Pricing Airbnb Apartments in Sicily, Italy

Introduction and Goal

This report investigates and explores the predictive models for forecasting Airbnb apartment prices in the top most populated and touristic cities in Sicily, Italy. The findings and models in this report will provide significant insights and help our client to price and operate small and middle-sized apartments hosting 2-6 guests. From the results, we will analyze and recommend the best model the client should use to price their company apartments. Since our client does not provide detailed information about the exact city and the rental purpose, this report will assume and focus on the most popular use case: leisure rentals in famous touristic cities in Sicily.

The code and both reports can be found on this GitHub.

---

1. Data

Data Preparation

The data in the report is the listing details of Sicily Airbnb apartments in March 2023 from http://insideairbnb.com/get-the-data/. Further data cleaning and filtering are applied to fit with our scope of analysis. The detailed steps are as follows:

1. Get the raw dataset (N = 51,679)
2. Apply filters (N = 16,240):
    - Drop observations with missing value in price, number of beds, number of bathrooms, the apartment is instant bookable.
    - Exclude apartment type 'Hotel' as our client is not operating hotel rooms.
    - Only keep apartments accommodating for 2-6 guests.
    - Only keep apartments located in Palermo, Catania, Messina, Taormina, Realmonte, Agrigento, Siracusa, Cefal, Monreale, Ragusa and Modica.
    - Exclude some extreme price values (top 1%) by keeping observation with price <= $400.
3. Feature engineering (N = 16,240):
    - Fill 0s as values for observations with no rating values.
    - Fill 1s as values for observations with no bedrooms values as these are studios.
    - Fill 'NAN' as values for observations with no license values.
    - Create boolean values for super host, the host has a profile picture, host identity verified, the apartment is instantly bookable and the host has a license.
    - Convert listing duration from text to number of days before 31/03/2023.
    - Convert price from text to numerical values.
    - Derive the number of bathrooms from the bathroom text descriptions.
    - Clean the property type to remove the redundant values of room type.
    - Derive various boolean variables for amenities from the amenities text descriptions.

From all the data features in the raw dataset, we will use a subset of those as predictors with the following rationales:

- Profile photo, verification status, license and the number of listings may influence the price as a host without those can price be lower to attract guests.
- Room details (location, type of property and room, number of accommodations, bedrooms, bathrooms, reviews, amenities, instant bookable, number of allowed nights) are chosen as rooms with higher value in any of those features can demand higher prices from intuition.
- Other features are excluded as they are presented in other predictors (like longitude) or are irrelevant to our scope.

Data Exploration

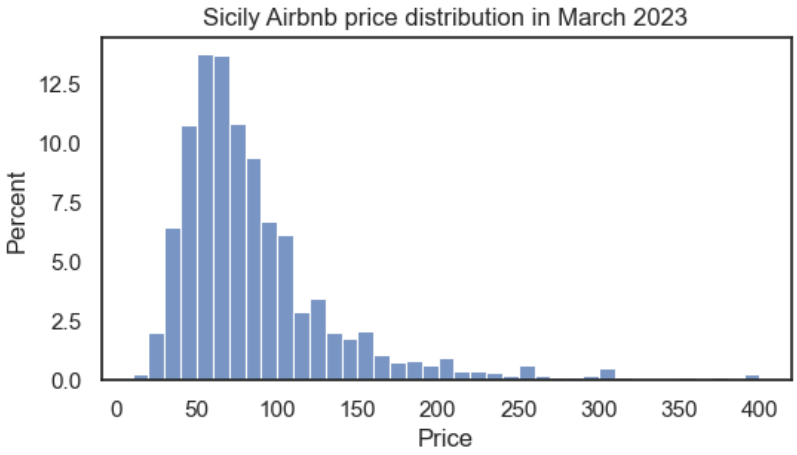

Fig. 1: Sicily Airbnb price distribution

From Fig. 1, it is clear that the price distribution is left-skewed. This means that the mean value price is not the typical value. Within this paper, the OLS models will assume that the mean is typical and will predict the mean value. It is worth noting that normally this is not the case and there might be impacts on the prediction results made by the OLS models.

2. Models

In this report, we will build 4 different models:

- Simple OLS model (OLS)
- Linear model with LASSO (LASSO)
- Random forest model (RF)
- Boosting with gradient descent model (GBM)

All models use the same set of features as discussed in the **Data** section, with the LASSO using all possible combinations of interaction terms. Since LASSO shrinks the coefficients for us and we do not know the interaction among the predictors from intuition, including all the interaction terms and letting the algorithm run will leave us with a prediction model with the optimal choice of interactions through cross-validation. Furthermore, all models, except for the simple OLS, will go through a 5-fold cross-validation with different tuning parameters to optimize for overfitting.

The data used to train and validate is the same across all models. In this report, we will split the working data (N = 16,240) into the training set and the holdout set with a ratio of 7:3.

Simple OLS

This model is a linear model with predictors as the features selected in the **Data** section. No interaction terms are included

LASSO

This linear model has predictors like the Simple OLS models but includes all possible combinations of interaction terms. The tuning parameter is the penalty term α in the range of [0.1, 0.5] inclusive, incremented by 0.05.

Random Forest

The random forest model includes all the selected features with no interaction as the random forest algorithm already handles the interactions among predictors. The tuning parameters are:

- Maximum number of features for each split: [8, 10, 12]
- Minimum number of samples in each terminal node: [5, 10, 15]

Boosting with Gradient Descent

The boosting model includes features similar to the random forest model. The tuning parameters are:

- Learning rate: 0.01
- Number of estimators: [200, 300, 500],
- Maximum tree depth: [5, 10, 15],
- Maximum number of features for each split: [5, 10, 15],
- Maximum number of features for each split: [8, 10, 12],
- Minimum number of samples to continue splitting: [10, 20, 30]

---

3. Model evaluation

Performances

The performance result of each model are as follows:

Out[33]:

	Model	Train RMSE	Holdout RMSE	Training time
0	Simple OLS	44.3212	45.0548	0m0s
1	LASSO	43.8334	44.6018	4m28s
2	Random Forest	42.9022	43.2020	0m23s
3	GBM	40.1745	40.6951	7m36s

The simple OLS model RMSE is the worst of all models, but it is not significantly worse than the LASSO model. Considering that the simple OLS completes almost instantly compared to the LASSO, the simple OLS might be better if speed is the top priority. However, the Random Forest model gives better results than the simple OLS while only taking under half a minute to complete might be a more balanced choice between accuracy and speed. The GBM model has the best accuracy and is significantly better than the rest of the models, but it also takes significantly more time to finish. If accuracy is the top priority, the GBM model is the best model to go forward. The LASSO, Random Forest and GBM models are cross-validated to prevent overfitting and they perform relatively well with the holdout set compared to that over the training set. The same can be said for the simple OLS model while it has not been cross-validated.

Diagnostic

From our analysis of the OLS and LASSO models' coefficients (discussed in more details in the **Model evaluation** in the technical report), these variables and their interactions consistently have significant power to the change in price prediction:

- Property type (f_property_type)
- Neighbourhood (f_neighbourhood_cleansed)
- Room type (f_room_type)
- Has pool (d_pool)
- Number of bathrooms (n_bathrooms)
- Number of bedrooms (n_bedrooms)
- Number of accommodates (n_accommodates)

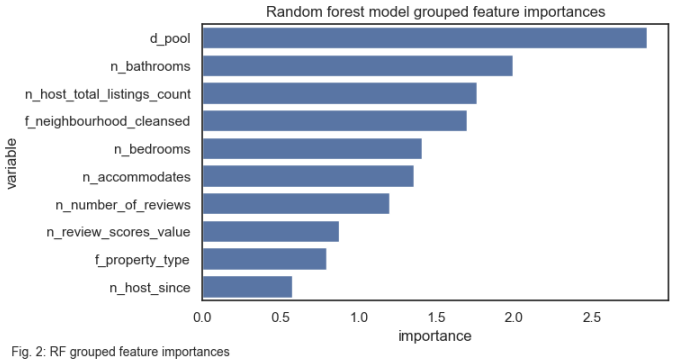

Fig. 2: RF grouped feature importances

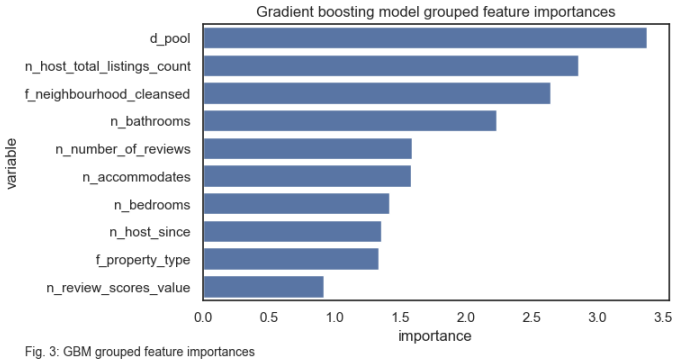

Fig. 3: GBM grouped feature importances

Figure 2 and Figure 3 also show the important features that impact the price prediction in the Random Forest and GBM models. The significant terms in the simple OLS and LASSO models are also important here. However, the Random Forest and GBM models have a few extra important predictors:

- Total number of listings that the host has (n_host_total_listings_count)
- Number of reviews (n_number_of_reviews)
- Review score (n_review_scores_value)
- Host operating duration (n_host_since)

From the feature importance analysis above, our client should provide the key information accurately for their apartment to get a good price prediction. Since our client's apartments are all new, the number of reviews and review scores may not be available. Thus, both of these values should be set at 0.

---

4. Conclusion

The choice of model is pivotal to support our client price their new apartments, as each offers unique advantages. In this report, the simple OLS and Random Forest models offer simplicity and speed, while the LASSO and GBM models offer accuracy with a trade-off in complexity. Among the evaluated models—Simple OLS, LASSO, Random Forest, and GBM—the GBM model excelled in accuracy over the rest. However, due to its computational demands, we might want to recommend the Random Forest to be the model of choice as it balances the complexity and usability.