Data Appendix: Airport Review Analysis

Dataset 1: Cleaned Reviews

This is our primary dataset. The unit of observation is a **single review from one traveler**.

- In `airport_clean.csv`
- Used to train linear regression model

Variables

10,849 reviews and **10 columns**:

- **Who & Where:** `airport_name`, `author`, and `author_country`
- **When:** `date` (Ranges from 2008 to 2015)
- **What:** `content` (The actual text of the review)
- **The Scores:**
  - `overall_rating`: The main score (1–10 scale)
  - Sub-ratings: `queuing`, `terminal_cleanliness`, `airport_shopping` (0–5 scale)
  - `recommended`: target variable, 1 = Yes, 0 = No

Stats

- **Average Rating:** 4.58 / 10, leans negative
- **Sub-ratings:** People were generally happier with Cleanliness (3.6/5) than Queuing or Shopping (both ~3/5)

Dataset 2: Processed Text

- In `preprocessed.csv`
- Used to train logistic regression model
- One additional column: `processed_content` The unit of observation is a **cleaned text string** ready for analysis

How we cleaned it

We ran the raw reviews through a script that: 1 **Tokenized:** Split sentences into individual words 2 **Lemmatized:** Converted words to their root form (ex. "flying" -> "fly") 3 **Removed Noise:** Deleted punctuation, numbers, and boring stop words like "the" or "and"

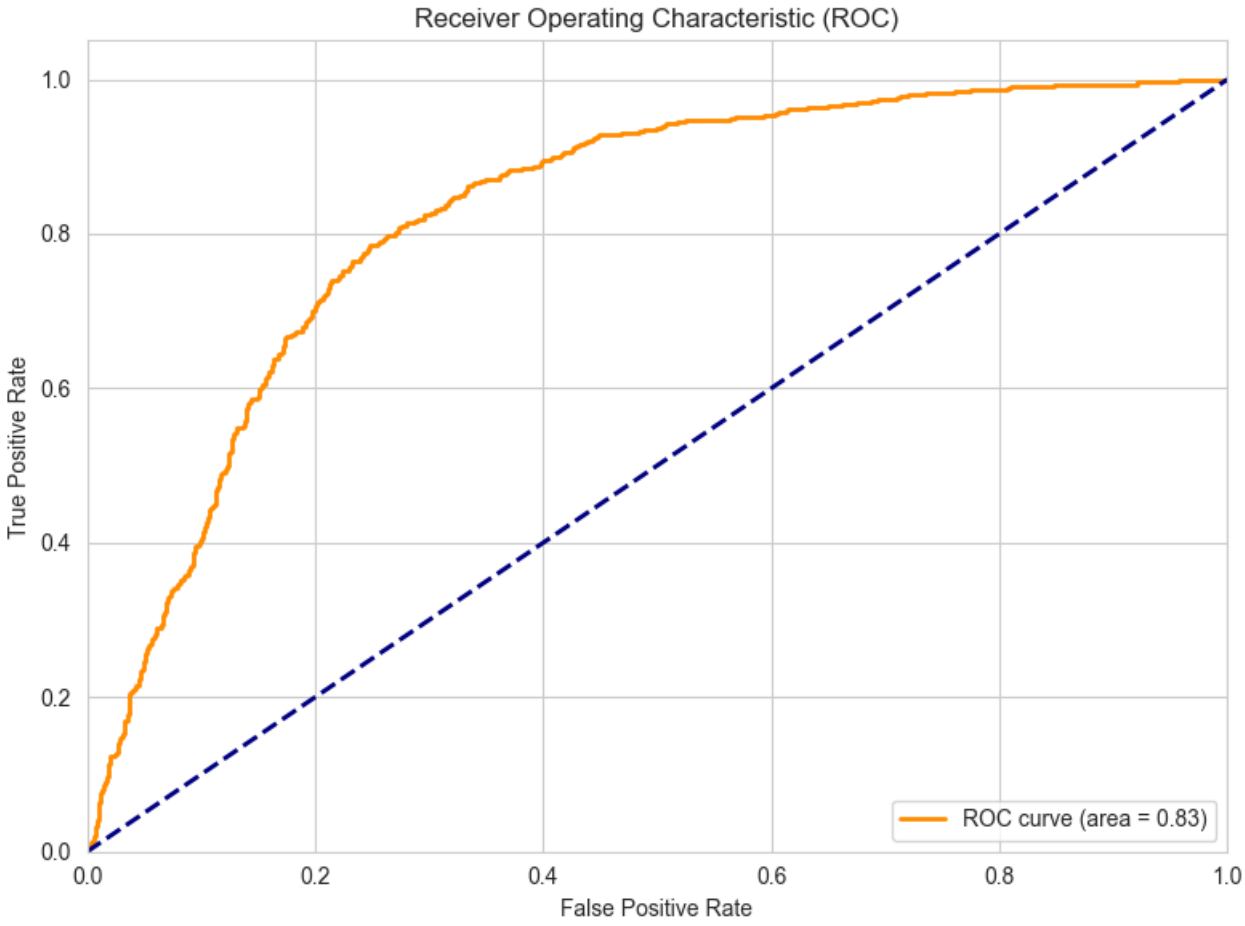
Class Balance

We are trying to predict the `recommended` column, which is imbalanced:

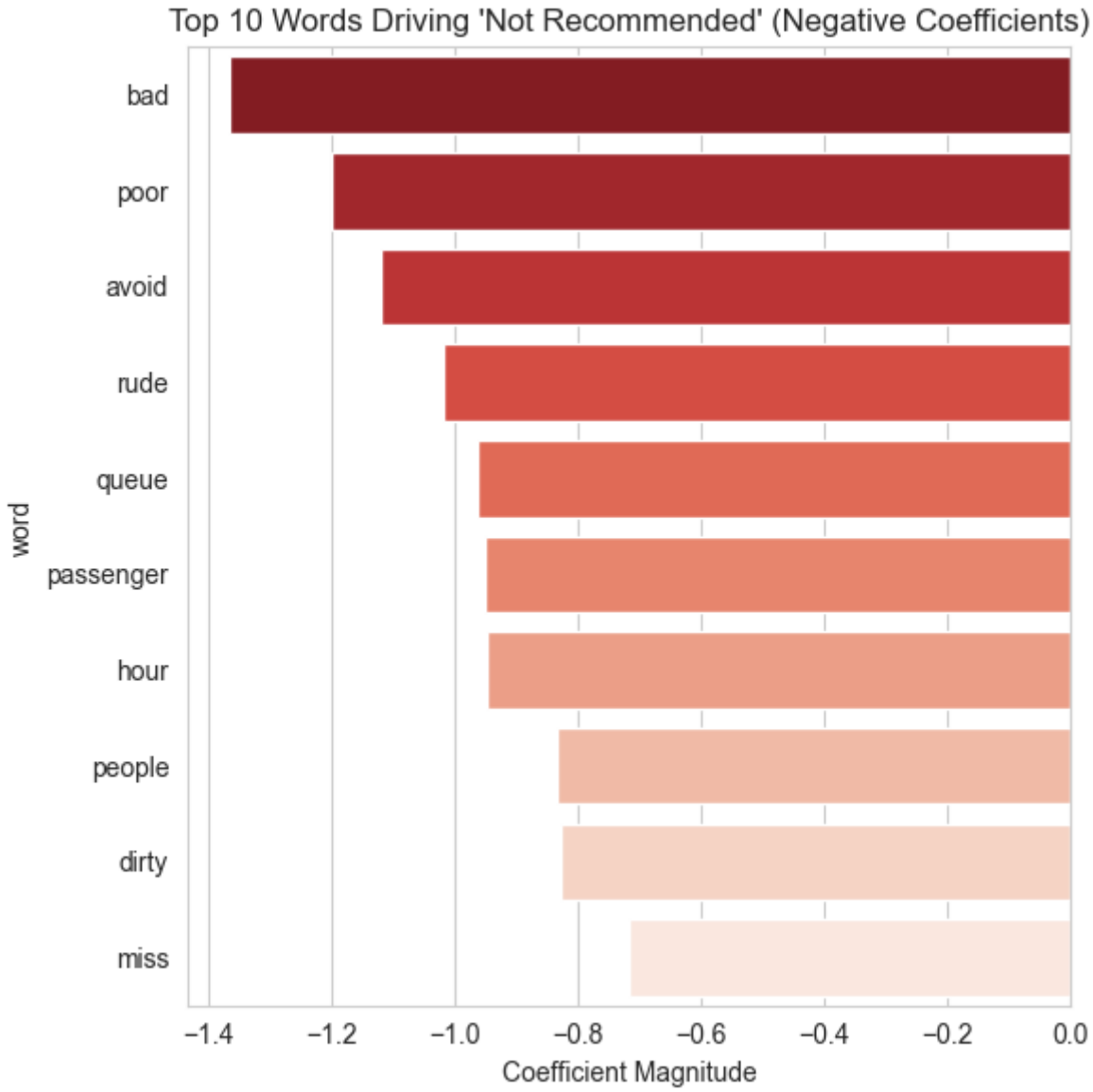
- **Not Recommended:** ~64%
- **Recommended:** ~36%

Figures & Outputs

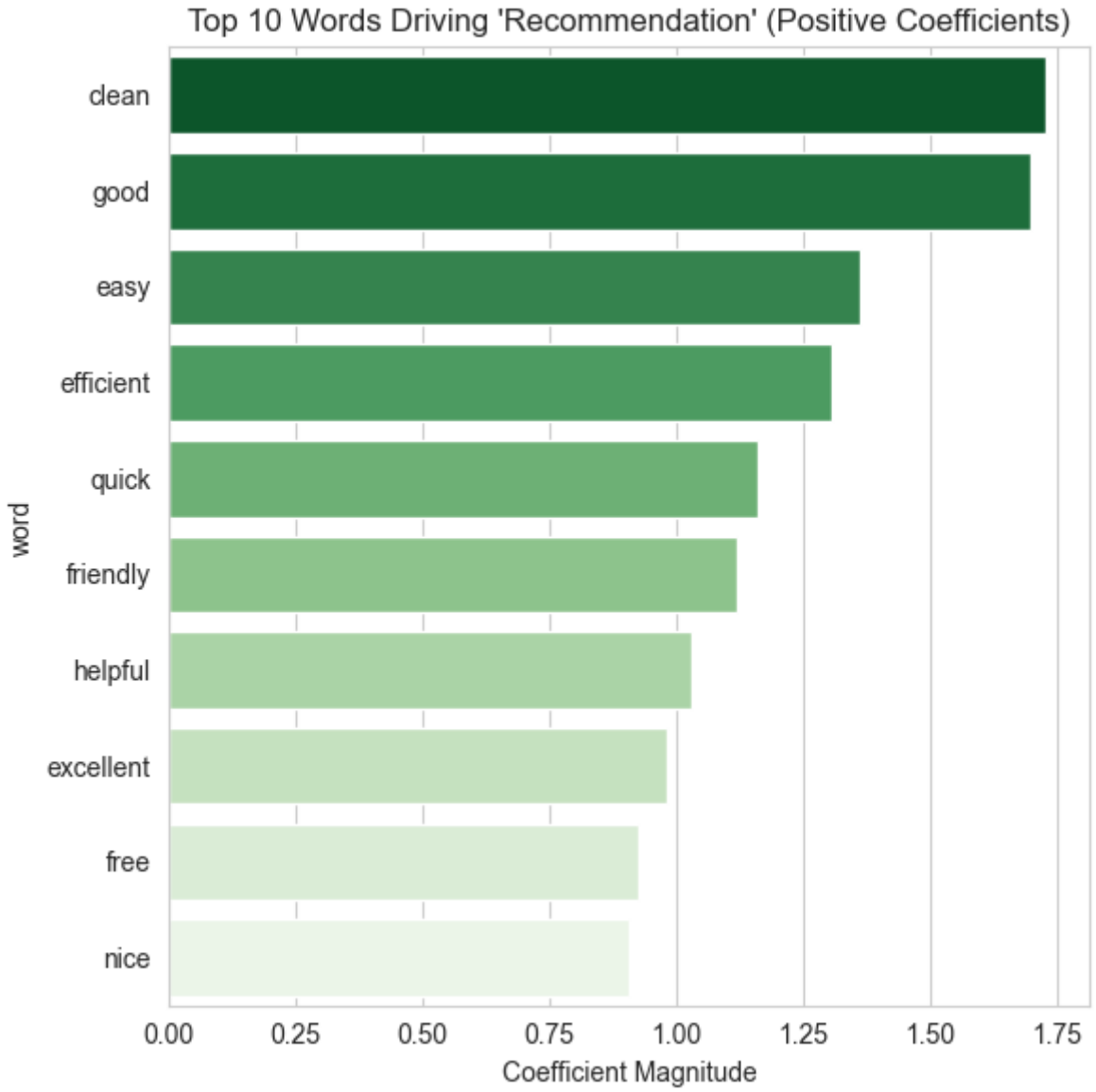
- **Figure 1 (ROC Curve):** Shows how good our log model is at distinguishing between recommends and non-recommends



- **Figure 2 (Positive Drivers):** The top words that boost a rating (ex. "clean", "quick", "friendly")



- **Figure 3 (Negative Drivers):** The top words that tank a rating (ex. "dirty", "rude", "slow")



- **Figure 4 (Rating Drivers):** A chart showing which sub-factors (like queuing vs shopping) actually drive the overall rating in our linear model's predictions