**Tourism Hospitality Industry Analysis Report**

Huirui （Evangeline） Song

Department of Information Systems & Operations Management, Suffolk University

ISOM 835: Predictive Analytics and Machine Learning

Professor: Hasan Arslan

April 29, 2025

**Table of Contents**

Introduction & Dataset Description ............................................................................................... 3

Exploratory Data Analysis (EDA) ................................................................................................. 4

Data Cleaning & Preprocessing ..................................................................................................... 6

Business Questions ........................................................................................................................ 8

Predictive Modeling ..................................................................................................................... 10

Insights ......................................................................................................................................... 12

Ethics & Interpretability .............................................................................................................. 15

Appendix: Code & Visuals .......................................................................................................... 18

**Introduction & Dataset Description**

The tourism and hospitality industry continues to be a critical contributor to economic development, with tourism revenue serving as a core indicator of success. Understanding the patterns and determinants of tourism revenue, customer retention, and visit purpose classification is essential for data-driven strategic planning in this sector. This report applies machine learning techniques to a curated tourism dataset comprising 500 records from diverse countries and cities, spanning from 2020 to 2024. The goal is to predict tourism revenue given travelers’ travel behavior and hotel characteristics, develop predictive models for customer return behavior, and classify visit purposes based on tourism and hospitality patterns.

The Kaggle dataset, [Tourism\_Hospitality\_Industry\_Analysis.csv](https://www.kaggle.com/datasets/smithmurphy/tourism-and-hospitality-industry-analysis-dataset), contains 23 features, both numerical and categorical. These include variables such as Number\_of\_Tourists, Average\_Length\_of\_Stay, Tourist\_Spending\_USD, Hotel\_Occupancy\_Rate, and Average\_Room\_Price\_USD, alongside categorical fields like Country, City, Purpose\_of\_Visit, and Hotel\_Rating. An additional target variable, Customer\_Return, was derived programmatically from the original column Tourist\_Satisfaction\_Score, with a binary assignment: 1 (likely to return) for scores ≥ 7, and 0 otherwise. This engineered feature enables binary classification for return behavior. The project implements regression, binary classification, and multi-class classification models, all trained using Python and scikit-learn libraries, to systematically address three central business questions:

1. Can we accurately predict tourism revenue based on travel behavior and hotel characteristics?
2. Can we predict customer return based on their experience?
3. Can we identify the purpose of visit based on traveler patterns?

**Exploratory Data Analysis (EDA)**

Exploratory analysis was conducted to understand the dataset's structure, feature distributions, and relationships, all of which directly inform our approach to modeling revenue, customer return, and purpose of visit classification.

We started by examining the distributions of numeric variables (Appendix Figure 1). Most variables such as Number\_of\_Tourists, Tourist\_Spending\_USD, and Eco\_Tourism\_Revenue\_USD display relatively even or moderately spread-out distributions, indicating a balanced dataset without extreme skews. The one exception is Tourism\_Revenue\_USD, which shows a clear right-skewed pattern—indicating a small number of instances with very high revenue outcomes. This supports the use of flexible models like Gradient Boosting Regressor that can capture such non-linearities effectively.

The dataset spans from 2020 to 2024 with reasonably balanced representation across years and months. The distribution of the Year feature confirms that the dataset is not overly concentrated in a single year, making it suitable for generalization across time periods. Similarly, the Month variable shows higher activity in January and December, which aligns with expected seasonal peaks during holiday and travel periods.

For categorical data (Appendix Figure 3), the dataset includes 20 countries, with Italy and China being the most frequently represented (33 records each), followed by Germany, South Africa, and Australia. The City variable includes 56 unique cities, with varied representation. Some cities appear multiple times, suggesting that data for those cities spans across different time periods, whereas others may be single-time entries. This variation is important to consider in modeling, as it introduces unequal sampling across locations.

The Purpose\_of\_Visit field reveals that Educational and Family Visit are the most common reasons for travel, each with 80 instances. This is followed by Medical (72), Religious (71), and Business (71), while Leisure is the least common purpose at 62 entries. This distribution contradicts common assumptions that leisure would dominate tourism data, but may reflect the dataset’s emphasis on structured travel activities such as education and family engagements.

Hotel ratings are also well distributed. Each rating level from 1-star to 5-star has substantial representation, with 5-star hotels being most common (91 entries). Notably, the ‘Unrated’ category is also sizable, with 82 entries, nearly equal to the 3-star and 4-star groups. This substantial count of unrated hotels suggests that some travelers may have stayed in informal accommodations or that data for these ratings was unavailable.

To explore variable relationships, we computed the correlation matrix (Appendix Figure 2). Tourism\_Revenue\_USD is strongly correlated with Tourist\_Spending\_USD (r = 0.78) and Number\_of\_Tourists (r = 0.56), confirming that both tourist count and individual spending are key revenue drivers. Meanwhile, the engineered variable Customer\_Return—based on whether satisfaction score ≥ 7—is highly correlated with its source, Tourist\_Satisfaction\_Score (r = 0.83). Other correlations are weak, suggesting that predictive modeling for return and visit purpose will likely require algorithms capable of capturing complex interactions, such as tree-based classifiers.

Overall, this EDA confirmed key trends and variable strengths relevant to our business questions. It also uncovered possible limitations, such as imbalanced city representation and the surprisingly low number of leisure travelers, which may influence model performance and interpretation.

**Data Cleaning & Preprocessing**

Before applying predictive models, the dataset was cleaned and prepared through a structured series of preprocessing steps using Python’s Pandas, NumPy, and Scikit-learn libraries. These steps ensured compatibility with machine learning models while preserving the integrity and interpretability of the data.

1. Missing Values and Duplicates

Initial inspection of the dataset revealed that there were no missing values and no duplicate rows. This allowed us to skip imputation and deduplication steps, providing a clean foundation for modeling. All entries in the dataset were assumed to be valid, complete records of tourist activity across multiple cities and countries.

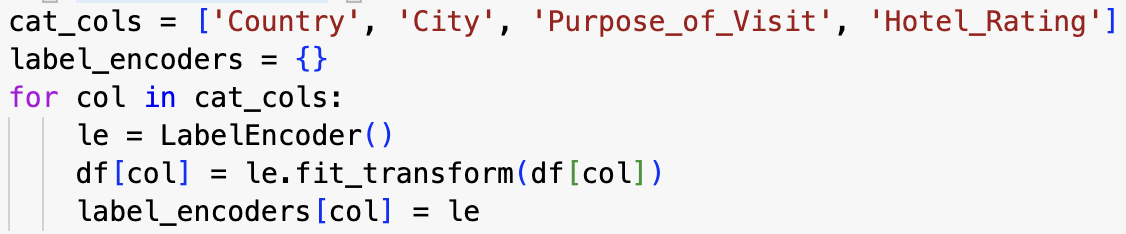
1. Label Engineering

To enable customer return prediction as a binary classification problem, we engineered a new variable called Customer\_Return. This variable was not part of the original dataset but was derived from Tourist\_Satisfaction\_Score. If the score was greater than or equal to 7, the customer was labeled as likely to return (1); otherwise, they were labeled as not likely to return (0). This transformation created a clear outcome variable for supervised learning related to customer loyalty, making it suitable for models like Naive Bayes, Decision Trees, and Gradient Boosting Classifiers.



1. Categorical Feature Encoding

Several important columns were categorical in nature: Country, City, Purpose\_of\_Visit, and Hotel\_Rating. These were label-encoded using Scikit-learn’s LabelEncoder(), which assigns a unique integer to each distinct category. This transformation preserves the distinctiveness of each class while converting them into a numeric format compatible with most machine learning algorithms. The encoders were also stored in a dictionary to reverse the transformation when interpreting prediction outputs, particularly in the multi-class classification task for purpose of visit.



1. Train-Test Splitting

For each of the three predictive tasks, we performed an 80/20 train-test split using train\_test\_split() with a fixed random\_state=42 for reproducibility. Each task required a tailored split depending on its target variable. For example:



This ensured that training and test data remained independent and that the test set could be used for final performance evaluation without leakage.

1. Target Selection Per task (Each business question led to a distinct target variable):
   1. For revenue prediction, the target was Tourism\_Revenue\_USD.
   2. For customer return prediction, the target was the engineered Customer\_Return column.
   3. For purpose of visit classification, the target was Purpose\_of\_Visit, using the label-encoded version for modeling and later decoding for interpretability.

In all cases, variables directly tied to the target (e.g., Tourist\_Satisfaction\_Score for return prediction) were excluded from the input features to prevent data leakage and artificial model inflation.

**Business Questions**

In the tourism and hospitality industry, business decision-makers are increasingly turning to data analytics to guide revenue optimization, customer retention, and service personalization. The following three business questions were formulated based on the structure of the dataset and the potential value that predictive insights could bring to industry stakeholders such as hotel managers, tourism boards, and city planners. Each question was selected with predictive modeling feasibility and strategic business relevance in mind.

1. Can we accurately predict tourism revenue based on travel behavior and hotel characteristics?

Rather than analyzing individual feature effects in isolation, this question focuses on building a high-performing predictive model of total tourism revenue (Tourism\_Revenue\_USD) based on travel patterns and contextual variables. These include Number\_of\_Tourists, Average\_Length\_of\_Stay, Tourist\_Spending\_USD, Hotel\_Occupancy\_Rate, and others. While Lasso regression helps estimate linear feature influence, the Gradient Boosting Regressor achieved the best overall performance (RMSE = 19,713.32, R² = 0.99) and was used as the final prediction model. This approach allows destination managers and tourism agencies to forecast future revenue with strong accuracy, supporting budget planning and performance benchmarking.

1. Can we predict customer return based on their experience?

Customer loyalty is critical in hospitality, where the cost of acquiring a new customer often exceeds that of retaining one. This question addresses whether it is possible to predict if a customer will return (binary variable Customer\_Return, engineered from Tourist\_Satisfaction\_Score) using features such as destination, spending, and satisfaction levels. Return prediction enables targeted retention campaigns, such as loyalty rewards or customized offers for at-risk groups. This task was modeled using Naive Bayes, Decision Trees, and Gradient Boosting, with moderate success (F1 scores around 0.66 for the best model).

1. Can we identify the purpose of visit based on traveler patterns?

Understanding the purpose behind each visit—whether Educational, Business, Family Visit, Medical, Religious, or Leisure—helps tailor destination services and marketing strategies. This classification task aimed to predict Purpose\_of\_Visit using customer spending, accommodation type, location, and timing. Although the model accuracy was relatively low across all algorithms tested (F1 scores around 0.11–0.12), the task remains conceptually valuable for visitor segmentation, especially with further feature refinement or expanded datasets.

**Predictive Modeling**

For each of the three business questions identified, I applied three predictive models and evaluated them using appropriate performance metrics. These are the models I used:

* Naive Bayes (as a baseline for classification tasks)
* Decision Tree (for its interpretability and ability to handle categorical data)
* Gradient Boosting (for superior predictive accuracy through ensemble learning)
* Lasso Regression (for linear interpretability in the regression task)

**1. Revenue Prediction (Regression Task)**

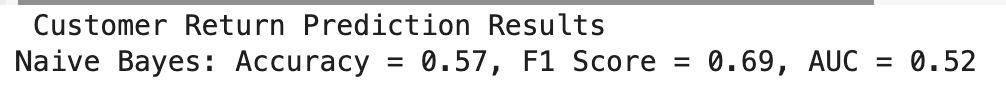
Target variable is Tourisom\_Revenue\_USD, I used models of Lasso Regression, Decision Tree, and Gradient Boosting with all features except target-related ones (Tourism\_Revenue\_USD, Customer\_Return, Tourist\_Satisfaction\_Score, and Hotel\_Rating\_Lable)

A math equations and formulas

Description automatically generated with medium confidence

As a result, Gradient Boosting Regressor outperformed the other models, providing near-perfect prediction accuracy on test data. This confirms that it effectively captures complex non-linear interactions between features such as Tourist\_Spending\_USD, Number\_of\_Tourists, and Hotel\_Occupancy\_Rate, which the correlation matrix in Appendix Figure 2 previously suggested were strong revenue predictors.

**2. Customer Return Prediction (Binary Classification)**

Target variable is Customer\_Return (engineered from Tourist\_Satisfaction\_Score). I used Models of Naïve Bayes, Decision Tree Classifier and Gradient Boosting Classifier by using all features except engineered or Target-related columns (Tourism\_Revenue\_USD, Customer\_Return, Tourist\_Satisfaction\_Score, and Hotel\_Rating\_Label) 

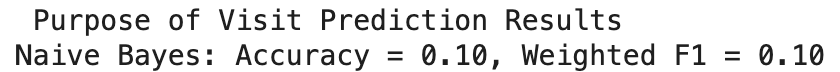




Naive Bayes had the highest F1 score and accuracy, though all models performed comparably. The lower Area Under the Curve (AUC) scores (all ~0.5) suggest limited ability to distinguish between returning and non-returning tourists. As shown in Appendix Figures 4–6, the confusion matrices reveal mild overprediction of the majority class. While predictive power was moderate, these models still provide a valuable starting point for targeted marketing strategies.

**3. Purpose of Visit Prediction (Multi-class Classification):**

Target variable is Purpose\_of\_Visit. I used models of Naïve Bayes, Random Forest, and Gradient Boosting by using all features except target and related columns (Purpose\_of\_Visit, Tourist\_Satisfaction\_Score, Customer\_Return, Tourism\_Revenue\_USD, Hotel\_Rating\_Label)







As shown in Appendix Figures 7–9, models struggled with this task due to overlapping feature distributions among visit purposes. Although accuracy was low, these results point to potential for improvement with more detailed features (e.g., traveler demographics, trip length, purpose-specific accommodations). The Random Forest classifier achieved the best balance across classes.

**Insights**

The predictive modeling efforts across the three business questions yielded varying levels of success, each with actionable implications and clear limitations. These insights help translate the technical findings into strategies that tourism and hospitality decision-makers can use to improve revenue forecasting, customer retention, and service segmentation.

**1. Predicting Tourism Revenue: Highly Accurate Forecasting with Gradient Boosting**

The Gradient Boosting Regressor achieved exceptional performance in predicting Tourism\_Revenue\_USD, with an RMSE of $19,713.32 and an R² of 0.99, suggesting that the model explains 99% of the variance in revenue outcomes. This high level of accuracy indicates that the dataset’s features—particularly Tourist\_Spending\_USD, Number\_of\_Tourists, and Hotel\_Occupancy\_Rate—are strongly predictive of revenue outcomes (Appendix Figure 2). These variables showed high positive correlations with revenue during EDA, which the Gradient Boosting model likely leveraged in its tree-based splits.

Business Implication: Tourism departments or city planners can use this model to build forecasting dashboards that predict city-level tourism revenue based on preliminary data inputs, such as visitor counts and accommodation statistics. This could help allocate budgets for infrastructure, staffing, or promotional campaigns more effectively.

Limitation: Despite the model’s accuracy, it offers limited interpretability. Stakeholders may need to complement it with simpler models (e.g., Lasso or Decision Trees) or SHAP (SHapley Additive exPlanations) values to understand specific feature contributions to revenue predictions.

**2. Predicting Customer Return: Moderate Performance, High Practical Value**

The second task explored the feasibility of predicting whether a tourist is likely to return, using the binary outcome variable Customer\_Return. Among the models tested, Naive Bayes achieved the best performance, with an F1 Score of 0.69, closely followed by Gradient Boosting at 0.66. Accuracy across models hovered around 0.55 to 0.57. While these results indicate moderate predictive power, the model still outperforms random guessing and could help identify groups with a high likelihood of repeat visits.

Business Implication: Hospitality firms can use this model to design retention campaigns targeted at tourists who show lower predicted likelihoods of return. For example, tourists from underperforming city-destination pairs could be offered personalized discounts, loyalty points, or service upgrades. Appendix Figures 4–6 show the confusion matrices, where the models were generally better at identifying non-returners than returners.

Limitation: The predictive model excludes the raw Tourist\_Satisfaction\_Score to avoid leakage, so it relies on secondary indicators of experience such as spending, location, and stay length. Without more detailed satisfaction-related features (e.g., service ratings, activity engagement), the model’s scope remains limited.

**3. Predicting Purpose of Visit: A Challenging Classification Problem**

The third modeling task aimed to classify the Purpose\_of\_Visit into one of six categories (e.g., Educational, Business, Family Visit). This task posed a significant challenge due to overlapping feature distributions and limited class-specific signals in the input data. The best performing model, Random Forest, achieved an accuracy of 0.13 and a Weighted F1 Score of 0.12, only marginally better than chance. Confusion matrices in Appendix Figures 7–9 highlight the widespread misclassifications, especially among mid-sized categories like Medical and Religious travel.

Business Implication: While the model itself is not production-ready, the exercise reveals gaps in data collection that, if addressed, could unlock valuable insights. For example, adding variables such as age group, travel group size, or duration of stay might provide stronger signals for distinguishing leisure from educational trips.

Limitation: The model was built using only quantitative and categorical data, with no behavioral or demographic input. This task likely requires a richer feature set or even natural language processing (e.g., on open-ended feedback) to meaningfully distinguish visit purposes.

**Ethics & Interpretability**

As machine learning becomes more embedded in decision-making within tourism and hospitality, ethical considerations must be addressed to ensure fairness, transparency, and responsible use of predictive technologies. While this project focuses on technical performance and accuracy, several ethical and interpretability concerns arise from both the modeling process and its potential deployment.

**1. Bias and Fairness**

The models developed in this project rely heavily on quantitative variables such as Tourist\_Spending\_USD, Hotel\_Occupancy\_Rate, and city/country encodings. This dependence raises concerns about unintentional bias. For instance, if certain cities or countries are underrepresented in the dataset, the models may learn to associate low return likelihood or lower revenue potential with those areas, not because of actual customer behavior but due to sampling limitations. Moreover, since the dataset does not include demographic data such as age, gender, or income, the potential for indirect bias still exists, especially when geographic features act as proxies for socio-economic background.

Similarly, the binary variable Customer\_Return was engineered based on an arbitrary satisfaction threshold (score ≥ 7), which—while intuitive—may not capture the full complexity of return behavior. This simplification, though necessary for modeling, may overlook subtler satisfaction patterns or cultural differences in rating behavior.

To mitigate these risks, future versions of the model should:

* Incorporate more representative data samples across locations and traveler types
* Use fairness metrics (e.g., disparate impact ratio) during evaluation
* Allow for periodic auditing of model predictions and retraining with updated data

**2. Model Interpretability and Stakeholder Transparency**

While Gradient Boosting was chosen for its high predictive performance, it is inherently less interpretable than linear models like Lasso. This creates challenges when communicating results to non-technical stakeholders such as tourism managers or government planners who must trust and act on these predictions. Black-box models may generate resistance or misapplication, especially in contexts involving resource allocation or customer targeting.

To improve interpretability:

* Future iterations should consider using tools like SHAP values or feature importance plots to explain individual predictions.
* Where feasible, simpler models (e.g., Decision Trees) could be deployed in tandem with complex models to provide explainable summaries.
* Documentation should clearly explain how features are engineered and how prediction outputs should (and should not) be used.

In conclusion, while the models developed in this project show strong technical promise, they should not be used without ethical oversight and ongoing evaluation. Predictive systems in tourism must balance accuracy with fairness and transparency to serve both business goals and public trust.

**Appendix**

**A. Code Snippets**

1. Customer Return Label Creation



2. Label Encoding for Categorical Variables

A white background with black text

Description automatically generated

3. Train-Test Splitting for Regression



4. Gradient Boosting Regression Fit and Evaluation

A screenshot of a computer code

Description automatically generated

5. Confusion Matrix Plotting Function

A screen shot of a computer code

Description automatically generated

**B. Visualizations**

This Figures listed below were generated in Colab and referenced throughout the report

EAD Visuals

Figure 1

A graph of different types of numbers

Description automatically generated with medium confidence

Figure 2

A graph with red and blue squares

Description automatically generated

Figure 3

A group of blue bars

Description automatically generated

Customer Return Classification

Figure 4:

A chart with blue squares

Description automatically generated

Figure 5:

A chart with blue squares

Description automatically generated

Figure 6

A chart with blue squares

Description automatically generated

Purpose of Visit Classification

Figure 7

A blue squares with white text

Description automatically generated

Figure 8

A blue squares with white text

Description automatically generated

Figure 9

A graph of a business results

Description automatically generated with medium confidence