Airbnb New User Bookings

***Report*: Predicting First Booking Destination for Airbnb New Users**

---

1. Introduction

Airbnb is a platform where users can find places to stay and unique travel experiences. This project focuses on predicting the first country a new user will book. By doing this, Airbnb can improve recommendations, make users happier, and plan resources better.

---

2. Data Overview

- **User Demographics:** Includes age, gender, and signup method.
- **Booking Information:** Contains the first booking destination, account creation dates, and first booking dates.
- **Session Details:** Tracks user actions, devices, and time spent on the platform.
- **Imbalance:** A significant proportion of data falls under the "NDF" category, representing users who did not select a travel destination.

Dataset Description:

- **Training Data:** `train_users.csv` contains user demographics, signup details, and target variable `country_destination`.
- **Test Data:** `test_users.csv` includes user information for which predictions need to be made.
- **Sessions Data:** `sessions.csv` tracks user actions, devices, and session durations.
- **Supplementary Files:**
    - `countries.csv`: Summary statistics of destination countries.
    - `age_gender_bkts.csv`: Statistics of users' age groups, gender, and destinations.
    - `sample_submission.csv`: Format for predictions submission.

Target Variable:
The target variable `country_destination` has 12 possible outcomes:
US, FR, CA, GB, ES, IT, PT, NL, DE, AU, NDF (no destination found), and other.

- **NDF:** No booking was made.
- **other:** Booking to a country not listed among the major destinations.

Evaluation Metric:
Normalized Discounted Cumulative Gain (NDCG@5) measures the quality of predictions based on their relevance and ranking.

---

3. Exploratory Data Analysis (EDA)

User Demographics:

- **Age Distribution:**
    - Most users are aged between 30 and 40 years.
    - The 46-55 age group has the highest number of active bookers.
    - Outliers (ages over 100) were replaced with the mean.
- **Gender:**
    - A significant portion of users are classified as "Unknown," while among the specified genders, females slightly outnumber males.

Booking Trends:

- **Timing:**
  Booking activity occurs mid-year, particularly in June and July.
  The busiest year for first bookings was 2013.
  Users are more likely to book around the 15th of any given month.

- **Destinations:**
  The US is the most frequently chosen destination, followed by France and "Other."
  Over 50% of users fall under the "NDF" category, indicating no travel destination selected.

Behavioral Insights from Sessions Data:

- **Devices:**
  Most users access Airbnb through Mac and Windows desktops, followed by iPhones.
  Android devices and tablets are less commonly used.

- **Actions:**
  Viewing and clicking actions dominate user sessions.
  Rare actions like "booking requests" are sparse but critical for conversion analysis.

- **Time Spent:**
  Younger users (18-25) spend the most time exploring the platform.
  Older users (above 55) tend to spend less time before making decisions.

Correlations and Patterns:

- Strong correlations exist between the timing of bookings (e.g `day_first_booking`) and user engagement metrics.
- Device usage correlates with total time spent on the platform, indicating exploratory behavior across multiple devices.
- Weak correlations were found between signup methods and engagement, suggesting uniform interaction across different pathways.

---

4. Proposed Solution

Data Preprocessing:

- Handled missing values and outliers (e.g, replaced unrealistic ages with the mean).
- Applied label encoding for categorical features (e.g., gender, language).
- Scaled numerical features using StandardScaler.

Feature Engineering:

- Retained only relevant features based on EDA insights and feature importance analysis.
- Dropped less relevant features (eg., language, signup_app) to improve model efficiency.

Model Selection:

- Evaluated models using **GridSearchCV** to optimize hyperparameters:
    - **Random Forest Classifier:** Achieved the best trade-off between accuracy and efficiency with a validation accuracy of **88%.**
    - **Other Models:** Decision Tree, XGBoost, and LightGBM models performed similarly but offered no significant improvement over Random Forest.

Feature Importance Analysis:

- **Key features:** `day_first_booking`, `month_first_booking`, `age`, and `total_secs`.
- **Less important features**: `language`, `num_devices`, and `signup_app`.

Advanced Techniques:

- **Imbalanced Data Handling:**
  SMOTE (Synthetic Minority Oversampling Technique) was tested but offered negligible improvement.

- **NDCG Evaluation:**
  Normalized Discounted Cumulative Gain (NDCG) was used to measure ranking performance (Score: **93%**).

- **Ensemble Learning:**
  Combined models using a **Voting Classifier** for ensemble learning.
  Achieved consistent validation accuracy of **88%.**

- **Neural Network:**
  Designed a multi-layer neural network using **Keras**.
  Achieved **88%** accuracy but showed signs of overfitting.

Deployment:

- Built a **Flask** web application for real-time prediction.
- Saved the trained Random Forest model as a `.pkl` file for integration.

---

5. Results and Insights

Model Performance:

- **Random Forest Classifier** provided the best performance with a validation accuracy of 88%.
- **NDCG Score:** 93%, indicating strong ranking performance.

Key Takeaways:

- **Timing of the first booking** (day, month, year) is the most critical predictor.
- **User engagement metrics** like total actions and time spent are important but secondary.
- **Imbalanced data** (high proportion of "NDF") requires further exploration for better insights.

Deployment:

- Successfully developed and tested a **Flask application** for local deployment.
- The model is ready for production deployment.

---

6. Recommendations and Next Steps

- **Production Deployment:**
    - Deploy the **Flask application** to a production server to enable broader access and real-time predictions.
- **Marketing Strategy:**
    - Use the insights to improve Airbnb's marketing strategies, focusing on peak booking seasons (e.g., June and July) and device usage patterns (e.g., desktop and iPhone preferences).
- **Model Improvement:**
    - Explore advanced techniques like **stacking** or **deep learning ensembles** for incremental improvements in prediction accuracy.
    - Further analyze and address the "NDF" category to derive actionable insights about non-traveling users, potentially improving booking rates.
- **Data Exploration:**
    - Investigate more on **users who do not select a destination** ("NDF"), which could lead to valuable insights into behavior and intervention strategies.

Conclusion: This comprehensive report outlines the dataset, EDA findings, modeling strategies, and deployment steps, providing a robust framework for predicting Airbnb's first booking destinations effectively.
