★★★★★

Airbnb Rating Classification

airbnb

By Ji Noh, Minwoo Sohn, and Tiffany Lee

Our Team

Minwoo Sohn

Tiffany Lee

Ji Noh

Agenda

1

Project Overview

Business overview and problem statement.

2

Dataset Overview

Overview of our dataset and EDA

3

Feature Engineering

Deepdive into how we created features

4

Modelling Pipeline

Decision Tree, Random Forest, Logistic Regression

5

Results

The best model and predicted zero review results

6

Conclusion

Key findings, conclusions, and limitations

airbnb

1

Project Overview

Project Overview

airbnb

1

Problem Statement

Lack of reliability for customers to book properties with zero reviews

2

Project Objective

Identify zero review properties' potential value

3

Benefits

Provide proper listing recommendations and mitigate the cold start problem

2

Dataset Overview

Dataset

Kaggle - Inside Airbnb USA Dataset

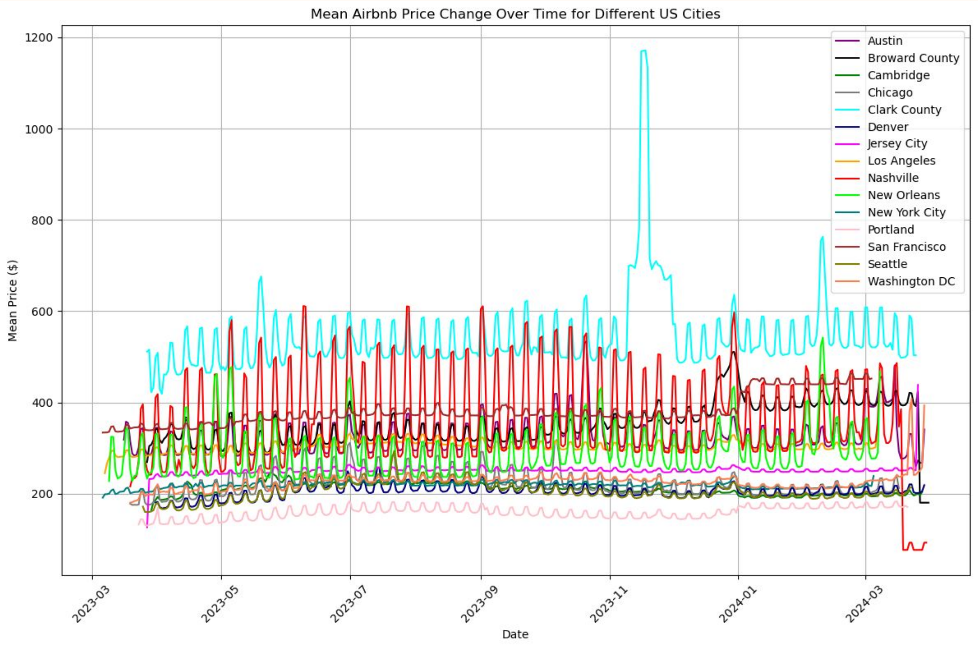
- Entire dataset is 8.94 GB
  - Included 30 U.S. cities worth of data
- Final dataset for project was 15 cities with 184,984 rows with 27 columns
  - Choose Five Cities for Each of Our Three Regions: East, Central, and West
- Narrowed 75 columns to 18 columns before doing feature engineering

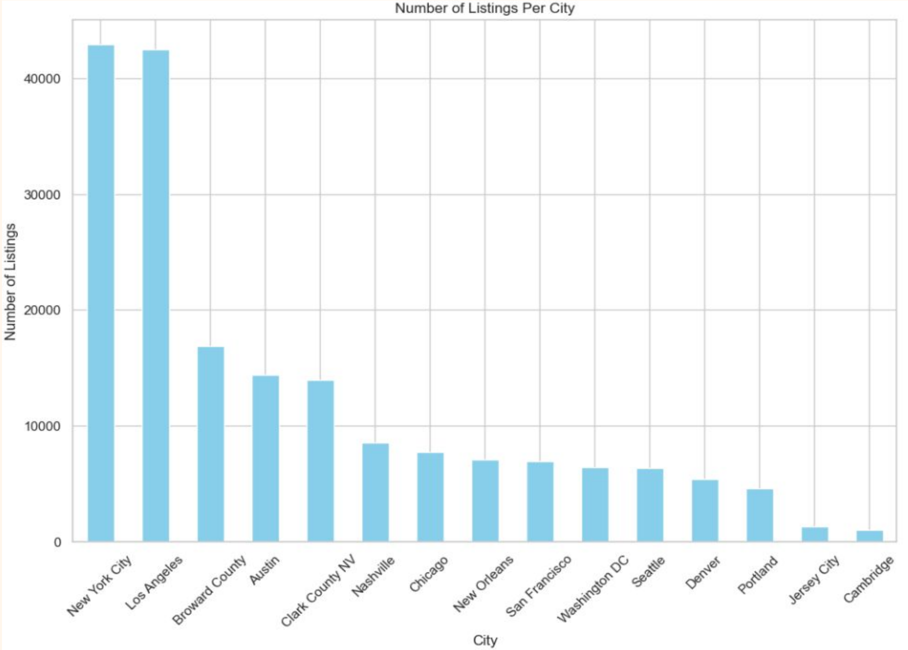
Exploratory Data Analysis

- Conducted EDA for each U.S. region and on the collective dataset encompassing all regions
    - Created plots for time-series, categorical, numerical, and geojson data
    - After exploring seven different tables for each city, we determined to focus on the listings_detailed tables for further EDA understanding.

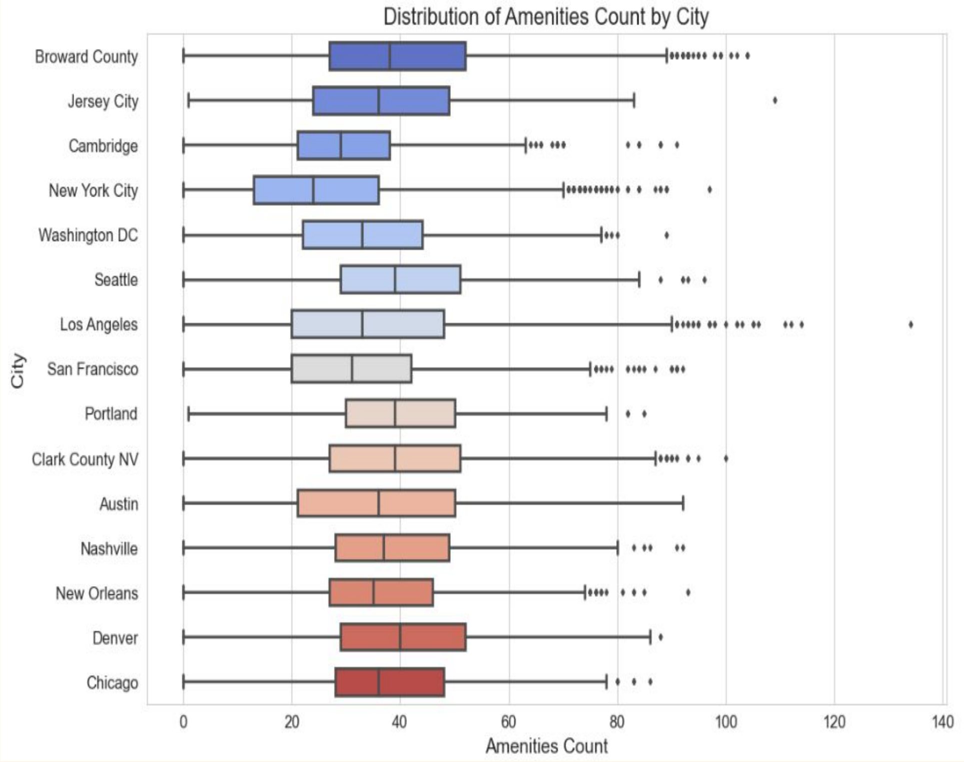
- Important Highlights:
    - Observed patterns in listing prices, with peaks occurring on weekends
    - Big cities such as NY and LA have the most listings and the most listings with 0 reviews
    - The median number of amenities per listing is similar for all cities
    - Despite regional differences, many listings contain the five essential amenities across cities
    - Airbnb experienced rapid business expansion in the mid-2010s through host sign-ups
    - Time-series price data shows seasonality or disruptions from COVID

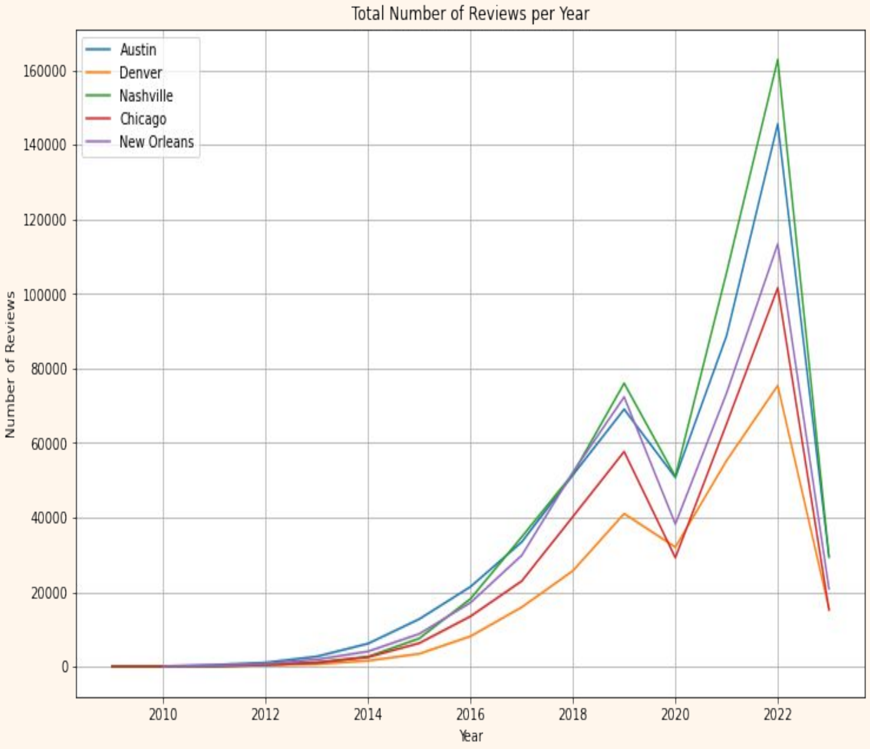
Exploratory Data Analysis





Exploratory Data Analysis





3

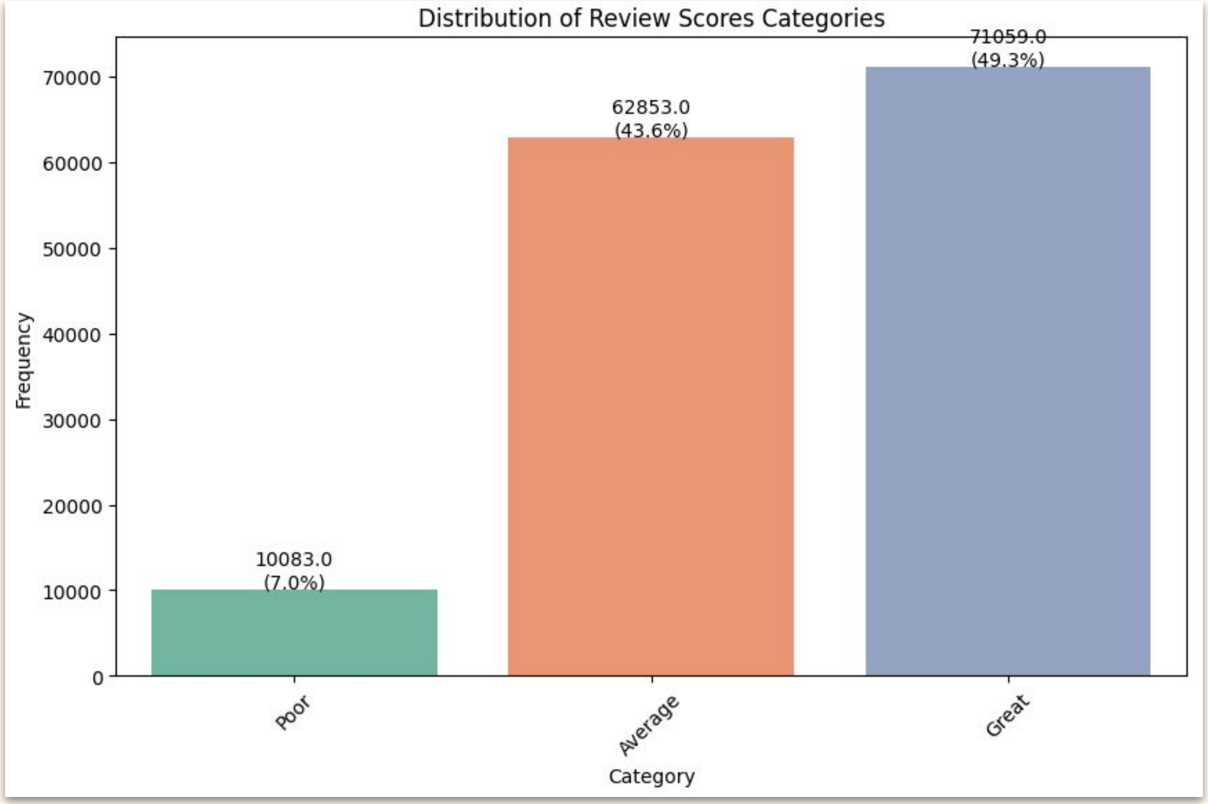
Feature Engineering

Feature Engineering

1	Amenities Count	Count number of amenities
2	Essential Amenities	Fridge, AC, Kitchen, WiFi, Essentials
3	Full-time Host	Hosts with more than 10 listings on Airbnb
4	Host Verification	Encode host_verification column from list object to string
5	Bathroom	Parse out characters to convert it to numeric column
6	Target	0 < Poor <= 4 4 < Average <= 4.8 4.8 < Great <=5

airbnb

Target Variable Distribution



Modeling

4

Modeling Pipeline

Transformations

- **18 features**
  - Categorial: 8 features → StringIndexer
  - Numerical: 10 features
  - Vector Assembler → Single features vector
- **Target outcome**:
  - ‘Poor’ == 0
  - ‘Average’ == 1
  - ‘Great’ == 2

features	target_label
[2000.0,1.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,3.0,8.0,3.0,4.0,6.0,500.0,2.0,5.0,1.0,14.0,3.0]	2.0
[644.0,0.0,1.0,0.0,1.0,0.0,0.0,8.0,1.0,12.0,6.0,2.0,2.0,4.0,186.0,129.0,4.68,3.0,22.0,4.0]	1.0
[262.0,0.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,20.0,7.0,2.0,2.0,5.0,297.0,27.0,4.44,6.0,17.0,3.0]	1.0
[869.0,0.0,1.0,0.0,1.0,0.0,0.0,1.0,1.0,5.0,4.0,1.0,1.0,2.0,162.0,162.0,4.64,5.0,69.0,5.0]	1.0
[1205.0,0.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,17.0,2.0,1.0,1.0,1.0,92.0,36.0,4.83,15.0,17.0,4.0]	0.0
[1747.0,1.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,1930.0,6.0,2.0,2.0,4.0,258.0,35.0,4.71,23.0,38.0,5.0]	0.0
[1205.0,1.0,1.0,1.0,1.0,2.0,0.0,4.0,1.0,1.0,4.0,1.0,1.0,2.0,100.0,156.0,4.89,1.0,72.0,4.0]	0.0
[486.0,0.0,1.0,1.0,1.0,2.0,0.0,8.0,1.0,19.0,4.0,1.0,1.0,2.0,189.0,12.0,4.17,16.0,60.0,5.0]	1.0
[1205.0,1.0,1.0,0.0,1.0,2.0,0.0,1.0,1.0,2.0,3.0,1.0,1.0,1.0,63.0,9.0,4.33,2.0,67.0,5.0]	1.0
[1205.0,0.0,1.0,1.0,1.0,2.0,0.0,1.0,1.0,13.0,2.0,1.0,1.0,1.0,127.0,49.0,4.57,6.0,51.0,5.0]	1.0
[34.0,0.0,1.0,1.0,1.0,2.0,0.0,1.0,1.0,112.0,4.0,2.0,1.0,1.0,300.0,1.0,5.0,5.0,81.0,5.0]	2.0
[1205.0,0.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,14.0,6.0,2.0,2.0,2.0,218.0,83.0,4.7,8.0,16.0,4.0]	0.0
[633.0,0.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,2.0,4.0,1.0,2.0,3.0,155.0,6.0,4.83,1.0,52.0,5.0]	0.0
[1547.0,1.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,32.0,9.0,3.5,4.0,6.0,1764.0,1.0,5.0,4.0,57.0,5.0]	2.0
[1909.0,1.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,10.0,2.0,1.0,1.0,1.0,91.0,205.0,4.78,4.0,52.0,5.0]	0.0
[1205.0,1.0,1.0,0.0,1.0,2.0,0.0,1.0,1.0,2.0,3.0,1.0,1.0,1.0,60.0,3.0,4.67,2.0,66.0,5.0]	1.0
[1909.0,0.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,292.0,6.0,1.0,2.0,3.0,313.0,20.0,3.7,28.0,24.0,4.0]	3.0
[386.0,1.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,72.0,2.0,1.0,1.0,1.0,101.0,17.0,4.94,70.0,54.0,5.0]	0.0
[644.0,1.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,12.0,4.0,2.0,2.0,2.0,236.0,23.0,4.87,10.0,63.0,5.0]	0.0
[644.0,1.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,23.0,6.0,3.0,3.0,3.0,440.0,64.0,4.8,19.0,52.0,5.0]	0.0

Decision Tree

A tree-structured model that represents decisions and their possible consequences, used to predict class labels by learning decision rules inferred from features

Logistic Regression

A predictive analysis model that uses logistic function to estimate the probabilities of multiple class outcomes

Random Forest

An ensemble of decision trees designed to improve classification accuracy by averaging predictions from various trees to determine the class of each input

Models

airbnb

Decision Tree

Baseline (13): 57.54%
After Tuning (13): 58.46%

After Feature Importance
Baseline (7): 57.19%
After Tuning (7): 58.05%

Logistic Regression

Baseline (13): 55.88%
After Tuning (13): 56.00%

After Feature Importance
Baseline (10): 55.51%
After Tuning (10): 55.53%

Random Forest

Baseline (13): 57.95%
After Tuning (13): 60.01%

After Feature Importance
Baseline (7): 58.02%
After Tuning (7): 59.65%

Baseline (10): 57.28%
After Tuning (10): 59.84%

With All Columns
Baseline (18): 50.93%

Model Comparison

05

Results

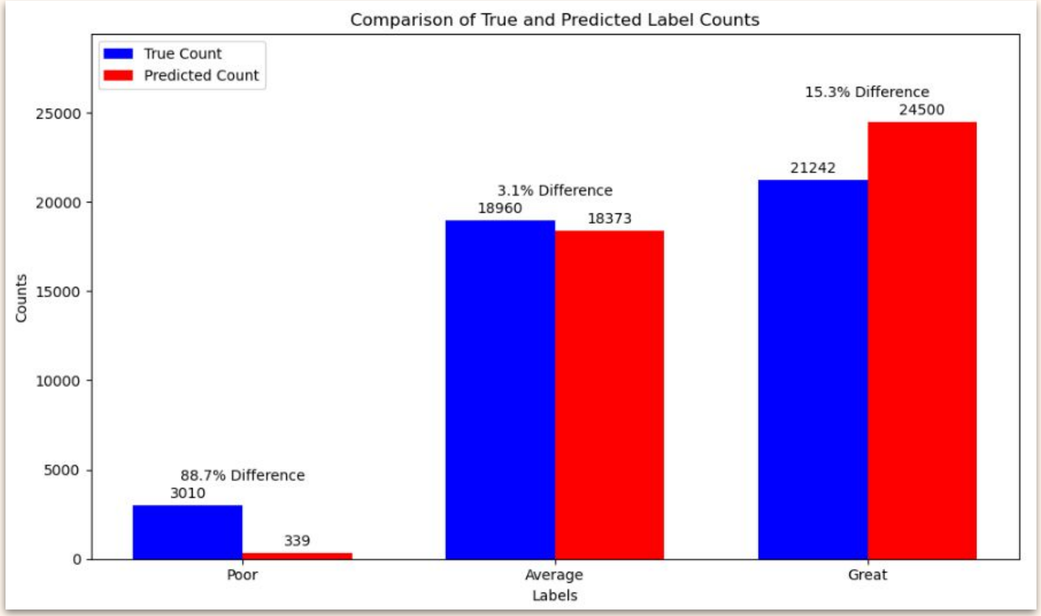
Random Forest Results

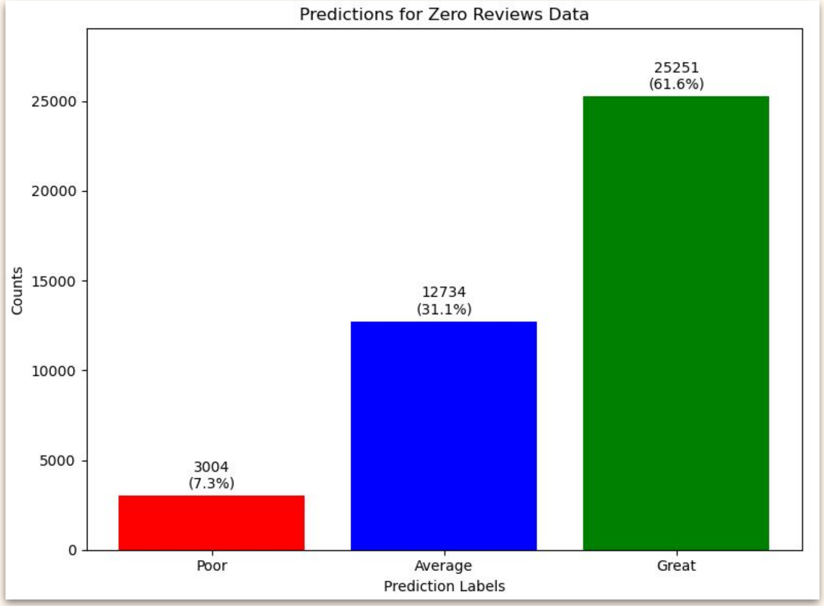
airbnb

Actual \ Predicted	Poor	Average	Great
Poor	166	1,711	1,133
Average	56	10,652	8,252
Great	117	6,010	15,115

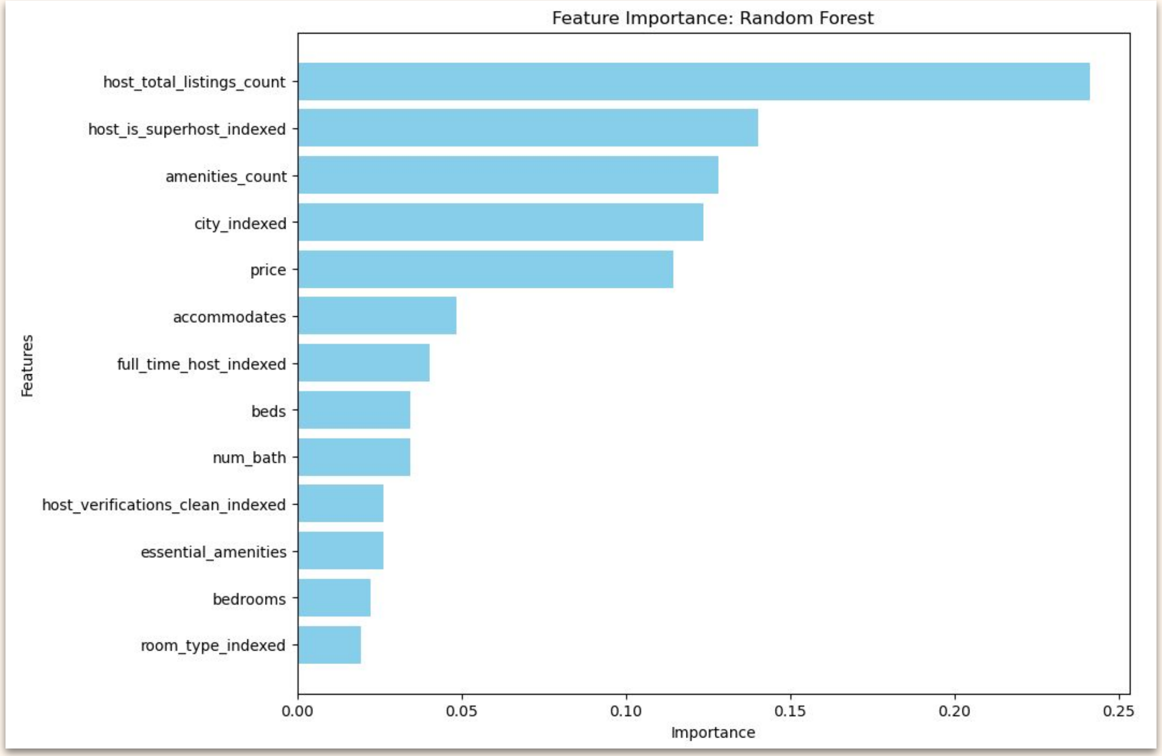
- Avg Precision: 59.18%
- Avg Weighted Recall: 60.01%
- Avg F1-Score: 58.22%

Random Forest Results





Random Forest Feature Importance



Conclusion

06

Conclusion

Model Outcome: 60% accuracy from Fine-tuned RF model classifying 3 categories

Real-world Application:

- **Airbnb Company**:
  - Internal tool for assessing the potential popularity of properties with zero reviews.
  - Guidelines for hosts on adjusting pricing or features to enhance property appeal.
- **Airbnb Customers:**
  - Guidance on evaluating the value of properties lacking reviews.
  - Increased confidence in booking decisions for properties without reviews.

Limitations:

- Lack of computational resources with GCS when fine-tuning (had to cut out 5 columns)
- Limited representative number of cities for each region
- Unbalanced 'Poor' rating data despite the high 4.0 maximum star rating threshold set
- Unaccountable additional regional differences, such as recent region growth
- Unable to see real ground truth on zero reviews dataset as it is for predictive purposes only

Thanks!