



AirBnb Ratings Classification

Matthew Zhang

December 2020



Overview

AirBnb success is often determined by customer experience. I hope to gain valuable insights to form strategy and demonstrate success.

In this case, I will be a company analyst looking to help AirBnb hosts enhance overall customer satisfaction by classifying listings into three categories: Subpar(0), Good(1) and Best(2); further, to ascertain which aspects of properties can be improved.



Understanding the Data

Source:

Inside AirBnb data set
from insideairbnb.com

Specifically, the Los
Angeles location using the
listings..csv.gz file.

Details:

The data set contains
around 31,000 rows
(properties) and 74
columns.

After cleaning, around
29,000 rows and 36
columns

Target:

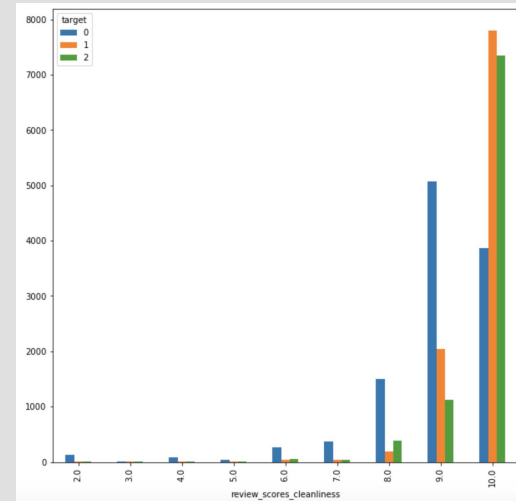
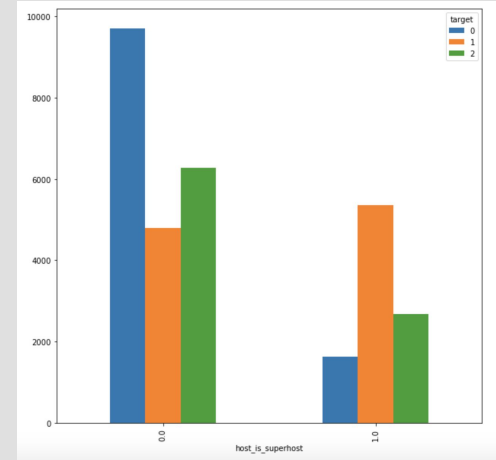
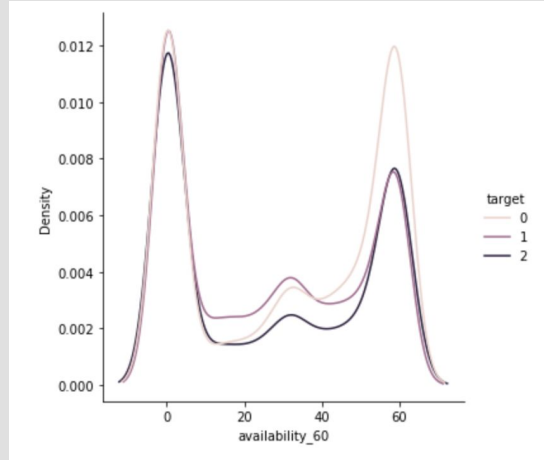
3 classes based on rating.
These classes were named
Subpar(0), Good(1) and
Best(2).

```
Out[130]: 0    0.370931  
          1    0.334184  
          2    0.294884  
          Name: target, dtype: float64
```

```
Out[55]: 0    11662  
          1    10479  
          2     9324  
          Name: target, dtype: int64
```

EDA

- Visualized and explored descriptive statistics
- Examined spreads of classes across each predictor
- Researched features, and processed accordingly
- Created dummy variables for relevant categoricals



Intermediate Model: Decision Tree Classifier

Vanilla Model: Perfect training metrics but accuracy of .60 for testing.

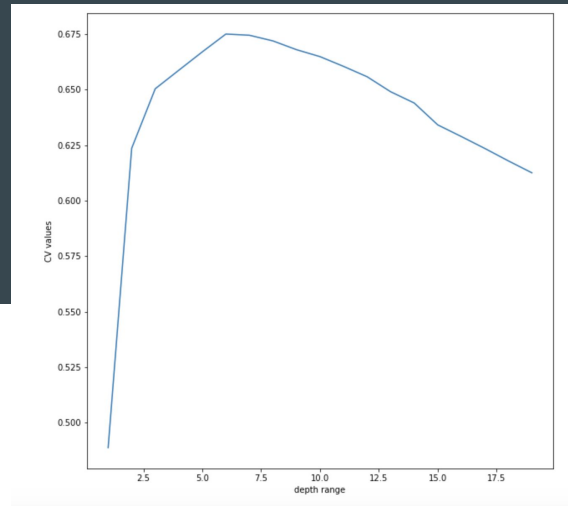
Max_depth and GridSearch CV optimal parameters:

Max_depth = 6, criterion = 'gini',
min_samples_split = 10

Improved model: .68 accuracy

	precision	recall	f1-score
0	0.70	0.74	0.72
1	0.64	0.71	0.67
2	0.76	0.60	0.67
accuracy			0.69
macro avg	0.70	0.68	0.69
weighted avg	0.69	0.69	0.69

	precision	recall	f1-score	support
0	0.68	0.72	0.70	2722
1	0.63	0.71	0.67	2504
2	0.74	0.59	0.66	2217
accuracy			0.68	7443
macro avg	0.68	0.67	0.67	7443
weighted avg	0.68	0.68	0.68	7443



Baseline Model: Random Forest Classifier

Without any tuning and running a vanilla model

Again, perfect training set metrics with an accuracy of 1.00

However, test set has an accuracy of .68

What I tuned: `n_estimators`, `max_depth`, `min_samples_leaf`, `min_samples_split`, `criterion` and `max_features`

	precision	recall	f1-score	support
0	1.00	1.00	1.00	8321
1	1.00	1.00	1.00	7445
2	1.00	1.00	1.00	6562
accuracy			1.00	22328
macro avg	1.00	1.00	1.00	22328
weighted avg	1.00	1.00	1.00	22328

	precision	recall	f1-score	support
0	0.70	0.71	0.70	2722
1	0.68	0.67	0.67	2504
2	0.66	0.65	0.66	2217
accuracy			0.68	7443
macro avg	0.68	0.68	0.68	7443
weighted avg	0.68	0.68	0.68	7443

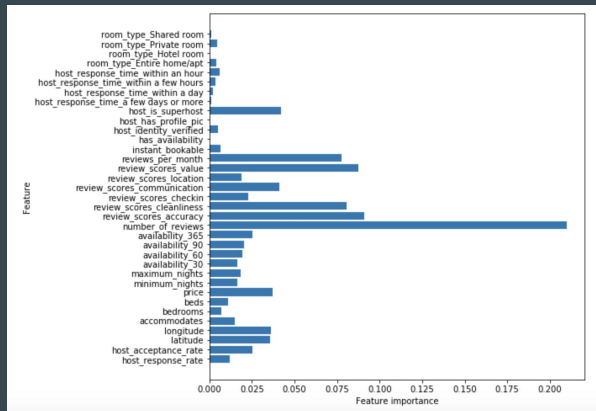
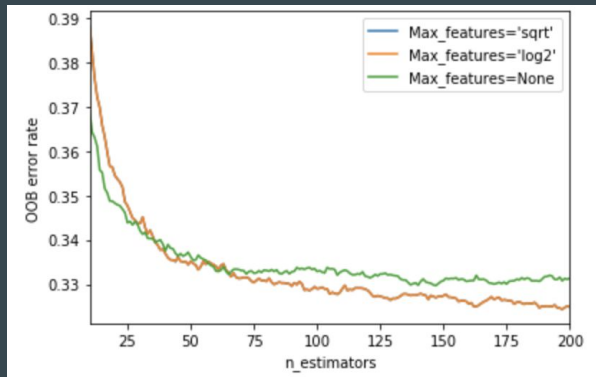
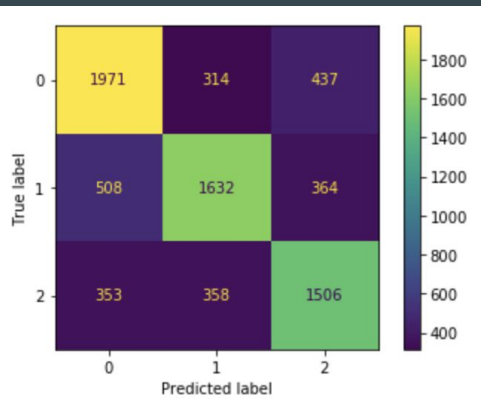
Final Model: Random Forest Classifier

After GridSearchCV and Out-Of-Bag comparison, final model had an accuracy and averages of .70 .

Confusion Matrix: Majority predicted correctly and less than half predicted incorrectly.

Illustration of features importances

Similar to decision trees, but the levels of importance changed as a result of ensemble of multiple decision trees considering different features



Conclusions

Effectively classify Airbnb listings according to each class Subpar(0), Good(1), Best(2)

Especially useful for predicting listings with not a lot of reviews

Recommendations:

- Properties should be cleaned thoroughly
 - Consider being an Airbnb superhost
 - Improve communication
-



Future Work

- Explore different types of property and how they affects a customer's visit
- Analysis of different cities provided by the Inside AirBnb website.



Thank You

GitHub Repository

- <https://github.com/mzcode98/airbnb-rating-classification>

Email

- mattzhang989@gmail.com

LinkedIn

- <https://www.linkedin.com/in/matthew-zhang-167859183/>

Acknowledgements

- Flatiron School, Instructor Yish Lim, Classmates