Tourism Experience Analytics: Classification, Prediction, and Recommendation System

Domain: Travel & Tourism Analytics | Machine Learning Deployment

1. Executive Summary

Tourism Experience Analytics is an end-to-end machine learning system developed to transform structured tourism data into predictive intelligence and personalized recommendations. The system integrates regression modeling, classification modeling, and a recommendation engine within a production-grade Streamlit application.

Using a fully merged master dataset of 52,898 records and 21 features, the solution predicts attraction ratings, classifies visit modes, and generates personalized attraction suggestions to improve business decision-making.

2. Business Problem

Travel platforms generate large volumes of user and attraction data but often fail to convert it into predictive insights.

Key challenges include predicting satisfaction before reviews, understanding visitor intent, personalizing recommendations, and managing seasonal demand fluctuations.

This project addresses these challenges using structured data engineering and supervised machine learning.

3. Data Architecture and Integration

The project is built on nine relational datasets covering transactions, users, attractions, visit modes, and hierarchical geographic data.

The Transaction dataset contains 52,930 records from 2013 to 2022, capturing user visits and ratings. User data includes 33,530 users linked across 6 continents, 22 regions, 165 countries, and 9,143 cities.

After cleaning, validation, and merging all tables using proper keys, the final master dataset contains 52,898 rows and 21 columns.

Duplicate rows, inconsistent types, redundant columns, and invalid IDs were removed. Mixed categorical values in AttractionTypeId were standardized to numeric IDs to ensure accurate integration.

The dataset is structured, complete, and suitable for regression, classification, and recommendation modeling.

4. Data Characteristics

1. Transaction Data:

- **Purpose:** Contains information on user visits to various attractions, including ratings and visit details.
- **Columns:**
  - TransactionId: Unique identifier for each transaction.
  - UserId: Identifier for the user making the transaction.
  - VisitYear: The year the visit occurred.
  - VisitMonth: The month the visit occurred.
  - VisitMode: The mode of visit (Business, Couples, Family, etc.).
  - AttractionId: Unique identifier for the visited attraction.
  - Rating: The user's rating for the attraction.
- **Usage:** This data will be used to understand user behavior, predict visit modes, and recommend attractions based on user preferences and ratings.

2. User Data:

- **Purpose:** Contains information about the users, such as their geographical location (continent, region, country, city).
- **Columns:**
  - UserId: Unique identifier for each user.
  - ContinentId: The continent to which the user belongs.
  - RegionId: The region within the continent.
  - CountryId: The country in which the user resides.
  - CityId: The city of residence for the user.
- **Usage:** This data will help in analyzing the user demographics, which can be used to predict their behavior, visit modes, and attractions they might prefer.

3. City Data:

- **Purpose:** Contains information about different cities, used to link users and attractions to specific geographical locations.
- **Columns:**
  - CityId: Unique identifier for each city.
  - CityName: Name of the city.
  - CountryId: The country associated with the city.
- **Usage:** This data helps to relate the CityId in the user and attraction data to a specific city name and country.

4. Type:

- **Purpose:** Contains details about type of tourist attractions.
- **Columns:**
  - AttractionTypeId: The type of the attraction (e.g., Beach, Museum, Park).
  - AttractionType: Type of the attraction.

5. Visit Mode Data:

- **Purpose:** Contains information about the types of visit modes users may have, such as business trips, family vacations, or solo travel.
- **Columns:**

    - VisitModeId: Unique identifier for the visit mode.
    - VisitMode: Name or description of the visit mode (e.g., Business, Couples, Family).
- **Usage:** This data will be used to predict the user's likely visit mode based on their transaction history and demographic information.

6. Continent Data:

- **Purpose:** Contains information about the continents, helping to link users to their respective continents.
- **Columns:**
    - ContinentId: Unique identifier for each continent.
    - Continent: Name of the continent (e.g., Africa, Asia, Europe).
- **Usage:** Helps associate users with their continent, which may influence their travel behavior and preferences.

7. Country Data:

- **Purpose:** Contains information about countries, linking users and attractions to specific countries.
- **Columns:**
    - CountryId: Unique identifier for each country.
    - Country: Name of the country.
    - RegionId: The region within the country.
- **Usage:** Helps in understanding user preferences and travel behavior in relation to different countries.

8. Region Data:

- **Purpose:** Contains information about regions within countries.
- **Columns:**
    - RegionId: Unique identifier for each region.
    - Region: Name or description of the region (e.g., North America, East Africa).
    - ContinentId: Unique identifier for each continent.
- **Usage:** This data helps classify users and attractions based on regional preferences and trends.

9. Item Data:

- **Purpose:** Contains information about regions within countries.
- **Columns:**
    - AttractionId**:** Unique identifier for each attraction, linking it to other datasets.
    - AttractionCityId**:** City where the attraction is located, connecting it to the City Data.
    - AttractionTypeId**:** Category of the attraction (e.g., beach, historical site, park) for personalized recommendations.
    - Attraction**:** Name of the attraction used to identify and suggest places.
    - AttractionAddress**:** Physical address, useful for mapping, distance calculations, and travel planning.
- **Usage:** This data will be used to recommend attractions based on user preferences, location, and visit modes.

5. Exploratory Data Analysis

Rating Distribution

Most ratings fall between 4 and 5, confirming strong customer satisfaction. Low ratings are rare but represent potential risk areas requiring monitoring.

Visit Mode Distribution

Couples and family trips dominate tourism activity. Business and solo travel segments are comparatively smaller, indicating untapped growth opportunities.

Visits by Year

Tourism increased from 2013 to 2016, declined afterward, and experienced a sharp drop in 2020 due to global disruptions. Recovery begins post-2021.

Visits by Month

Mid-year months, especially July and August, show peak travel activity. Early and late months experience lower traffic, indicating strong seasonality.

Geographic Distribution

America and Asia generate the highest number of users and visits. Africa shows the lowest participation, representing a potential expansion opportunity.

Attraction Type Distribution

Nature, wildlife, beaches, and religious sites are most visited. Niche categories such as specialty museums and spas receive lower traffic.

Top Attractions

Sacred Monkey Forest Sanctuary and Waterbom Bali are among the most visited attractions, showing concentration in specific high-demand locations.

6. Bivariate and Multivariate Insights

Rating by Visit Mode

Business and family travelers give slightly higher ratings. Solo travelers show marginally lower satisfaction, suggesting room for service enhancement.

Rating by Attraction Type

Water parks and spas receive the highest average ratings. Historic sites and beaches show slightly lower satisfaction but remain above 4 overall.

Rating by Year

Ratings remained stable early on, dipped slightly in 2017–2018, then improved, peaking around 2021 before a minor decline in 2022.

Visit Mode by Month

Families and couples dominate during peak months. Business travel remains consistent but low throughout the year.

Popularity vs Rating

Highly popular attractions tend to receive higher ratings, confirming a relationship between demand and satisfaction.

Correlation Analysis

Most numerical features show weak linear correlation with rating. This confirms that satisfaction is influenced by complex, non-linear behavioral factors.

7. Regression Modeling – Rating Prediction

The objective was to predict attraction ratings using supervised regression.

Models evaluated:

- Linear Regression
- Random Forest (Default & Tuned)
- Gradient Boosting (Default & Tuned)

Final Results:

Model	R^2 Score
Linear Regression	0.7328
Random Forest (Tuned)	0.7420
Gradient Boosting (Default)	0.7450
Gradient Boosting (Tuned)	**0.7453**

Final Selected Model: **Gradient Boosting (Tuned)**, $R^2 = 0.7453309385744564$

The model explains approximately 74.53% of rating variance, demonstrating strong predictive capability.

8. Classification Modeling – Visit Mode Prediction

The objective was to classify visitor type using supervised classification.

Models evaluated:

- Logistic Regression
- Random Forest
- Gradient Boosting
- XGBoost
- LightGBM

Final Results:

Model	Accuracy
Logistic Regression	43.59%
Random Forest	50.98%
Gradient Boosting	49.86%
LightGBM	50.40%
XGBoost (Tuned)	**51.93%**

Final Selected Model: **XGBoost (Tuned)**, Accuracy = 0.51923%.

Tree-based ensemble methods outperformed linear models, confirming non-linearity in visitor behavior patterns.

9. Recommendation Engine

The recommendation system uses behavior-based filtering.

It identifies preferred attraction types and ranks similar attractions using popularity scoring.

The engine supports dynamic Top-N selection and safe fallback mechanisms if user history is unavailable.

This enables personalized tourism experiences and improved engagement.

10. Automated Model Selection Framework

The application uses JSON configuration files to automatically load the best-performing models.

Regression Configuration:

- Model: Gradient Boosting (Tuned)
- Metric: R^2
- Score: 0.7453

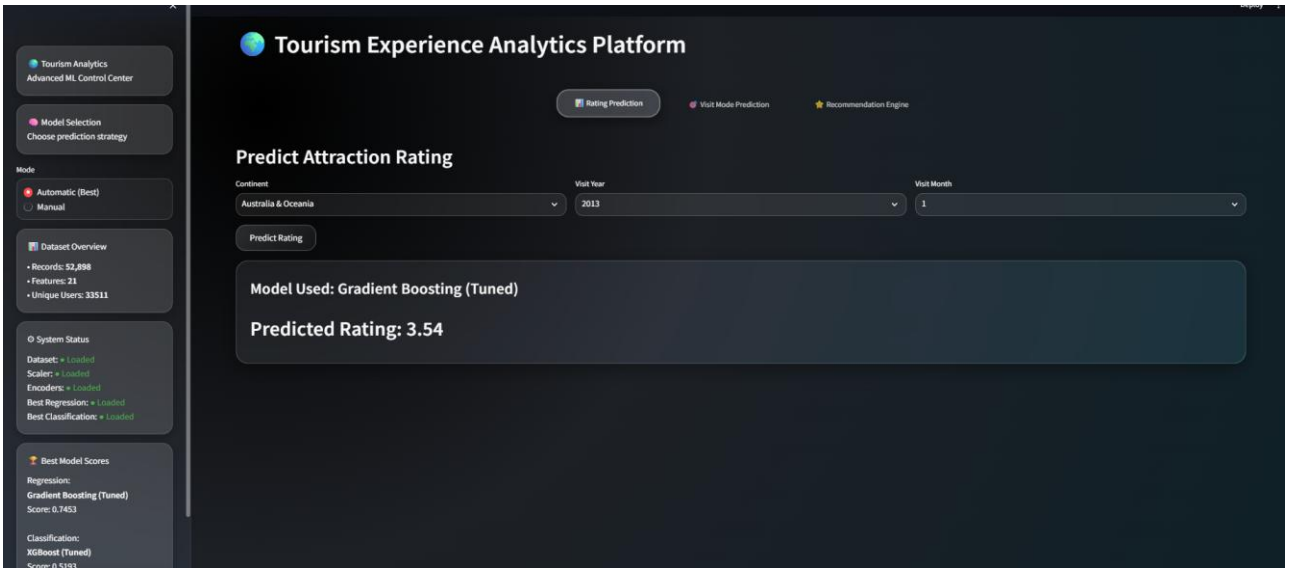
Classification Configuration:

- Model: XGBoost (Tuned)
- Metric: Accuracy
- Score: 0.5193

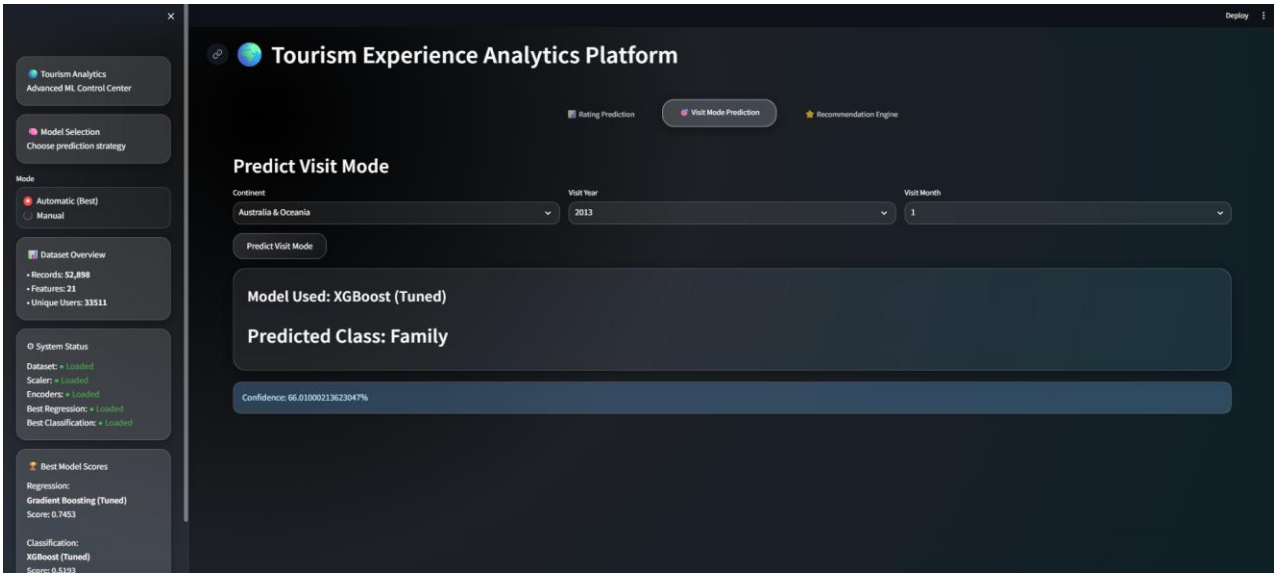
This ensures reproducibility, flexibility, and stable deployment.

11. Output

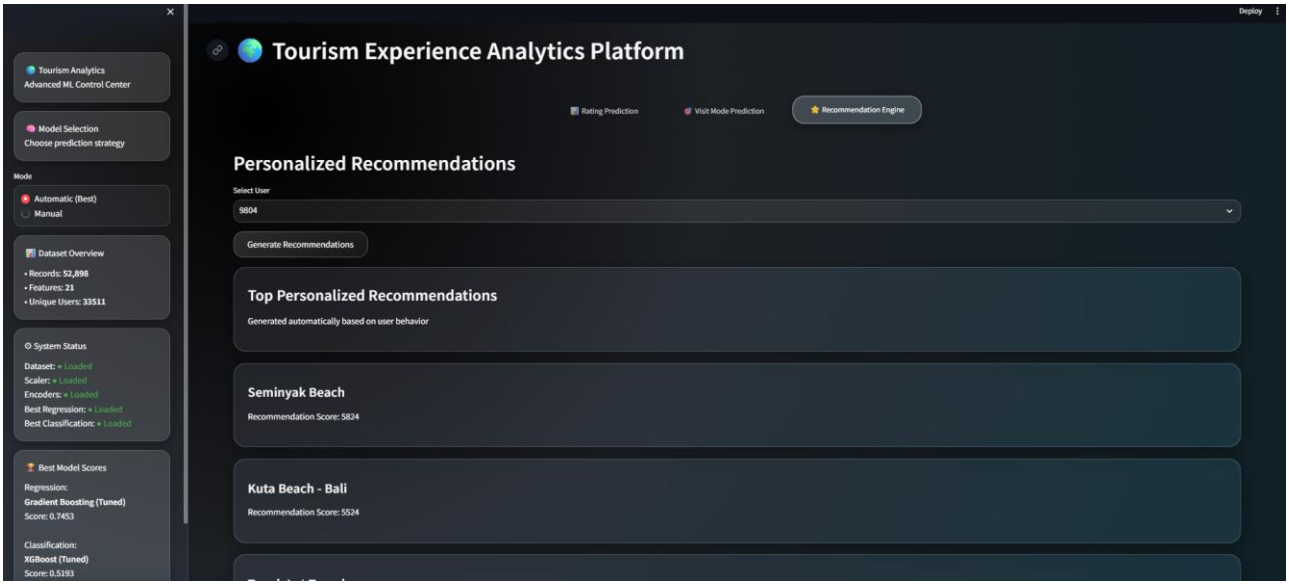
Rating predication:



Visit Mode Predication:



Recommendation Engine:



12. Deployment Architecture

The system is deployed through a production-grade Streamlit dashboard featuring:

- Automatic best model selection
- Manual override capability
- Model health monitoring
- Confidence score tracking
- Safe fallback architecture

The system prevents crashes even if models or scalers are missing, ensuring operational reliability.

13. Key Business Insights

1. Customer satisfaction is generally high, supporting strong brand reputation.
2. Tourism demand is geographically concentrated in America and Asia.
3. Family and couples segments drive core revenue.
4. Solo and African markets represent growth opportunities.
5. Tourism is highly seasonal, requiring dynamic pricing strategies.
6. Popular attractions reinforce satisfaction but risk overcrowding.
7. Satisfaction is influenced by complex behavioral factors, justifying advanced ML models.

14. Strategic Recommendations

- Expand marketing in underrepresented regions.
- Develop tailored packages for solo travelers.

- Promote secondary attractions to reduce overcrowding.
- Implement predictive monitoring for low-rated attractions.
- Optimize seasonal pricing and promotions.
- Continue using ensemble models for future improvements.

15. Conclusion

Tourism Experience Analytics successfully integrates large-scale data engineering, predictive modeling, classification intelligence, and personalized recommendation systems into a production-ready solution.

With 52,898 integrated records, strong model performance, and a stable deployment framework, the system demonstrates how tourism behavioral data can be converted into actionable business intelligence.

The project represents a complete real-world machine learning lifecycle — from data integration to deployment — with measurable strategic value for travel platforms.