CMPS140 Become a Top Airbnb Earner

Tangni Wang, Tung Hoi Man, Yunxiang Fu
University of California Santa Cruz,
twang63@ucsc.edu, tuman@ucsc.edu, yfu7@ucsc.edu

Baskin Engineering
UC SANTA CRUZ

Motivation

Airbnb hosts' earning could vary quite significantly in the same area. We want to know what factors help low performers to get making more profit than the others.

Goals

The goal is to make recommendations for the low earner hosts based on the comparison of different performers' data.

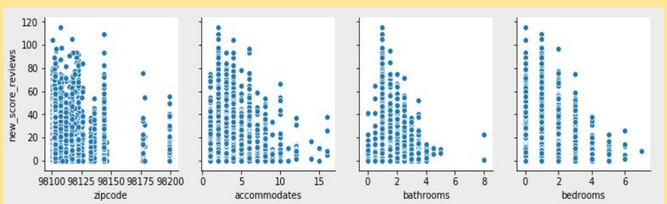
Data

- The data comes from Seattle Airbnb Open Data (https://www.kaggle.com/airbnb/seattle)
- Dataset: 3818 samples
- Features: 92

Data preprocessing:

- Dropped irrelevant features: 75
- Convert string to number
- Convert boolean to 1 / 0
- Fill missing values with 0

Dataset after preprocessing: 3805 samples
Training set: 3044 (80%)
Testing set: 761 (20%)



Methodology

- **Linear Regression**: we use it to predict the target variable (y) based on independent variable (X). So, we can find out a linear relationship between features (X) and target variable (y). Also we can know which feature is more important by looking at the coefficient of each feature.

  Improve model with cross validation (cv): cv can ensure our data split randomly when we split it to train/test set. Here, we performed 9-fold cross validation for linear regression. The score was improved by 20%.

- **Random Forest Regression:** use multiple decision trees in determining the outputs instead of using one decision tree.
  - Reason: very easy to measure the relative importance of each feature on the prediction.

Features

- **Relevant: 17**
  zipcode, accommodates, bathrooms, bedrooms, beds, price, weekly_price, monthly_price, security_deposit, cleaning_fee, extra_people, minimum_nights, host_response_time, host_is_superhost, host_identity_verified, instant_bookable, cancellation_policy

- **Correlate to new_score_reviews: 10**
  accommodates, bathrooms, bedrooms, beds, price, cleaning_fee, host_response_time, host_is_superhost, host_identity_verified, instant_bookable

- **Making a new feature as target variable:**
  new_score_reviews = *(reviews_per_month * review_scores_rating) / 10*

- **Top and low performers threshold:**
  Top performers: new_score_reviews >= 90% quartile (44.11)
  Low performers: new_score_reviews <= 25% quartile (6.48)

Results and Evaluation

Baseline

Model	Features	Metrics				
		r2	ev	mae	mse	med_ae
Dummy Regressor (median)	Relevant (17)	-0.12	0	13.02	353.18	9.94

Stretch Model

Model	Features	Metrics				
		r2	ev	mae	mse	med_ae
Linear Regression	Relevant (17)	0.322	0.322	10.66	213.63	8.21
Linear Regression with cv	Relevant (17)	0.392	0.315	10.8	215.84	8.2
Lasso Regression	Relevant (17)	0.318	0.318	10.70	215.05	8.25
Bayesian Ridge Regression	Relevant (17)	0.321	0.321	10.68	214.17	8.24
Random Forest Regression	Relevant (17)	0.368	0.368	9.99	199.31	7.21

Conclusion

Recommendations for low performers:
Based on coefficient of each feature from linear regression model and feature importances in random forest regression. We conclude the following 9 features are crucial to host's earning:

zipcode, bedrooms, price, cleaning_fee, host_response_time, host_is_superhost, host_identity_verified, instant_bookable, cancellation_policy

Discussion:

- Is the new feature (new_score_reviews) a good indicator for host's earning?
- Future work: Incorporate top and low performers threshold to build models.

Reference

- Project repo: https://github.com/tonyman316/airbnb
- Kaggle: https://www.kaggle.com/yogi045/how-to-become-top-earner-in-airbnb/
- Scikit-learn: https://scikit-learn.org