Saida Huseyn
saida.huseyn@outlook.com
https://github.com/SaidaHuss
July, 2022 Brain Station: Diploma in Data Science Final Project

Short-term rental analysis using New York City Airbnb rental example

What makes a super host?

The project aims to help short-term rental hosts optimize their performance, increase income and understand why some hosts succeed more on short-term rental platforms than others. For Airbnb NY City hosts specifically, hosts with the super-host label get three times more reviews (proxy to bookings) than non-super hosts. Many projects using this dataset targeted price prediction, but almost none focused on host performance. From the EDA in this project, we can see that the price does not have a linear relationship with any of the factors; thus, linear regression performs poorly. On the other hand, this project focuses on identifying classification patterns and explaining what makes some hosts super or successful on that platforms. Airbnb/Nyu York City dataset has a good variety of features (74 columns) and a decent number of listings (38k) to apply ML models.

The Data

The data is downloaded from the Insideairbnb website and placed here. It has the shape of (37631, 74). The dataset is a mix of numeric, categorical and text data.

The below table describes the data in summary:

	Original Data		Final Data	
Topic	**N of Features**	**Dtype**	**N of Features**	**Dtype**
Host	23	object/number	5	int/bool
Location	6	object/number	7	float/int/bool
Property	17	object/number	19	float/int/bool
Price	1	float	1	float
Reviews	13	float/int	10	float/int
Availability	14	object/number	8	float/int/bool
Total	74		50	

Data summary description:

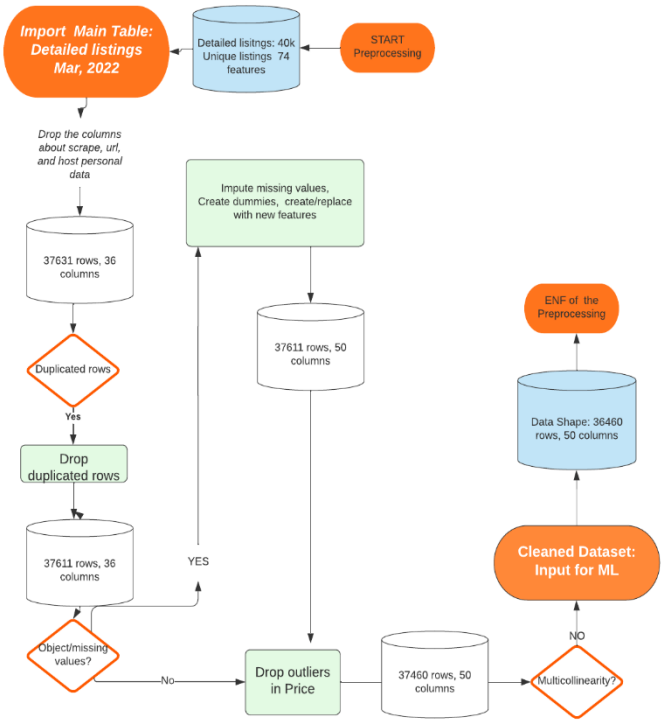
- Each data point is a single listing active on Airbnb NYC for the date of data scraping.
- For the analysis, two columns are used as the target of interest: Price and host_is_superhost.

Saida Huseyn
saida.huseyn@outlook.com
https://github.com/SaidaHuss
July, 2022 Brain Station: Diploma in Data Science Final Project

Data Cleaning:

The main challenge of the given data is that it has many missing values and text data, and some of the features are either irrelevant or can be replaced with a better representative of the same quality. Price outliers were also removed.

- The following techniques were used to fill the missing values: mean, the most representative value, median, and KNN imputer.
- A few duplicates were removed from rows, and duplicated information in columns (repetitive with different names) was dropped.



EDA Results:

Main observations for price:

- Price is not linearly related to any of the variables

- Price is mainly related to the location, size of the property, type of the property, and amenities.

- Independent variables don't have multicollinearity; however, for dummies classes, one class can be dropped as a reference before modelling.

Main observations for Superhost:

- Being a super host is not related to price or to being active for a long time in the market

- Superhosts tend to have more amenities, more number of reviews, higher review scores

Data Preprocessing:

- Before any modelling, the data was split twice, creating train, validation and test subsets.
- After the split, the training set was scaled using StandardScaler, and missing values were imputed using KNN; all steps fit in the training set and transformed both validation and test sets.

Saida Huseyn
saida.huseyn@outlook.com
https://github.com/SaidaHuss

Modelling Results:

	Logistic Regression	KNN	Decision Tree	Random Forest	Bagging:DTs
Hyperparameters	C=100	k.n=16	depth=7	n.estimators=100	n.estimators=100
Accuracy	83	84	86	88	88
Precision	68	69	71	78	77
Recall	36	37	58	55	61
F1-score	47	48	64	65	68

Supervised Classification:

- Four basic models were applied and optimized for this project; Logistic Regression, KNN, Decision Tree and Random Forest.
- The best performing model is Random Forest, based on combined features. However, a single decision tree results in a similar performance.

Based on our Random Forest model **feature importance estimations:**

The number of reviews and number of **reviews in the recent 12 months** is the best predictors of a host being a super host

Followed by **review_scores_rating**, a combined rating based on many qualities of the property, location and host service. Specifically, **review_scores_cleaneliness** stands out. The initial observation is; that people also write the most about cleanliness in their text reviews. Further NLP analysis is needed to understand the main topics discussed in reviews and even use only those parameters to predict a super host.

The **number of amenities** plays the main role in pricing and the host becoming a super host. This is more significant than a specific type of amenity.

Hosts being longer in the market and having more **available days** also influences performance positively.

Host listings count seems to be an influential feature, although this feature is one that customers don't care about. Probably Airbnb weights host with more listings positively.

General Conclusion and Business implications:

- Considering that the label "super-host" is not automatically generated on the given features but rather assigned by Airbnb, it was an interesting topic to analyze and see how Airbnb's labels are aligned with customers' perceptions and demands of a listing. The result shows that classifiers struggle to predict the classes; however, after some optimization, performance is improved.

- Even with poor prediction capacity, models like Random forest can give us reasonable estimations on what hosts can improve to become super hosts. Surprisingly, performance and price are not strictly related to a place's location or size. Super hosts have many opportunities to improve the experience for guests:
    - Improve cleanliness and accuracy in the given information
    - Improve, and increase the number of amenities
    - Motivate customers to leave reviews: super hosts get three times more reviews than others. It might be related to the fact that they are rented more, but also they might have better communication with guests, and thus, motivate them for better/more reviews.