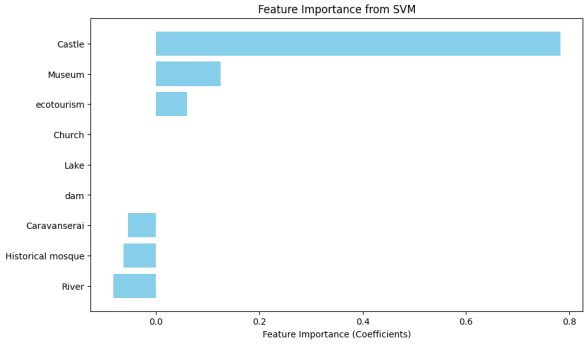

Figure 3: Feature Importance from SVM between 2015-2018

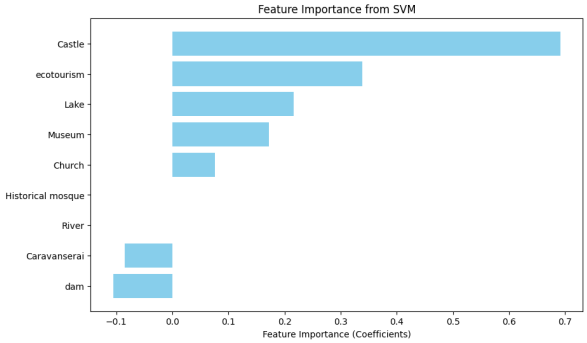

Figure 4: Feature Importance from SVM between 2019-2021

means that tourists pay attention to history (museums and historical places) alongside natural places with good special weather (ecotourism and dams).

4.2 Support Vector Machine

In Support Vector Machines (SVMs), especially when using a linear kernel, feature importance can be inferred from the weights of the model. These weights are part of the linear decision boundary that the SVM creates to separate classes. Here's how feature importance is calculated[12]: Formula for Feature Importance in Linear SVM: For a linear SVM, the decision function can be expressed as[12]:

$$f(x) = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b$$

Where:

- w_i are the weights assigned to each feature x_i,
- b is the bias term (or intercept),
- n is the number of features.

Feature Importance: The importance of each feature can be approximated by the absolute value of the weights w_i. Features with larger absolute weights are considered more important because they contribute more to the decision boundary.

The formula for importance is:

$$\text{Feature Importance} = |w_i|$$

Where:

- w_i is the weight associated with feature x_i,
- The larger $|w_i|$ is, the more important the feature x_i is in predicting the target.

As it shown, Figure3 and Figure4, SVM finds Castles the most effective variable for attracting tourists in 2015-2108 and 2019-2021. Ecotourism is the second most effective variable that attracted tourists in 2019-2021 and the third in 2015-2018.

4.3 Gradient Boosting

In Gradient Boosting, feature importance is calculated similarly to Random Forests, but it involves the boosting framework. Gradient Boosting builds trees sequentially, where each tree tries to correct the errors made by the previous ones, and feature importance is derived from how often and how effectively features are used to reduce the loss[13].

Gain-based Importance (Mean Decrease in Impurity): The most common way to compute feature importance in Gradient Boosting is by looking at how much a feature contributes to improving the model's performance across all the trees. Gain-based importance measures the average reduction in loss (or impurity) when a feature is used to split the data. It considers how much each feature contributes to minimizing the overall error during training[14].

The formula for importance is:

$$\mathrm{I}(X_i) = \frac{\sum_{\text{trees}} \sum_{\text{nodes where } X_i \text{ is used}} \text{Reduction in Loss}}{\text{Total number of samples passing}}$$

This is similar to Mean Decrease in Impurity (MDI) in Random Forests, but here the reduction in loss (instead of impurity) is the key metric.

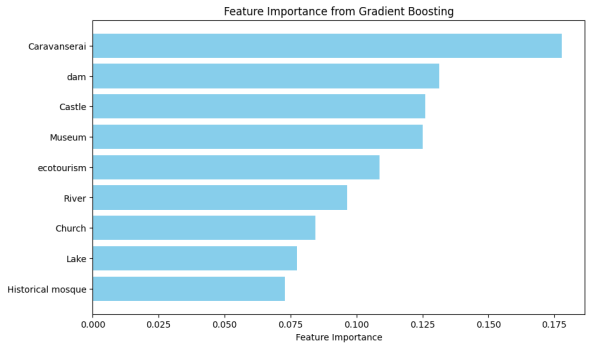

Figure 5: Feature Importance from Gradient Boosting 2015-2018

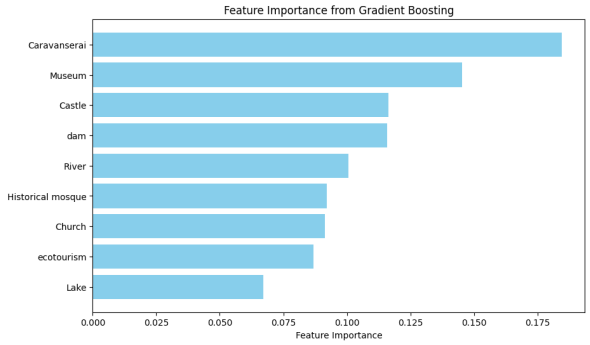

Figure 6: Feature Importance from Gradient Boosting 2019-2021

As is shown in Figure5 and Figure 6, the Carvanserai is the most effective variable in attracting tourists by Gradient Boosting in 2015-2018 and 2019-2021. Between 2015 and 2018, the Dam was the second most effective factor in attracting customers, while from 2019 to 2021, the Museum took on that role. The number of castles is the third key factor attracting tourists during both periods.

5 Correlation

Correlation measures the strength and direction of a linear relationship between two variables. It is represented by the correlation coefficient (r), which ranges from -1 to 1. A value of 1 indicates a perfect positive relationship, meaning as one variable increases, the other also increases. A value of -1 shows a perfect negative relationship, where one variable increases as the other decreases. A value of 0 indicates no linear relationship. Commonly used methods to compute correlation include Pearson's correlation for linear relationships and Spearman's rank correlation for non-linear but monotonic relationships[15]. As is shown in Figure 7 and

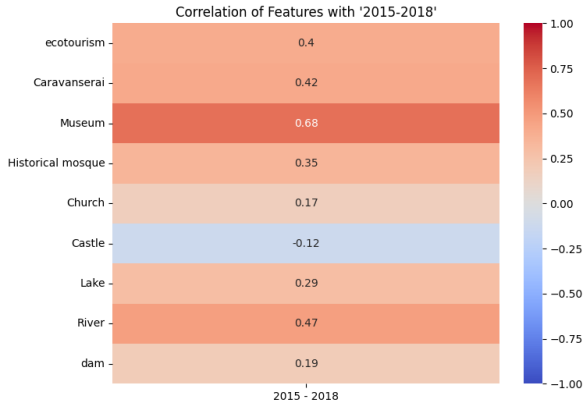

Figure 7: Correlation between Number of tourists in 2015-2018 and variables

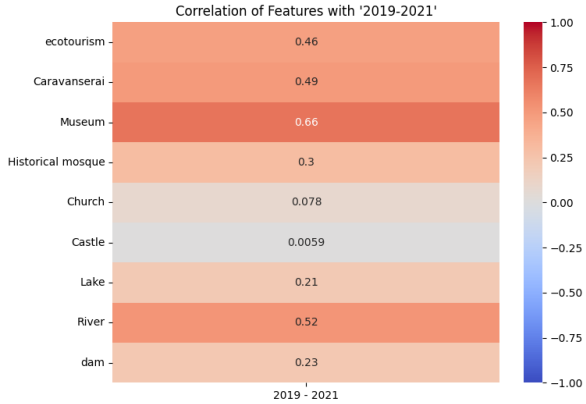

Figure 8: Correlation between Number of tourists in 2019-2021 and variables

Figure 8, the number of attracted tourists has the highest correlation to museums (number of museums). After the museums, the river has the highest correlation to the number of attracted tourists to each province. Carvaserai and ecotourism ranked third and fourth in correlation with tourist attraction numbers from 2015-2018 and 2019-2021.

6 Conclusion

Tourism significantly impacts the economy, influenced by various factors that affect tourist destinations. Businesses must understand these variables to thrive. The

wealth of data on customer travel patterns requires investments in data analysis. This article examines key factors that attract tourists, applying supervised algorithms like SVM and Random Forest to analyze the data, revealing distinct performance differences among them. In the final section, the correlation between these variables and the number of attracted tourists is presented, showing a strong relationship with the number of active or visitable museums in each province. The results indicate that tourists are very interested in historical sites (museums, castles, historical places) while also being attracted to locations with pleasant weather, such as dams. Caravanserais are another appealing feature for tourists and travelers.

The researcher could continue by using an unsupervised learning approach, like clustering, to group data based on effective features or provinces and analyze each cluster's characteristics.

7 Appendix

The data and the codes is available at the link below:
drive.google.com/file/d/1784Lbyt1hd3VMsSgCWGdlsicbFUZYZE9

References

[1] H. Song, L. Dwyer, G. Li, and Z. Cao. Tourism economics research: A review and assessment. *Annals of tourism research*, 39:1653–1682, 2012.

[2] D. Cronjé, E. du Plessis. A review on tourism destination competitiveness. *Journal of Hospitality and Tourism Management*, 45:256–265, 2020.

[3] C. Iorio, G. Pandolfo, A. D'Ambrosio, R. Siciliano. Mining big data in tourism. *Quality & Quantity*, 54:1655–1669, 2020.

[4] A. Khade. Performing customer behavior analysis using big data analytics. *Procedia Computer Science*, 79:986–992, 2016.

[5] R. Egger. Machine Learning in Tourism: A Brief Overview. *Applied Data Science in Tourism: Interdisciplinary Approaches*, 85–107, 2022.

[6] J. Li, L. Xu, L. Tang, S. Wang, and L. Li. Big data in tourism research: A literature review. *Tourism management*, 68:301–323, 2018.

[7] V. Shapoval, M.C. Wang, T. Hara, and H. Shioya. Data mining in tourism data analysis: inbound visitors to Japan. *Journal of Travel Research*, 57:310–323, 2018.

[8] M.C. Chou. Does tourism development promote economic growth in transition countries? A panel data analysis. *Economic Modelling*, 33:226–232, 2013. https://doi.org/10.1016/j.econmod.2013.04.024.

[9] S. Dolnicar. A Review of Data-Driven Market Segmentation in Tourism. *Journal of Travel & Tourism Marketing*, 12:1–22, 2002. 10.1300/J073v12n01_01.

[10] V. Bhatnagar, M. Goyal, and M.A. Hussain. A Proposed Framework for Improved Identification of Implicit Aspects in Tourism Domain Using Supervised Learning Technique. In *Proceedings of the International Conference on Advances in Information Communication Technology & Computing*, pages 1–4, 2016.

[11] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[12] S. Abe. Support vector machines for pattern classification. Springer, 2005.

[13] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54:1937–1967, 2021.

[14] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

[15] J. Lee Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42:59–66, 1988.