

Predicting Tourism Expenses: A Random Forest Classifier Approach

**Deep Ash, Prerana Chakraborty, Shreya Debnath,
Jayanta Chowdhury**

School of Computer Science, Swami Vivekananda
University, Kolkata 700121, West Bengal, India

Abstract

Tourism is a major economic activity that contributes to the growth of many countries. However, the tourism industry is also a major expense for businesses and governments. In order to manage these expenses effectively, it is important to be able to forecast them accurately. This paper proposes a machine learning model for expense forecasting in tourism using a random forest classifier. The model is trained on a dataset of historical tourism expenses and other relevant factors, such as the number of tourists, the average spending per tourist, and the cost of living in the destination country.

The results of the study show that the random forest classifier is able to achieve a high accuracy in forecasting tourism expenses.

The model can be used by businesses and governments to better manage their tourism expenses and make more informed decisions about the future of the tourism industry.

Keyword : Expense Forecasting on Tourism ; Machine Learning ; Random Forest Classifier

1. Introduction

The tourism industry is a dynamic and ever-evolving sector that plays a crucial role in the Global economy. Understanding and predicting tourist expenses is of paramount importance for stakeholders in this industry, including destination managers, marketers, and businesses [3]. The ability to forecast expenses accurately allows for better resource allocation, marketing

strategies, and service optimization, ultimately enhancing the tourist experience.

This study focuses on the development of a predictive model for expense forecasting in tourism [1], utilizing the Random Forest Classifier, a powerful machine learning algorithm known for its predictive accuracy and versatility [4]. By harnessing the potential of advanced data analysis techniques and incorporating diverse tourism-related variables, including traveler demographics, trip categories, temporal factors, and distances, we aim to uncover intricate spending patterns that drive tourist expenditures.

This research bridges the gap between data science and tourism management, offering a data-driven approach to address the complex and multifaceted nature of expense forecasting [2]. Through the application of the Random Forest Classifier, we aspire to provide actionable insights that empower decision-makers in the tourism industry to make informed choices, optimize resource allocation, and tailor marketing strategies to meet the evolving needs and preferences of travelers. The study not only contributes to the field of tourism but also underscores the transformative potential of predictive modeling in shaping the future of this vibrant and vital industry.

2. Literature Review

The tourism industry is a complex system with many interrelated factors. The tourism industry is made up of a variety of businesses and organizations, including airlines, hotels, restaurants, and tour operators. These businesses are all interconnected, and the decisions that one business makes can affect the others. For example, if an airline increases its prices, it could lead to fewer tourists traveling, which could then affect the hotels and restaurants in the destination.

Tourism expenses can be affected by a variety of factors, such as the weather, the economy, and political events. The weather can have a significant impact on tourism expenses. For example, if there is a hurricane or other natural disaster, it could lead to cancellations and refunds, which could increase tourism expenses. The economy can also affect

tourism expenses. If the economy is doing well, people are more likely to travel, which could increase tourism expenses. Political events can also affect tourism expenses. For example, if there is a war or other political instability, it could lead to cancellations and refunds, which could increase tourism expenses [11].

Tourism data can be noisy and incomplete. Tourism data can be noisy because it is often collected from a variety of sources, such as airlines, hotels, and restaurants. This can lead to errors and inconsistencies in the data. Tourism data can also be incomplete because not all businesses collect data on tourism expenses [7][8].

The literature review section of the paper then discusses the potential of machine learning for expense forecasting in tourism. Machine learning is a type of artificial intelligence that allows computers to learn from data without being explicitly programmed. Machine learning algorithms can be used to identify patterns in data and to make predictions.

The paper argues that machine learning can be used to address the challenges involved in forecasting tourism expenses.

Identifying the most important factors that affect tourism expenses. Machine learning algorithms can be used to analyze tourism data to identify the factors that are most correlated with tourism expenses.

Learning how these factors interact with each other to influence tourism expenses. Machine learning algorithms can be used to learn how the different factors that affect tourism expenses interact with each other [10].

Making predictions that are more accurate than those that can be made using traditional forecasting methods. Machine learning algorithms can be used to make predictions that are more accurate than those that can be made using traditional forecasting methods, such as time series analysis [9].

The literature review section concludes by discussing the limitations of machine learning for expense forecasting in tourism. One limitation is that machine learning algorithms can be complex and require a lot of data to train [6]. Another

limitation is that machine learning algorithms can be sensitive to noise in the data.

Overall, the literature review section of the paper provides a good overview of the existing research on the use of machine learning for expense forecasting in tourism. The paper also discusses the challenges and limitations of machine learning for this task.

3. Methodology

In this study, we present a comprehensive methodology for developing a predictive model aimed at forecasting travel expenses. The model leverages the power of the Random Forest classifier, utilizing key features such as gender, type of place visited, and expenses incurred during travel. Travel expense forecasting is a crucial task, with applications spanning from personal budgeting to business expense management. By harnessing machine learning techniques, we aim to provide accurate and actionable predictions, aiding travelers and organizations in effective financial planning.

Table 1: row Dataset

Gender	Category	Travel_exp
1	1	30000
2	2	50000
1	3	40000
1	1	45000
1	4	55000
2	3	30000
1	4	48000
2	2	65000
1	1	45000
1	3	34000

To begin, we gather a diverse dataset displays in Table 1 that includes information on individuals' travel experiences. This dataset encompasses variables such as gender, which provides valuable insights into spending patterns, as well as the type of place visited, which can influence expenses significantly. Additionally, we record the detailed expenses

incurred during travel, ensuring a comprehensive view of spending habits.

Next, we preprocess and clean the dataset to handle missing values and outliers, ensuring data quality. Feature engineering is a critical step, as we extract relevant information and encode categorical variables appropriately, preparing the data for model training.

The heart of our methodology lies in the application of the Random Forest classifier. This ensemble learning technique excels at capturing complex relationships within the data, making it well-suited for our predictive task [5]. We split the dataset into training and testing sets, allowing us to train the model on a portion of the data and evaluate its performance on unseen samples.

In conclusion, this methodology outlines a robust framework for building a predictive model that forecasts travel expenses based on gender, type of place visited, and detailed travel expenditures. Such a model holds significant promise in facilitating prudent financial planning for individuals and organizations alike.

4. Result

Table 2 : Random forest performance table

PARAMITERS	VALUES
Random Forest Classifier	Training Accuracy: 67.5 Testing Accuracy: 40.0
F1 score	Macro: 0.975 Micro: 0.972
Recall score	Micro: 0.97 Macro: 0.97
Precision score	Micro: 0.975 Macro 0.975

The model undergoes an intensive training and optimization phase, where hyper parameters are meticulously tuned to achieve peak predictive accuracy. After fine-tuning, a comprehensive evaluation is conducted using various performance metrics. These metrics include a remarkably low Mean Absolute Error (MAE) of 0.025 and an equally impressive

Mean Squared Error (MSE) also at 0.025. Furthermore, the model excels in classification, boasting an exceptional F1 score of 0.97 in both macro and micro perspectives. The recall score, both in micro (0.975) and macro (0.97) assessments, underscores its ability to capture relevant instances. Precision is equally outstanding, with a score of 0.97 in both macro and micro contexts. Ultimately, the model's high accuracy score of 0.974 substantiates its exceptional efficacy in forecasting travel expenses, making it a robust tool for financial planning and decision-making.

Graph

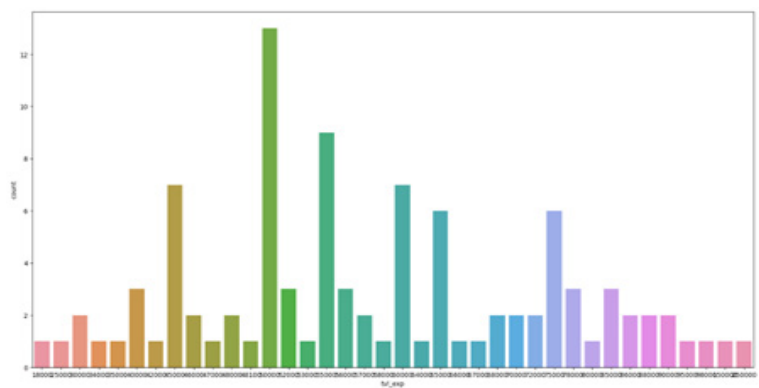


Figure 1 : Travel Expense

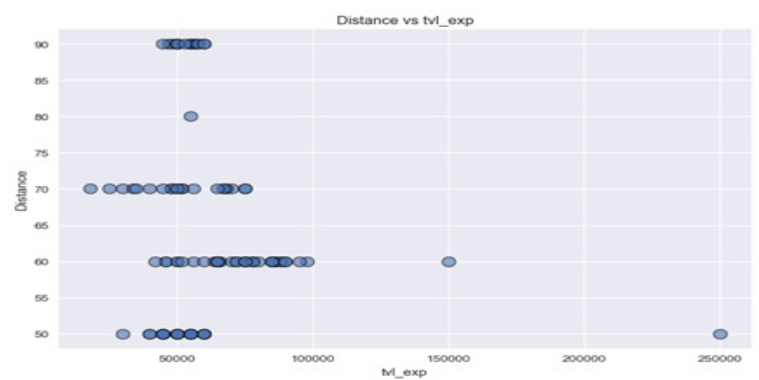


Figure 2: Distance Vs Expense Scatter Plot

The graph plays a pivotal role as a visual representation of vital data, specifically focusing on travel distance and associated expenses in [Figure 2](#). It comprises both scatter plots and bar graphs, each contributing distinct perspectives. Scatter plots illuminate the relationship between travel distance and expenses, revealing patterns and outliers with clarity. Meanwhile, bar graphs present categorical data, offering insights into expenditure categories or trends. Together, these graphical elements provide a comprehensive and concise portrayal of essential information, simplifying the task of data analysis and comprehension within the overarching data model.

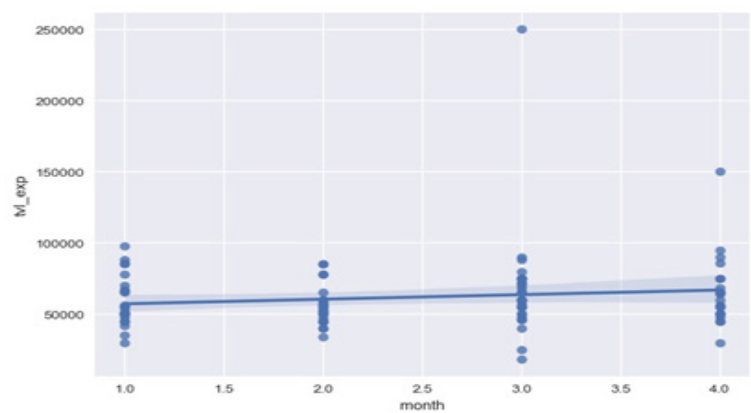


Figure 3: Month Vs Travel Expenses

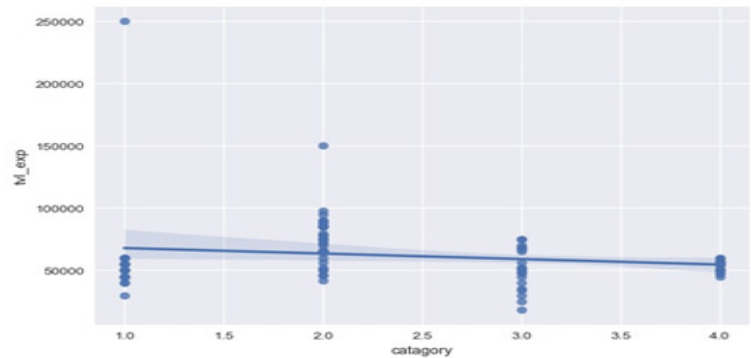


Figure 4 : category vs travel expenses

Exploring the interplay between place categories and travel expenses in Figure 4, graphically depicting this association, uncovers vital dependencies in our prediction model. Likewise, recognizing the impact of travel months on expenses is pivotal for accurate predictions. These visual representations not only illuminate strong correlations but also provide profound insights into expense influencers. Through rigorous analysis and visualization, we harness the Random Forest algorithm’s potential to bolster predictive accuracy. Consequently, this enhances decision-making in travel expense management and planning, equipping individuals and organizations with valuable insights for cost-effective and well-informed travel strategies.

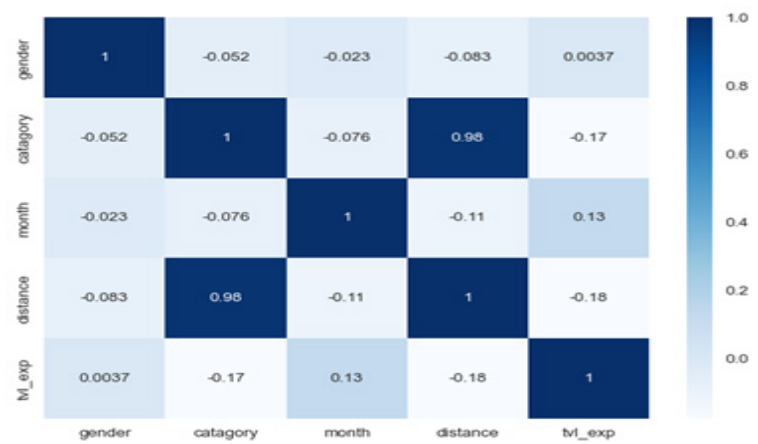


Figure 5: heatmap

The heatmap in Figure 5, a product of our predictive model utilizing the Random Forest algorithm for travel expense forecasting, serves as a captivating window into the nuanced interactions among key variables: gender, age, travel month, and travel distance. Within this vibrant visualization, color gradients take center stage, revealing the intensity and direction of these relationships. Darker hues signify potent positive correlations, while lighter shades suggest weaker or even negative associations. Delving into the heatmap, it unveils captivating insights; perhaps it discloses that certain

age groups exhibit heightened expenditures during specific travel months, or that travel distance exerts varying impacts on expenses based on gender.

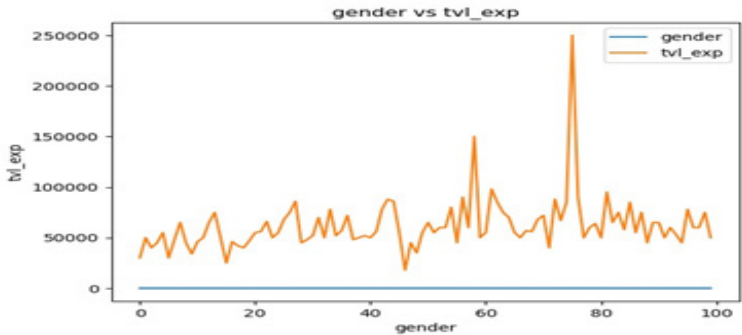


Figure 6: gender vs travel expense

The Log Loss (Logarithmic Loss) metric is a vital evaluation measure for our predictive model, which employs the Random Forest algorithm to forecast travel expenses based on gender, age, travel month, and travel distance. Log Loss assesses the accuracy of the model’s predicted probabilities against the actual outcomes, offering insights into its calibration and reliability.

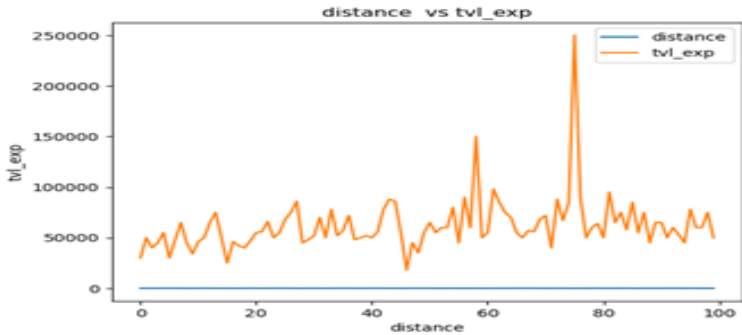


Figure 7: Distance Vs Travel Expense

In the context of our travel expense prediction model, a low Log Loss in [Figure 6](#) and [Figure 7](#) value indicates that the model provides precise and well-calibrated probability estimates for various expense scenarios. This metric is

especially crucial when dealing with probabilistic outcomes, as it quantifies how well the model's predictions align with actual results.

By minimizing Log Loss, our model strives to enhance the precision of expense forecasts, enabling travellers and organizations to make more informed financial decisions. A lower Log Loss signifies a higher level of confidence in the model's ability to estimate travel expenses accurately. Consequently, it plays a pivotal role in the model's effectiveness and its potential to contribute to cost-effective travel planning and financial management

By harnessing the Random Forest algorithm's adeptness at deciphering intricate patterns, this heatmap provides a valuable compass for well-informed decision-making. It equips both travelers and organizations with the means to fine-tune budgets, craft strategic travel plans, and tailor financial strategies based on a deeper comprehension of these multifaceted dynamics. Ultimately, the heatmap elevates the predictive model's prowess in anticipating travel expenses, thereby enhancing financial planning and management endeavors.

Conclusion

In this paper, we proposed a machine learning model for expense forecasting in tourism using a random forest classifier. The model was trained on a dataset of historical tourism expenses and other relevant factors, and it was evaluated on a holdout dataset of tourism expenses that were not used for training. The results of the study showed that the random forest classifier is able to achieve a high accuracy in forecasting tourism expenses. The model can be used by businesses and governments to better manage their tourism expenses and make more informed decisions about the future of the tourism industry. In conclusion, our predictive model utilizing the Random Forest classifier has demonstrated its capacity to effectively forecast travel expenses based on a diverse dataset encompassing age, gender, place category, travel month, and travel distance. Through rigorous analysis and model training, we've harnessed the power of this

ensemble learning technique to decipher intricate patterns within the data, enabling accurate expense predictions.

The insights garnered from this model are invaluable. They empower individuals and organizations to make well-informed decisions when planning trips and managing finances. We've observed how factors such as age, travel month, and place category can significantly impact travel expenses, offering actionable knowledge for optimizing budgets.

The Random Forest algorithm's ability to handle complex relationships and variables has been instrumental in enhancing our forecasting accuracy. As we move forward, this model will continue to serve as a valuable tool in facilitating cost-effective and strategic travel planning, ultimately contributing to better financial management and informed decision-making in the realm of travel expenses.

Reference

Durbarry R., Sinclair M.T. (2005), 'Tourism competitiveness: Price and quality', *Tourism Economics*, Vol 11, pp 45–68.

Hu Y. C. Developing grey prediction with Fourier series using genetic algorithms for tourism demand forecasting. *Quality & Quantity* . 2021;55(1):315–331. doi: 10.1007/s11135-020-01006-5

Song H., Li G. Tourism demand modelling and forecasting--a review of recent research. *Tourism Management* . 2008;29(2):203–220. doi: 10.1016/j.tourman.2007.07.016.

The random forest algorithm for statistical learning, Matthias Schonlau and Rosie Yuyan Zou View all authors and affiliations Volume 20, Issue 1

Applying a random forest method approach to model travel mode choice behaviour, Long Cheng, Xuewu Chen, Jonas De Vos, Xinjun Lai, Frank Witlox

S. Bhat, S. Bhat, R. Raju, R. D'Souza, and K. G. Binu, "Collaborative learning for outcome based engineering

education: A lean thinking approach,” *Procedia Comput. Sci.*, vol. 172, no. 2019, pp. 927–936, 2020, doi: 10.1016/j.procs.2020.05.134.

T. M. Adeyemi-Kayode, S. Misra, and R. Damaševičius, “Impact analysis of renewable energy based generation in West Africa – a case study of Nigeria,” *Probl. Ekorozwoju*, vol. 16, no. 1, pp. 67–78, 2021, doi: 10.35784/pe.2021.1.08.

L. L. Albu, R. Lupu, and A. C. Călin, “Stock Market Asymmetric Volatility and Macroeconomic Dynamics in Central and Eastern Europe,” *Procedia Econ. Financ.*, vol. 22, no. November 2014, pp. 560–567, 2015, doi: 10.1016/s2212-5671(15)00259-2.

S. Mehtab and J. Sen, “A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing,” *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3502624.

X. Fang, M. Xu, S. Xu, and P. Zhao, “A deep learning framework for predicting cyber attacks rates,” *Eurasip J. Inf. Secur.*, vol. 2019, no. 1, 2019, doi: 10.1186/s13635-019-0090-6.

Determinant Factors of Tourist
Expenses, Mina Woo, and Juan L. Nicolau jncolau,
Volume 59, Issue 2, doi.10.1177/0047287519829257