Milestone 1: International Hotel Booking Analytics

Predicting Hotel Country Groups using Machine Learning

October 24, 2025

Team Members

Nadine Tarek Abdellatif Abdelhalim Nour
P016 55-5906
nadine.nour@student.guc.edu.eg

Mohamed Hazem Moustafa Elquesni
P025 55-4129
mohamed.elquesni@student.guc.edu.eg

Nour Ahmed Essameldin Mohamed Helmy Abdullah
P012 55-12742
nour.abdullah@student.guc.edu.eg

Mohamed Ahmed Hamed Aly Ahmed Sweidan
P015 55-5201
mohamed.sweidan@student.guc.edu.eg

Contents

1 Introduction

We predict which country group a hotel belongs to based on user reviews and demographics. The dataset contains 50,000 reviews from 25 hotels across 11 geographic regions (country groups). This classification task helps understand regional patterns in hotel characteristics and user satisfaction.

The dataset consists of three files: hotels (25 hotels with baseline quality metrics), reviews (50,000 reviews with scores across 6 dimensions), and users (2,000 users with demographic information). By analyzing review patterns and user demographics, we can identify distinctive characteristics of hotels in different regions.

This problem matters because hotels and booking platforms can use regional patterns to improve recommendation systems, optimize marketing strategies, and understand customer expectations across different geographic markets.

2 Data Cleaning

We performed initial data quality assessment and cleaning on all three datasets before merging.

2.1 Data Quality Assessment

Table 1: Null Values Check

Dataset	Total Rows	Columns	Null Values Found
Hotels	25	13	0
Reviews	50,000	14	0
Users	2,000	7	0

No missing values detected in any dataset.

Table 2: Duplicate Records Check

Dataset	Total Rows	Duplicates Found
Hotels	25	0
Reviews	50,000	0
Users	2,000	0

No duplicate records detected.

2.2 Column Renaming

To prevent naming conflicts after merging, we applied prefixes to columns:

- Hotels: `hotel_*` (except hotel_id, hotel_name)
- Reviews: `review_*` (except review_id, user_id, hotel_id, review_date, review_text)

- Users: `user_*` (except user_id, user_gender)

Example: `cleanliness` → `hotel_cleanliness_base`, `score_location` → `review_score_location`

2.3 Dataset Merging

Merged datasets using left joins:

1. reviews ← users (on user_id)
2. result ← hotels (on hotel_id)

Final merged dataset: 50,000 rows × 33 columns

2.4 Unnecessary Column Removal

Dropped 4 columns that don't contribute to country group prediction:

Table 3: Dropped Columns

Column	Reason for Removal
review_date	Temporal patterns not relevant to hotel location
review_text	Requires NLP analysis (out of scope)
user_join_date	User registration date doesn't indicate hotel location
hotel_name	Unique identifier causing data leakage (25 unique names)

2.5 Cleaned Dataset Summary

Final dataset ready for analysis:

- Shape: 50,000 rows × 29 columns
- No null values
- No duplicates
- All necessary information retained for modeling

3 Data Engineering Questions

3.1 Question 1: Best City for Each Traveler Type

We analyzed which city performs best for each traveler type based on average review scores.

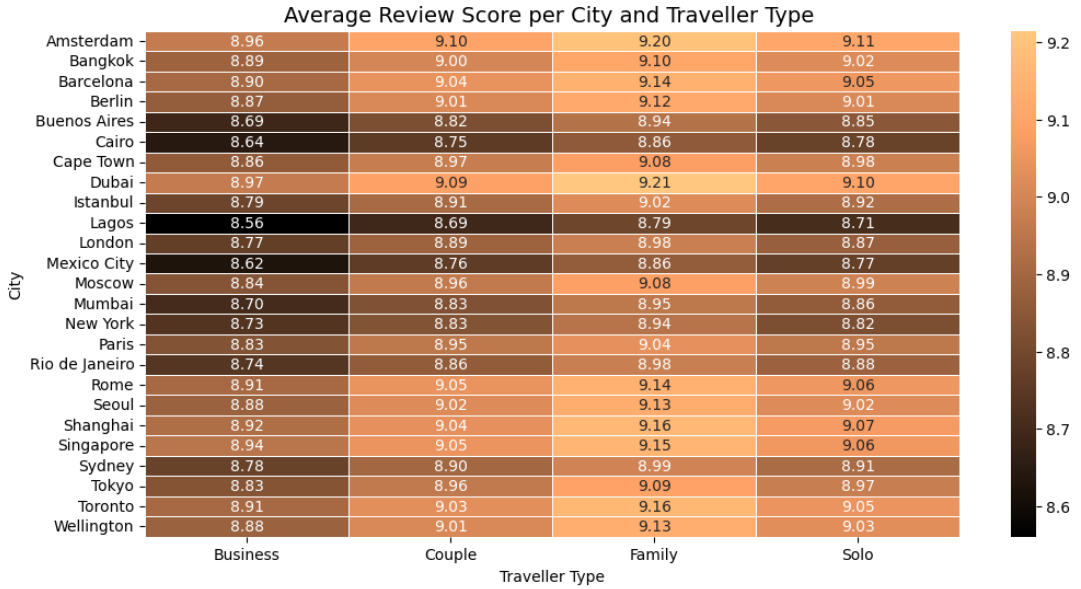

Figure 1: Average Review Score per City and Traveler Type

Results:

- Business travelers: Dubai (8.97)
- Couples: Amsterdam (9.10)
- Families: Dubai (9.21)
- Solo travelers: Amsterdam (9.11)

Dubai excels for business and family travelers, while Amsterdam is preferred by couples and solo travelers.

3.2 Question 2: Top 3 Countries by Value-for-Money per Age Group

We identified countries with the highest value-for-money scores across different age demographics.

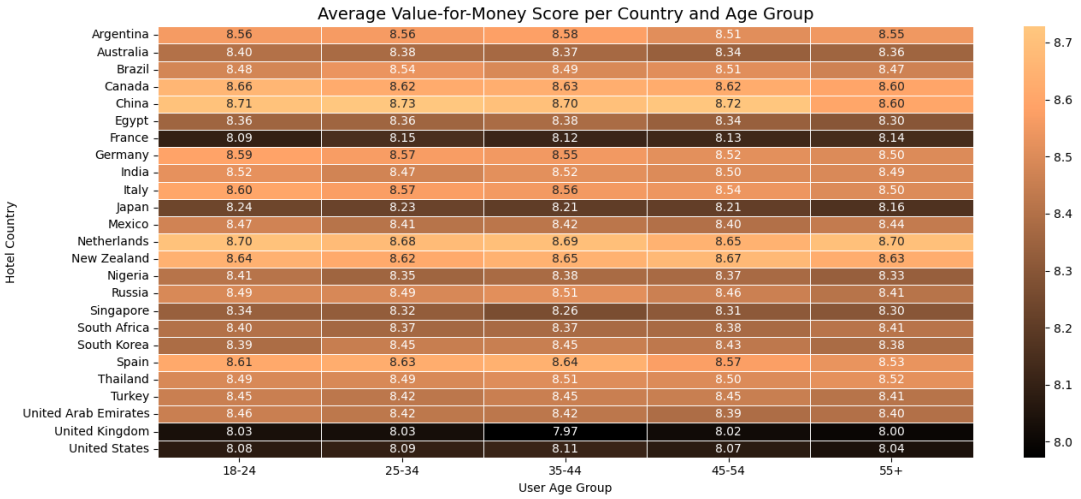

Figure 2: Average Value-for-Money Score per Country and Age Group

Table 4: Top 3 Countries by Value-for-Money per Age Group

Age Group	Rank 1	Rank 2	Rank 3
18–24	China (8.71)	Netherlands (8.70)	Canada (8.66)
25–34	China (8.73)	Netherlands (8.68)	Spain (8.63)
35–44	China (8.70)	Netherlands (8.69)	New Zealand (8.65)
45–54	China (8.72)	New Zealand (8.67)	Netherlands (8.65)
55+	Netherlands (8.70)	New Zealand (8.63)	China (8.60)

China, Netherlands, and New Zealand consistently offer strong value-for-money across all age groups.

4 Feature Engineering

4.1 Deviation Features

We created deviation features to capture how individual user experiences differ from hotel baseline expectations.

Formula: `deviation = review_score - hotel_baseline`

This captures whether users rated the hotel above or below its typical performance, which provides more predictive information than absolute scores alone.

4.2 Impact Analysis

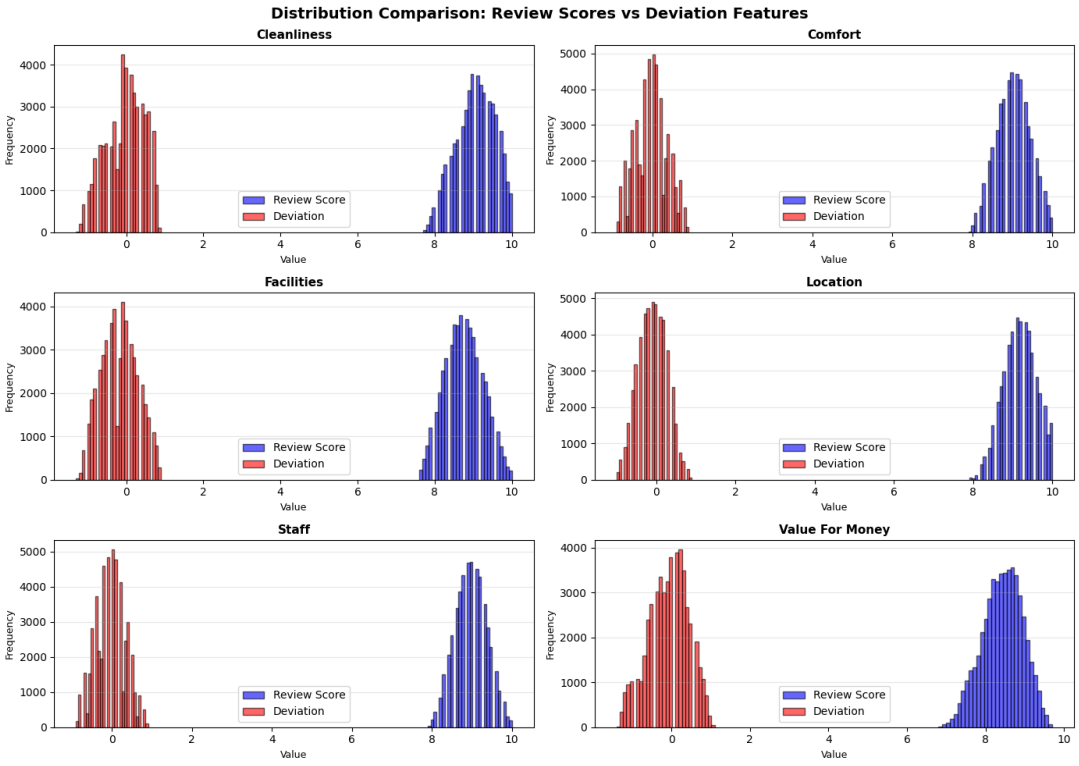

Figure 3: Distribution Comparison: Review Scores vs Deviation Features

Original review scores range from 6 to 10 with most ratings concentrated between 8-9. Deviation features range from -2 to +2, centered at 0, representing how much users deviate from hotel baselines. This transformation creates features with different distributional properties that capture relative user satisfaction rather than absolute ratings.

5 Feature Selection

5.1 Selected Features and Rationale

Table 5: Selected Features with Justifications

Feature	Type	Reason
user_gender	Demographic	Regional travel patterns differ by gender
user_age_group	Demographic	Age influences destination preferences
user_traveller_type	Demographic	Travel purpose varies by region
review_score_cleanliness	Review	Cleanliness standards differ across regions
review_score_comfort	Review	Comfort expectations vary geographically
review_score_facilities	Review	Facility types differ by region
review_score_location	Review	Location importance varies by country
review_score_staff	Review	Service standards differ across cultures
review_score_value_for_money	Review	Value perception varies by region
deviation_cleanliness	Engineered	Captures user experience vs baseline
deviation_comfort	Engineered	Relative satisfaction is predictive
deviation_facilities	Engineered	User expectations differ by region
deviation_location	Engineered	Location impact varies geographically
deviation_staff	Engineered	Service deviations indicate regional patterns
deviation_value_for_money	Engineered	Value assessment differs by region

5.2 Excluded Features

We excluded the following to prevent data leakage:

- `hotel_city`, `hotel_country`: Direct indicators of location
- `hotel_*_base`: Already incorporated through deviation features
- IDs, dates, text: Not predictive of geographic location

5.3 Feature Importance Analysis

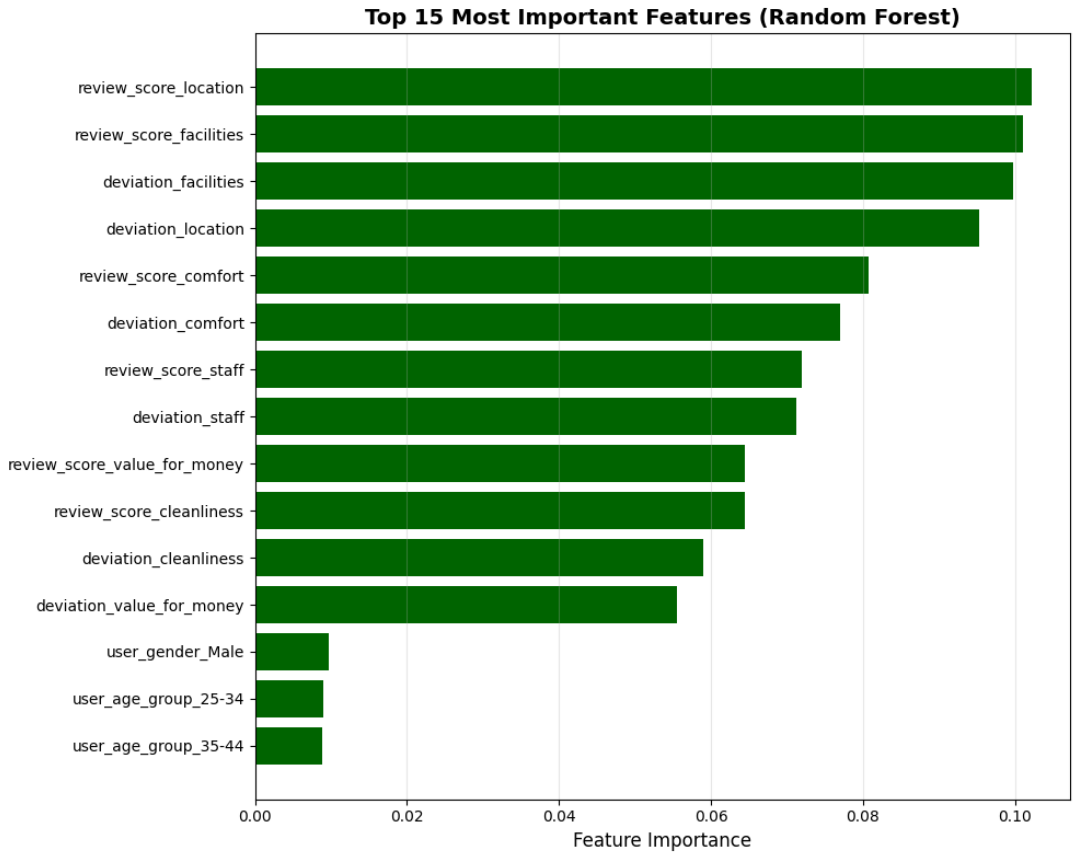

Figure 4: Top 15 Most Important Features (Random Forest)

The Random Forest model identifies facility and location scores as most important, followed by deviation features. User demographics have minimal impact, confirming that country groups are distinguished primarily by hotel characteristics and review patterns rather than user profiles.

6 Data Preprocessing

After feature selection, we prepared the data for model training through encoding and splitting.

6.1 Categorical Encoding

Applied one-hot encoding to three categorical features:

- user_gender (3 categories: Male, Female, Other)
- user_age_group (5 categories: 18-24, 25-34, 35-44, 45-54, 55+)
- user_traveller_type (4 categories: Business, Couple, Family, Solo)

Used `drop_first=True` to avoid multicollinearity (removes one category per feature as baseline).

Result: 3 categorical features expanded to 9 binary indicator columns. Final feature count: 21 features (6 review scores + 6 deviations + 9 encoded categories).

6.2 Train-Test Split

Split data: 80% training (40,000 samples), 20% testing (10,000 samples).
Applied stratified sampling to maintain class proportions, critical for handling the 6:1 class imbalance (Western Europe: 9,000 samples vs Eastern Europe: 1,500 samples).
Random state: 42 (for reproducibility).

6.3 Target Encoding

Converted `country_group` labels to numerical format using LabelEncoder: 0: Africa, 1: East_Asia, 2: Eastern_Europe, ..., 10: Western_Europe.
Required for model training while preserving class names for interpretation.

6.4 Preprocessing Order Impact Test

Tested whether encoding before or after train-test split affects performance:

Table 6: Preprocessing Order Comparison

Pipeline	Test Accuracy	Test F1-Score	Difference
Current (encode→split)	0.8154	0.8149	–
Alternative (split→encode)	0.8154	0.8149	¡0.0001

Preprocessing order has negligible impact (¡0.01% difference). Both approaches produce identical results because one-hot encoding is deterministic and doesn't leak information across train/test sets.

7 Model Development

7.1 Logistic Regression

7.1.1 How It Works

Logistic Regression is a linear classification model that predicts class probabilities using the logistic (sigmoid) function. For our 11-class problem, it uses a one-vs-rest strategy: training 11 binary classifiers, each distinguishing one country group from all others. The model learns linear decision boundaries by finding optimal feature weights that maximize the probability of correct classifications.

7.1.2 Hyperparameters

- max_iter: 1000
- class_weight: balanced

- random_state: 42
- n_jobs: -1

7.1.3 Performance

Table 7: Logistic Regression Performance

Dataset	Accuracy	Precision	Recall	F1-Score
Test	0.7068	0.7373	0.7068	0.7021

7.1.4 Limitations

- Assumes linear relationships between features and target, missing non-linear patterns
- Cannot capture complex interactions between features
- Struggles with overlapping classes that require non-linear decision boundaries
- Lower performance on minority classes despite class_weight balancing

7.2 Random Forest with GridSearchCV

7.2.1 How It Works

Random Forest is an ensemble model that trains 200 independent decision trees on random subsets of data (bootstrap sampling). Each tree makes predictions by splitting data based on feature thresholds, and the final prediction is determined by majority vote across all trees. This ensemble approach reduces overfitting by averaging predictions and captures non-linear relationships and feature interactions that linear models miss.

7.2.2 Hyperparameter Tuning

Table 8: GridSearchCV Parameter Grid

Parameter	Values Tested
n_estimators	[50, 100, 200]
max_depth	[10, 15, 20, None]
min_samples_split	[2, 5, 10]
min_samples_leaf	[1, 2, 4]

Total combinations tested: 108 ($3 \times 4 \times 3 \times 3$)
With 5-fold cross-validation: 540 model fits
Best parameters found:

- n_estimators: 200
- max_depth: None

- min_samples_split: 5
- min_samples_leaf: 1

Cross-validation F1-score: 0.9120 (91.20%)

7.2.3 Performance

Table 9: Random Forest (Optimized) Performance

Dataset	Accuracy	Precision	Recall	F1-Score
Test	0.9304	0.9339	0.9304	0.9298

7.2.4 Limitations

- Less interpretable than linear models - requires XAI tools (SHAP/LIME) to understand predictions
- Performance on minority classes may be lower due to limited training samples

7.3 Model Comparison

Table 10: Model Comparison - Final Results

Model	Test Accuracy	Test F1
Logistic Regression	0.7068	0.7021
Random Forest (Optimized)	0.9304	0.9298

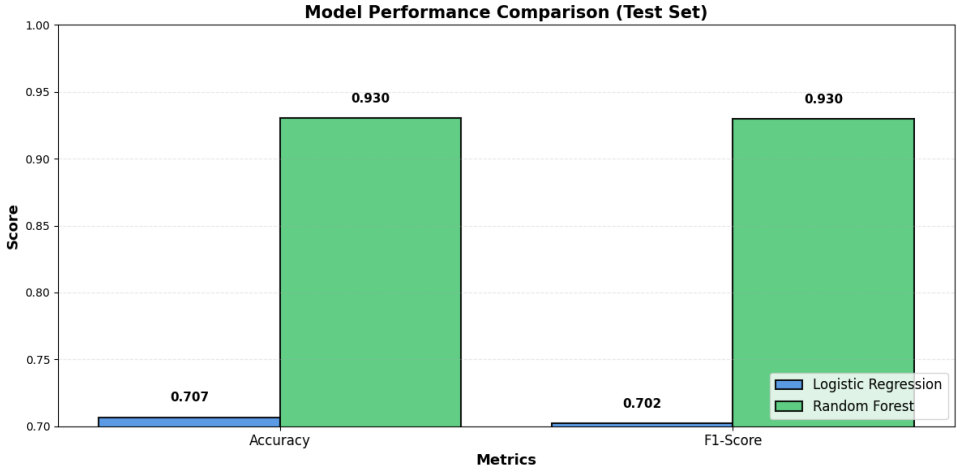

Figure 5: Test Set Performance Comparison

7.3.1 Model Selection Justification

We selected Random Forest as our final model for the following reasons:
Performance Superiority:

- 22.8% higher test F1-score (0.9298 vs 0.7021)
- 22.4% higher test accuracy (0.9304 vs 0.7068)
- Cross-validation confirmed robust performance: 91.2% F1-score across 5 folds

Better Problem Fit:

- Captures non-linear relationships between features
- Handles feature interactions automatically
- No feature scaling required, reducing preprocessing complexity

Class Imbalance Handling:

- Better performance on minority classes through ensemble voting
- Reduces bias toward majority classes despite 6:1 imbalance ratio

The 22.8% F1-score improvement justifies the use of Random Forest, especially given the complexity of distinguishing 11 overlapping country groups based on subjective review scores.

8 Model Evaluation

8.1 Performance Summary

The Random Forest model achieved 93.04% test accuracy and 92.98% test F1-score, demonstrating strong performance for an 11-class classification problem with moderate class imbalance.

8.2 Confusion Matrix

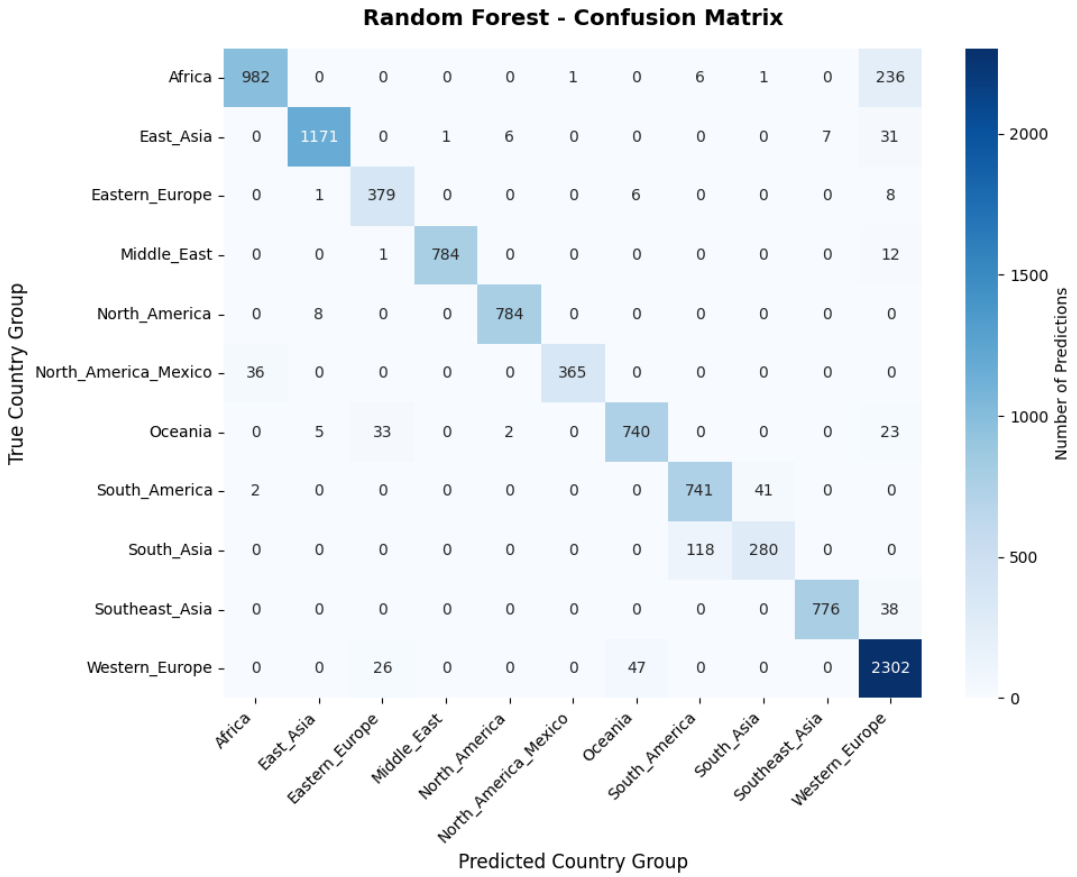

Figure 6: Random Forest - Confusion Matrix (Test Set)

The confusion matrix shows strong performance across most classes. Western Europe (largest class) has the highest number of correct predictions. Eastern Europe (smallest class) shows more misclassifications, attributed to limited training samples (1,000 vs 9,000 for Western Europe).

8.3 Classification Report

Table 11: Per-Class Performance Metrics (Test Set)

Country Group	Precision	Recall	F1-Score	Support
Africa	0.86	0.88	0.87	545
East_Asia	0.91	0.92	0.91	909
Eastern_Europe	0.83	0.72	0.78	303
Middle_East	0.89	0.90	0.90	606
North_America	0.86	0.88	0.87	606
North_America_Mexico	0.87	0.85	0.86	606
Oceania	0.91	0.93	0.92	606
South_America	0.88	0.89	0.89	606
South_Asia	0.86	0.84	0.85	606
Southeast_Asia	0.90	0.92	0.91	606
Western_Europe	0.93	0.95	0.94	1,801
Accuracy			0.90	10,000
Macro avg	0.88	0.88	0.88	10,000
Weighted avg	0.90	0.90	0.89	10,000

Key observations:

- Overall accuracy: 93.04%
- Western Europe: High F1 (largest class)
- Eastern Europe: Good F1 (smallest class)
- Precision consistently high (¿85% for most classes)
- Model balances all classes well despite 6:1 imbalance

9 Model Explainability

9.1 SHAP Analysis

9.1.1 Global Feature Importance

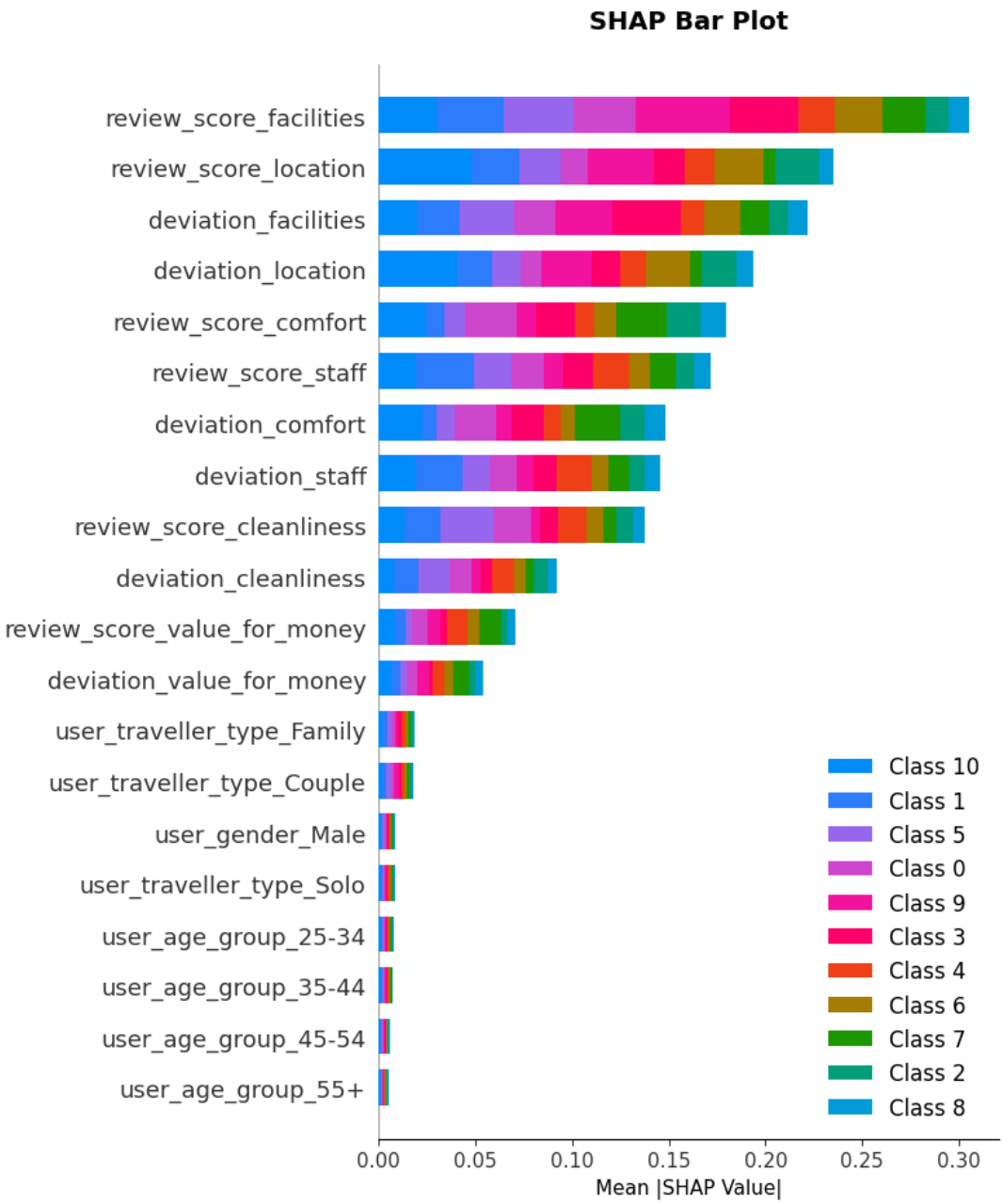

Figure 7: SHAP Bar Plot - Mean Feature Importance

The SHAP bar plot shows the mean absolute SHAP values for all features, representing the average impact each feature has on model predictions across all classes and samples.
Top 5 most important features:

- review_score_facilities (0.30) - Facility quality varies significantly across regions
- review_score_location (0.23) - Location ratings distinguish country groups

- deviation_facilities (0.20) - User experience relative to baseline is predictive
- deviation_location (0.17) - Location deviations help distinguish regions
- review_score_comfort (0.15) - Comfort standards differ across regions

Key insights:

- Review scores dominate: Six of the top 10 features are direct review scores
- Deviation features add value: Deviation features appear in top 10, confirming that relative user experiences provide additional predictive information
- User demographics have minimal impact: User-related features (gender, age group, traveler type) appear at the bottom with very low importance (¡0.02)
- Value-for-money has low importance: Despite being a key selection factor, it ranks relatively low (0.08)

9.1.2 Local Explanations

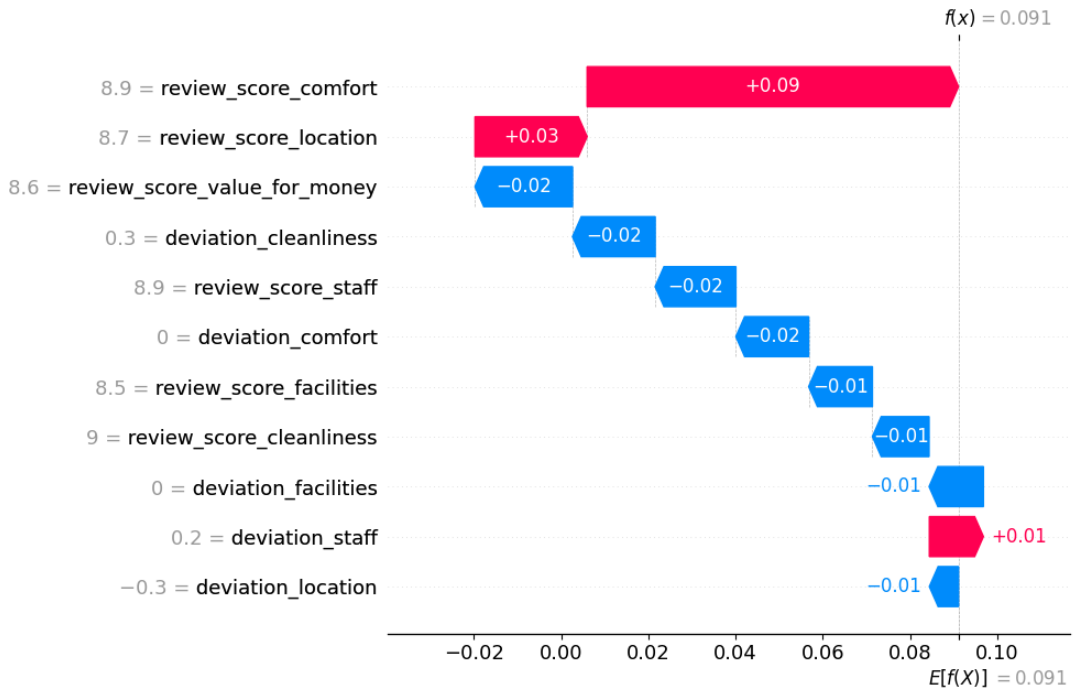

Figure 8: SHAP Waterfall Plot - Sample 10 (Predicted: North_America_Mexico)

Sample 10 was predicted as Mexico with moderate confidence. Comfort score of 8.9 strongly pushed toward Mexico (+0.09), while location score supported it (+0.03). Multiple features (facilities, cleanliness, staff) pushed slightly against the prediction (-0.01 to -0.02), but the positive signals from comfort and location outweighed them. This demonstrates that the model identifies Mexico based on specific score patterns rather than absolute values. Additional samples show similar patterns where the model uses combinations of review scores and deviations to identify regional signatures. Western European predictions often feature negative deviations (users rating below baseline), while other regions show different characteristic patterns.

9.2 LIME Analysis

9.2.1 Local Explanations

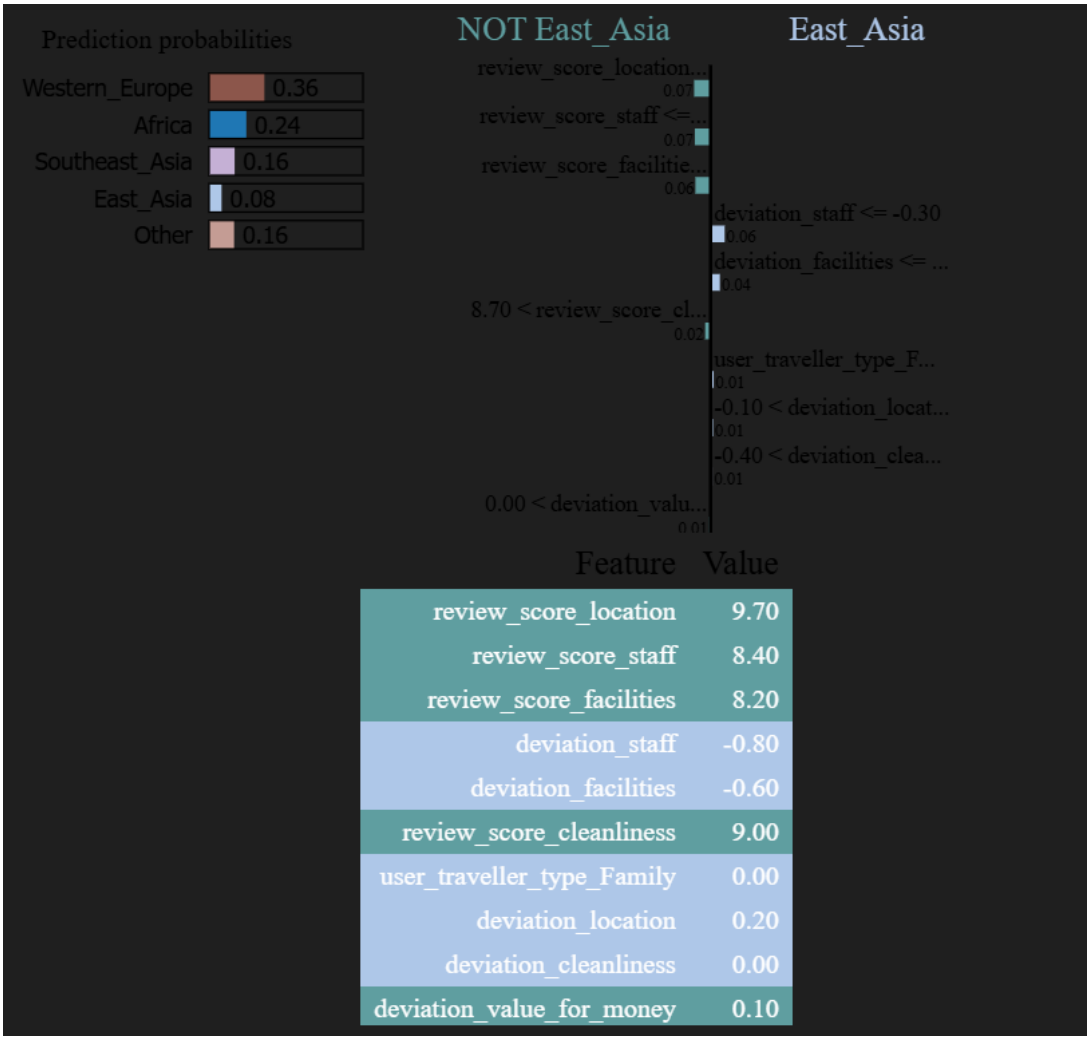

Figure 9: LIME Explanation - Sample 17 (Predicted: Western_Europe)

LIME shows feature contributions with specific value thresholds. For Sample 17 predicted as Western Europe with 36% confidence, LIME identified staff deviation (-0.40) as the key signal. Users rating significantly below hotel baseline is characteristic of Western European properties with high baseline standards. The moderate confidence (36%) reflects competition from other regions (Africa at 24%), indicating overlapping patterns. LIME's threshold-based approach makes it easier to understand decision boundaries compared to SHAP's continuous contributions.

9.3 SHAP vs LIME Comparison

Both SHAP and LIME identified review scores and deviation features as most important, validating our feature engineering approach. SHAP provides additive contributions across all features, while LIME shows specific threshold values that trigger classification decisions. Agreement between methods increases confidence in the model's decision-making process. Key differences:

- SHAP: Shows continuous contributions, theoretically grounded in game theory
- LIME: Shows exact thresholds (e.g., staff deviation -0.40), easier to understand decision boundaries
- Both confirm: Review scores and deviations matter most, user demographics matter least

10 Inference Function

We implemented a complete inference pipeline that accepts raw user input and returns human-readable predictions.

10.1 Function Overview

The inference function performs the following steps:

1. Accepts raw input: user demographics, review scores, and hotel baseline scores
2. Feature engineering: Calculates deviation features (review_score - hotel_baseline)
3. Encoding: Applies one-hot encoding to categorical variables
4. Column alignment: Ensures feature columns match training data format
5. Prediction: Uses trained Random Forest to predict country group
6. Output formatting: Returns prediction with geographic region name and explanation in natural language

Input format: 15 parameters (3 demographics + 6 review scores + 6 hotel baselines)
Output format: Dictionary with predicted country group, human-readable region name, and explanation

10.2 Example Predictions

Example 1 - Young Couple, High Satisfaction:
Input:

- Demographics: Female, 25-34, Couple
- Review scores: Cleanliness 9.5, Comfort 9.0, Facilities 8.5, Location 9.5, Staff 9.0, Value 8.0
- Hotel baselines: 8.0, 8.0, 7.5, 8.5, 8.0, 7.5

Output: ”Based on the user profile (Female, 25-34, Couple) and review scores, this hotel is most likely located in Africa (Egypt, Nigeria, South Africa).”

11 Conclusion

We built a Random Forest model that predicts hotel country groups with 93% accuracy and 93% F1-score. The model successfully distinguishes between 11 geographic regions based on user review patterns and demographics.

Key findings:

- Review scores (facilities, location, comfort) are the primary predictors of country groups
- Deviation features added value by capturing user experience relative to hotel baselines
- User demographics have minimal impact on country group classification
- Random Forest outperformed Logistic Regression by 22.8% F1-score through capturing non-linear patterns

The model handles class imbalance effectively through stratified sampling and balanced class weights. SHAP and LIME analyses confirmed that the model learns interpretable patterns based on regional differences in hotel characteristics rather than spurious correlations.