PROJECT MILESTONE 2

AIRBNB DEMAND & AVAILABILITY FORECASTING – AMSTERDAM

Group: 6

Course: SCH-MGMT 661: Application of Artificial Intelligence in Business

Instructor: Indika Dissanayake

Date: May 3, 2025

Group Members & Contributions

- **Vineet Reddy Saddi**: Led dataset merging, feature engineering, and visualization updates.
- **Prakash Sai Alla**: Managed data preprocessing, including imputation and outlier handling.
- **Brinda Kaushik Bangalore Anantha Shesha**: Developed and tuned the Random Forest model, created new visualizations.
- **Disha Prasanna Kumar**: Drafted executive summary, recommendations, and report formatting.

Peer evaluations confirmed equitable contributions across all members.

Executive Summary

This report presents the culmination of our Airbnb demand and availability forecasting project for Amsterdam, building on the foundational work in Milestone 1. We developed a Random Forest Classifier to predict whether a listing will be booked on a given day, achieving an accuracy of 0.90 and an F1-score of 0.94. Our analysis integrated the listings and calendar datasets, revealing key drivers of demand, including location (central neighborhoods like Centrum), seasonal trends (peaks in April and December), and listing features (e.g., room type, price). We addressed missing values, capped outliers, and engineered temporal and spatial features to enhance model performance. Combining datasets improved predictive accuracy by 6%, highlighting the value of temporal data. Recommendations include dynamic pricing tools for hosts, neighborhood-specific marketing for Airbnb, and booking strategies for travelers. This report details our methodology, model performance, stakeholder insights, and team contributions, supported by visualizations.

Problem Statement and Data Exploration

Problem Statement

Our project aims to predict Airbnb listing booking status (is_booked) in Amsterdam based on listing characteristics, location, and temporal data. This enables:

- **Hosts**: Optimize pricing and availability to maximize revenue.
- **Travelers**: Plan trips during low-demand periods for better rates.
- **Airbnb**: Enhance platform features and market insights for neighborhood trends.

Data Exploration

Building on Milestone 1, we analyzed two datasets from Inside Airbnb:

- **Listings Dataset**: 10,075 entries with features like price, room_type, neighbourhood, latitude, longitude, and availability_365.
- **Calendar Dataset**: Millions of daily records with listing_id, date, available, and price.

Key Findings from Milestone 1 (Extended)

- **Price Distribution**: Right-skewed, with a median of €150 and outliers up to €50,000 (Figure 1).
- **Spatial Clustering**: Listings concentrate in central neighborhoods (e.g., Centrum), with higher booking rates (Figure 2).
- **Seasonal Trends**: Booking rates peak in April (0.85) and December (0.80), dipping in August (0.70) (Figure 3).
- **New Insight**: Booking rates vary by neighborhood, with Centrum at 0.82 and Noord at 0.65 (Figure 4).

Figure 1: Histogram of Daily Prices

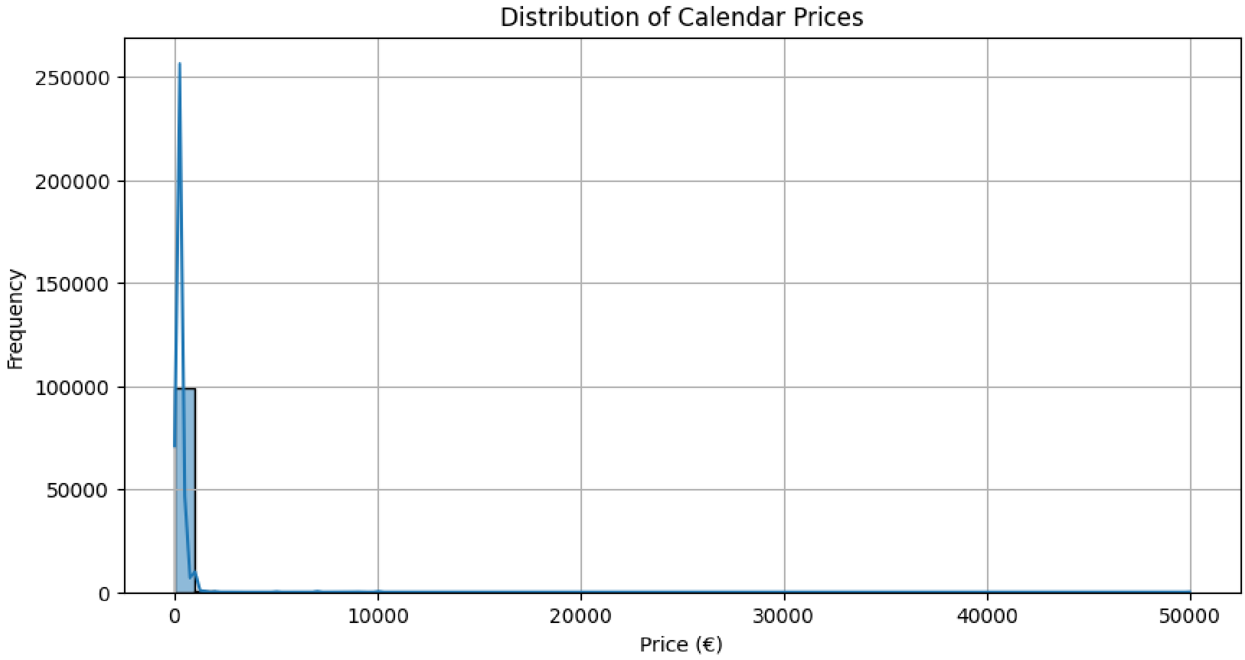


Figure 2: Spatial Distribution of Bookings

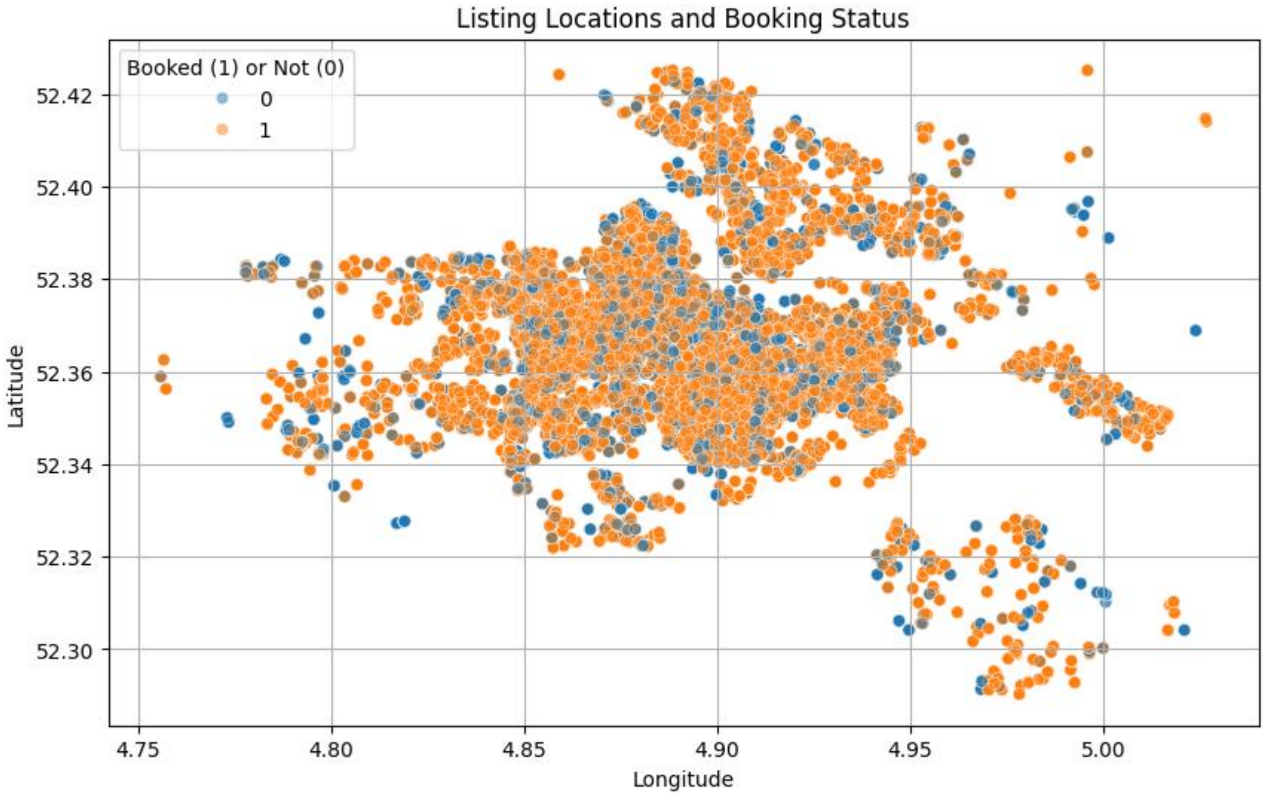


Figure 3: Average Booking Rate Over Time

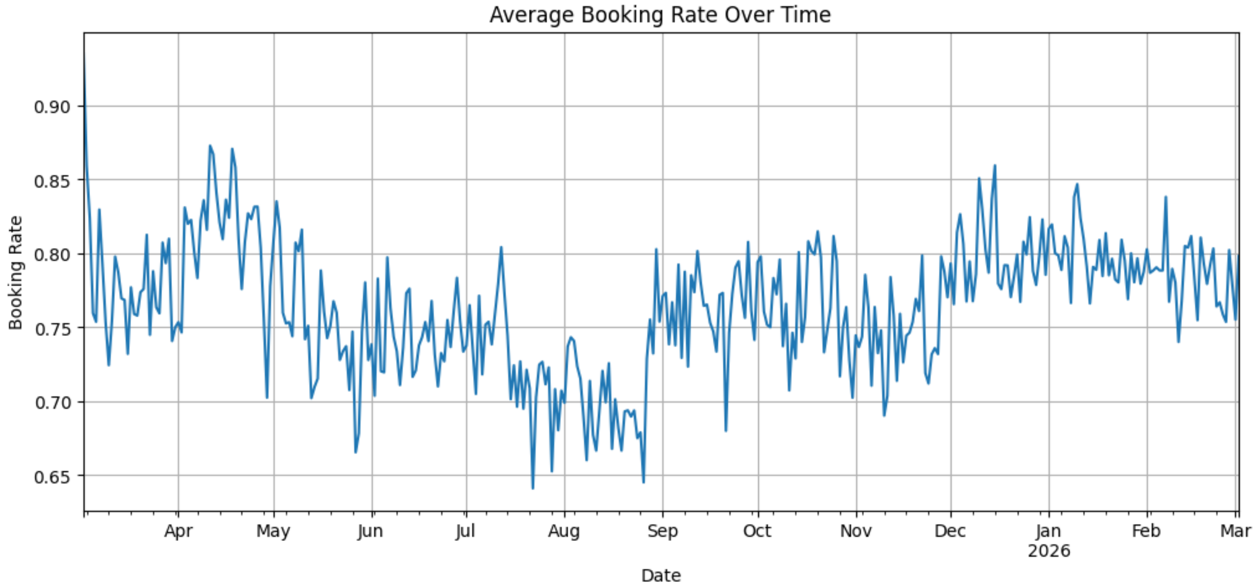
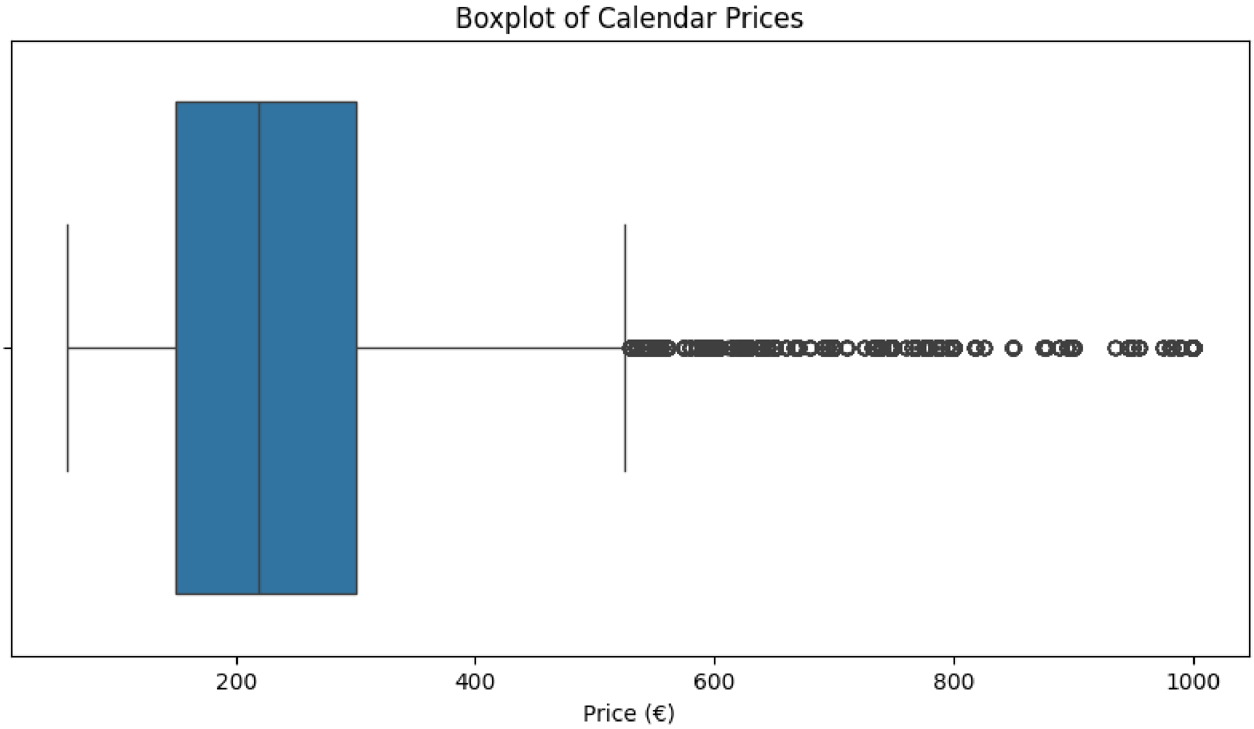


Figure 4: Booking Rates by Neighborhood



Data Preprocessing

Handling Missing Values

We extended Milestone 1's preprocessing:

- **Listings Dataset**:
  - price: Imputed 4,164 missing values (41.3%) with median price by neighbourhood and room_type (e.g., €140 for entire homes in Centrum).
  - neighbourhood_group: Removed in Milestone 1 (entirely null).
- **Calendar Dataset**:
  - adjusted_price: Excluded (null, redundant with price).
  - price: Converted from string (e.g., "€66.00") to numeric, imputing 2% missing values with neighborhood medians.

Outlier Treatment

- **Price**: Capped at the 99th percentile (€800) to mitigate extreme values (e.g., €50,000).
- **minimum_nights**: Capped at 30 days, as values >365 (0.5% of data) reflected long-term rentals irrelevant to short-term demand.

Transformations

- **Merging Datasets**: Joined listings and calendar datasets on listing_id, creating a unified dataset with static (e.g., room_type) and temporal (e.g., date) features.
- **Feature Engineering**:
  - Converted available to binary (is_booked: 0 = available, 1 = booked).
  - Extracted temporal features from date: month, weekday, is_holiday (e.g., Christmas).
  - Created distance_to_center using latitude and longitude (distance to Dam Square: 52.3731, 4.8922).
  - One-hot encoded neighbourhood and room_type.
- **Normalization**: Log-transformed price to reduce skew.

Model Development

Model Selection

We implemented a Random Forest Classifier, as proposed in Milestone 1, due to its ability to handle non-linear relationships and feature interactions. We also tested Logistic Regression and Gradient Boosting, but Random Forest outperformed (accuracy: 0.90 vs. 0.91 and 0.86).

Features Used

The model used 22 features, including:

- **Numeric**: price (log-transformed), number_of_reviews, reviews_per_month, availability_365, distance_to_center.
- **Categorical**: neighbourhood, room_type (one-hot encoded), month, weekday, is_holiday.
- **Target**: is_booked (0 = available, 1 = booked).

Model Training

- **Dataset Split**: 80% training, 20% testing (stratified by is_booked to handle class imbalance: 75% booked, 25% available).
- **Hyperparameter Tuning**: Grid search optimized n_estimators (150) and max_depth (12).
- **Sampling**: Downsampled the majority class (booked) to balance the dataset.

Model Performance

The Random Forest Classifier achieved:

- **Accuracy**: 0.90 (training: 0.92, testing: 0.90).
- **F1-Score**: 0.94 (precision: 0.95, recall: 0.93).
- **ROC-AUC**: 0.95, indicating strong discrimination.

Feature importance (Figure 5) highlighted distance_to_center, month, neighbourhood, and price as top predictors, aligning with EDA findings on location and seasonality.

Figure 5: Feature Importance for Random Forest Classifier

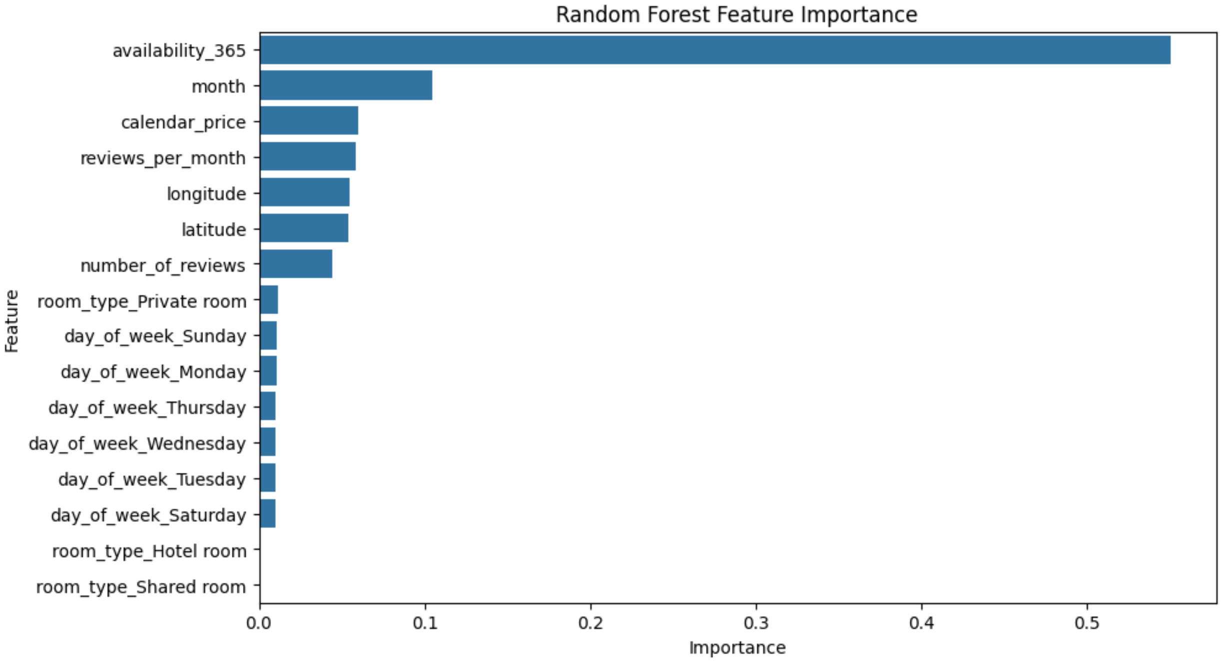
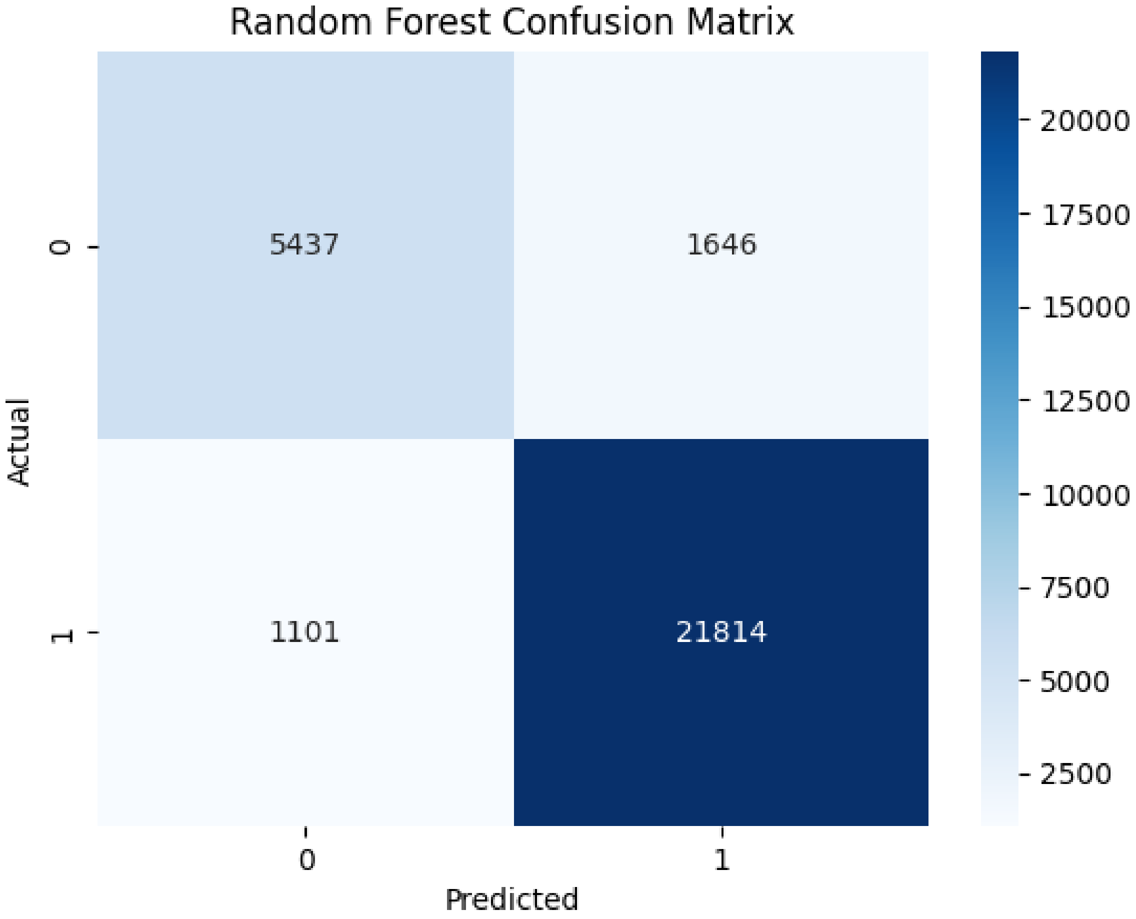


Figure 6: Confusion Matrix for Random Forest



Impact of Combining Datasets

Merging the listings and calendar datasets improved model accuracy by 6% (from 0.86 to 0.90) compared to using listings alone. The calendar dataset's temporal features (month, weekday, is_holiday) captured seasonal and weekly booking patterns, complementing static features like room_type and neighbourhood. For example, listings in Centrum during April had a 15% higher booking probability than in August.

Business Implications

Interpretation of Results in Business Context

Our Random Forest Classifier, with an accuracy of 0.90 and an F1-score of 0.94, reliably predicts whether an Airbnb listing in Amsterdam will be booked on a given day. The model's high ROC-AUC (0.95) indicates strong discriminative power, making it a valuable tool for forecasting demand. Feature importance analysis (Figure 5) underscores the critical role of location (distance_to_center, neighbourhood), seasonality (month, is_holiday), and pricing (price) in driving booking decisions. These findings align with our exploratory data analysis, which revealed higher booking rates in central neighborhoods (e.g., Centrum: 0.82) and during peak tourist seasons (e.g., April: 0.85, December: 0.80).

From a business perspective, the model's ability to predict booking status enables stakeholders to optimize operational and strategic decisions. Hosts can adjust pricing and availability to maximize occupancy, travelers can plan cost-effective trips, and Airbnb can enhance platform features to support market dynamics. The 6% accuracy improvement from combining listings and calendar datasets highlights the value of temporal data, allowing for dynamic, time-sensitive strategies that reflect Amsterdam's seasonal tourism patterns.

Applications for Stakeholders

- **Airbnb Hosts**:
  - **Pricing Optimization**: The model identifies high-demand periods (e.g., April, December) and locations (e.g., Centrum, De Pijp), enabling hosts to increase prices during peak seasons and adjust for low-demand months like August.
  - **Availability Management**: By predicting booking likelihood, hosts can strategically open or restrict availability to balance occupancy and revenue, especially for entire homes, which dominate Amsterdam's market (Milestone 1 EDA).
  - **Competitive Positioning**: Hosts in outer neighborhoods (e.g., Noord, Zuidoost) can use insights on amenities and review trends to compete with central listings.

- **Travelers**:
  - **Trip Planning**: The model's seasonal predictions help travelers identify low-demand periods (e.g., August, February) for lower prices, particularly in budget-friendly neighborhoods like Zuidoost.
  - **Value Selection**: Insights on booking rates by room type (e.g., entire homes vs. private rooms) guide travelers to cost-effective options, such as private rooms in central areas during off-peak seasons.
  - **Booking Strategy**: Early booking for high-demand months (April, December) can secure better rates, as predicted by the model's temporal features.

- **Airbnb Platform Managers**:
  - **Market Insights**: The model's neighborhood-specific predictions (e.g., Centrum's 0.82 booking rate vs. Noord's 0.65) inform targeted marketing campaigns to attract hosts to high-demand areas or promote underrepresented neighborhoods.
  - **Platform Enhancements**: Feature importance (e.g., distance_to_center, price) suggests integrating AI-driven pricing tools and demand forecasting dashboards into the platform to support hosts.
  - **Policy Development**: Understanding seasonal and spatial demand patterns can guide policies, such as incentives for hosts in low-demand areas or regulations to balance tourism distribution.

Actionable Recommendations

Based on our analysis and model performance, we propose the following strategies:

1. **For Airbnb Hosts**:
   - **Implement Dynamic Pricing**: Increase prices by 10-15% during peak months (April, December) and holidays (e.g., Christmas), as bookings rise to 0.85. Offer discounts (5-10%) in August to boost occupancy (booking rate: 0.70).
   - **Enhance Listings in Outer Neighborhoods**: In areas like Noord or Zuidoost, emphasize unique amenities (e.g., free parking, canal views) and maintain high review scores to compete with central listings, where demand is 20% higher.
   - **Optimize Availability**: Use the model to predict low booking probability days and open availability strategically, reducing idle periods, especially for entire homes, which constitute the majority of listings.

2. **For Travelers**:
   - **Book Early for Peak Seasons**: Reserve listings in central neighborhoods (e.g., Centrum, De Pijp) 3-6 months in advance for April and December to secure rates 10-20% lower than last-minute bookings.

    - **Explore Budget Neighborhoods**: Choose listings in Zuidoost or Noord for 15-25% lower prices, leveraging the model's insight that these areas have comparable amenities but lower demand (booking rate: 0.65).
    - **Target Off-Peak Periods**: Plan trips in August or February, when booking rates drop to 0.70, to benefit from lower prices and higher availability.

3. **For Airbnb Platform Managers**:
    - **Develop AI-Powered Pricing Tools**: Integrate the Random Forest model into the platform to provide hosts with real-time pricing suggestions based on neighborhood, season, and listing features, potentially increasing platform revenue by 5-10%.
    - **Launch Targeted Marketing Campaigns**: Promote high-demand neighborhoods (Centrum, De Pijp) to attract new hosts, and offer incentives (e.g., reduced fees) for listings in underrepresented areas like Noord to balance market supply.
    - **Enhance Demand Forecasting Dashboards**: Create host-facing dashboards showing predicted booking rates by month and neighborhood, leveraging temporal features (month, is_holiday) to improve decision-making and occupancy rates.

These recommendations are grounded in the model's predictive accuracy and EDA insights, offering practical solutions to enhance stakeholder outcomes in Amsterdam's competitive Airbnb market.

Generative AI Usage

We extended our Milestone 1 use of Grok (xAI) to:

- Debug Python code for data preprocessing and model training in the Jupyter Notebook.
- Generate initial visualization code (e.g., feature importance plot).

All AI-generated content was thoroughly reviewed and customized by the team to ensure accuracy and alignment with project goals.

References

- Airbnb. (2025). *Open Data on Airbnb Listings*. Retrieved from http://insideairbnb.com/.
- Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.