A. Company Scoreboard

Company name: Airbnb is a global online marketplace that connects hosts offering accommodations with guests seeking short-term stays.

Company product / service: Airbnb provides a digital platform that enables users to list, discover, and book accommodations worldwide.

Company business model: Airbnb operates a two-sided marketplace, earning revenue by charging service fees to both hosts and guests per booking.

Company size: Airbnb is a large-scale global company with millions of active listings and users across more than 190 countries.

Company mission: Airbnb's mission is to create a world where anyone can belong anywhere by enabling meaningful travel experiences.

Company sector: Airbnb operates in the travel, hospitality, and sharing economy sectors.

Company location(s): Airbnb is headquartered in San Francisco, California, with operations and listings worldwide.

B1. Problem Statement

Business problem: Airbnb needs to proactively identify hosts who are likely to become Superhosts in order to retain high-quality hosts in an increasingly competitive short-term rental market and differentiate itself from competing platforms (Namely Wimdu in Berlin).

Business metric(s): The primary business metric we aim to impact is the Superhost Conversion Rate. Utilizing the model, we plan to increase the rate through accurately identifying the high-potential hosts, which Airbnb can provide targeted resources and incentives to help them meet Superhost criteria. The current Superhost percentage in Berlin is 54.32%, and we aim to increase this to 65%.

Algorithm: Random Forest Classifier is our main algorithm. It conducts supervised learning with a labeled Berlin Airbnb dataset. Its task is to give a binary classification to distinguish between "Superhost" (1) and "Regular Host" (0).
We chose Random Forest for its ability to handle non-linear features and provide Feature Importance rankings, which are essential for identifying the key drivers of host success and informing business decisions. Beyond binary classification, the model provides class probabilities, allowing Airbnb to identify 'high-potential' hosts who are nearing the Superhost

threshold. This enables precise marketing interventions to bridge the gap and maximize conversion efficiency.

Chosen metric(s) for the algorithm: To evaluate the effectiveness of our Random Forest model, we utilize two complementary metrics:

- F1 Score: This is our primary metric. It provides a balance between Precision (avoiding false alarms for potential Superhosts) and Recall (ensuring we don't miss actual high-potential candidates). Given the business cost of misallocating marketing resources, the F1 score ensures a robust trade-off.
- ROC-AUC: This measures the model's ability to rank hosts. Since our strategy involves identifying "potential" Superhosts based on probability, a high ROC-AUC ensures that a host with a higher predicted probability is indeed more likely to be a Superhost than one with a lower probability.

Dataset description: Source: Kaggle: https://www.kaggle.com/datasets/thedevastator/berlin-airbnb-ratings-how-hosts-measure-up/data

The dataset contains detailed information on Airbnb listings in Berlin, including host characteristics (e.g., response time, response rate, tenure), listing attributes (e.g., property type, room type, capacity, price), and guest feedback (e.g., number of reviews and rating scores). This data enables analysis of host behavior and performance patterns that are relevant for predicting Superhost status using supervised machine learning methods.

```
[14]: # column names
      df.columns
```

```
[14]: Index(['index', 'Review ID', 'review_date', 'Reviewer ID', 'Reviewer Name',
             'Comments', 'Listing ID', 'Listing URL', 'Listing Name', 'Host ID',
             'Host URL', 'Host Name', 'Host Since', 'Host Response Time',
             'Host Response Rate', 'Is Superhost', 'neighbourhood',
             'Neighborhood Group', 'City', 'Postal Code', 'Country Code', 'Country',
             'Latitude', 'Longitude', 'Is Exact Location', 'Property Type',
             'Room Type', 'Accomodates', 'Bathrooms', 'Bedrooms', 'Beds',
             'Square Feet', 'Price', 'Guests Included', 'Min Nights', 'Reviews',
             'First Review', 'Last Review', 'Overall Rating', 'Accuracy Rating',
             'Cleanliness Rating', 'Checkin Rating', 'Communication Rating',
             'Location Rating', 'Value Rating', 'Instant Bookable',
             'Business Travel Ready'],
            dtype='object')
```

D2. Baseline Model Performance

Baseline Strategy: The baseline model utilizes a "Most Frequent Class" strategy, which always predicts the majority class found in the training data (in this case, 'Superhost' = 't'). This provides

a minimum performance threshold that any subsequent machine learning model must exceed to be considered valuable.

Performance Summary Table:

Metric	**Trainig Set**	**Validation Set**
Accuracy	0.5432	0.5432
F1 Score	0.7040	0.7040
ROC-AUC	0.5000	0.5000

Interpretation:

The baseline metrics reflect a "Most Frequent Class" strategy. While the Accuracy (54.32%) and F1 Score (0.7040) appear moderate due to the majority class bias, they represent a model with zero predictive power, as confirmed by the ROC-AUC of 0.5000.

These figures establish our performance floor: they prove that simply guessing the most common category provides no strategic insight. In Part 2, our Random Forest model must significantly outperform these benchmarks—particularly in ROC-AUC—to demonstrate that it has successfully learned to distinguish potential Superhosts from regular ones using host behavior data.