NEW YORK CITY AIRBNB HOSTS RATING PREDICTION

Instructor: Prof. Ying Lin

Group5: Xinxia Song, Yunxia Zhao, Dahai Liu

Problems & Motivations

Since 2008 people have used Airbnb to travel in a more personalized way, and they wish to have some unique experience. Also, it is increasingly common that guests engage in Aibnb reviews.
For hosts of Airbnb, those data are voice of their customers. To interpret those data can help hosts to improve the service and then improve their review scores. However, there are so many features and hosts have no idea which features matter.

Objectives

- Create different models and tune the parameters to find the best model which has lowest RMSE.
- Based on the results of the best model, we'd like to give hosts some suggestions to manage their house better.

Data Description & Data Preprocessing

- **Raw Dataset**: 48377*106
- Date Complied: 12 September,2019
- Data Source: http://insideairbnb.com/get-the-data.html
- **Target Variable**: review_scores_value
- Remove irrelevant columns and columns that have more than 5% NAs
- Remove duplicate columns (e.g. accommodates VS guests_included)
- Remove columns that have same values for all rows
- Remove NAs in rows
- Regroup values in columns that have so many categories like property type, room type, etc.
- Create dummy variables for columns like property type, room type, etc.
- Winsorize outliers
- Standardize data in different scales
- **Dataset after cleaning**: 37680*30

Model Description & Parameter Tuning

Regression: To estimate the relationships between a dependent variable and one or more independent variables.
Models: Linear Regression (OLS Regression and Ridge Regression), Rand Forest Regression, and Gradient Boosting Regression.

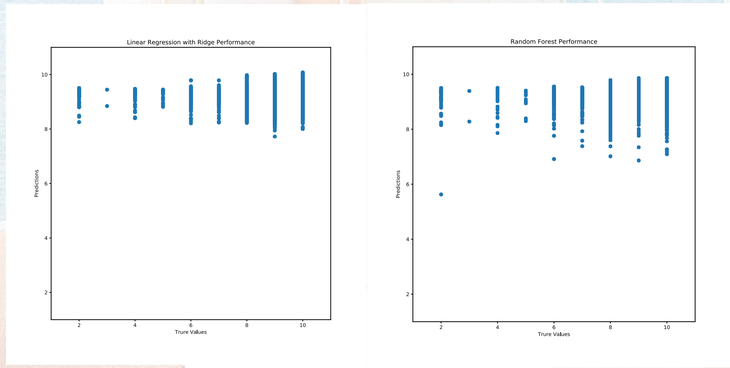
Model	Parameters we tried	Best Parameter
Ridge Regression	alpha: [0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 5, 10, 20]	alpha = 10
Rand Forest Regression	bootstrap: [True], n_estimators: [32, 64, 100, 200], max_depth: [2, 5, 10], max_features: [3, 5, 10]	bootstrap = True, max_depth = 10, max_features = 5, n_estimators = 200
Gradient Boosting Regression	learning_rate': [0.1, 0.5], n_estimators': [100, 200], max_depth': [2, 5], max_features': [3, 5]	learning_rate = 0.1, max_depth = 5, max_features = 3, n_estimators = 100

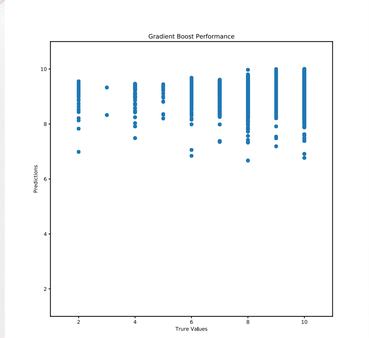
Model Comparison Metrics

In this part, we choose RMSE(Root Mean Squared Error) to evaluate the performance of each model, and then find the best model with lowest RMSE. The results are as follows. As we can see, **Rand Forest Regression** is our best model after tuning.

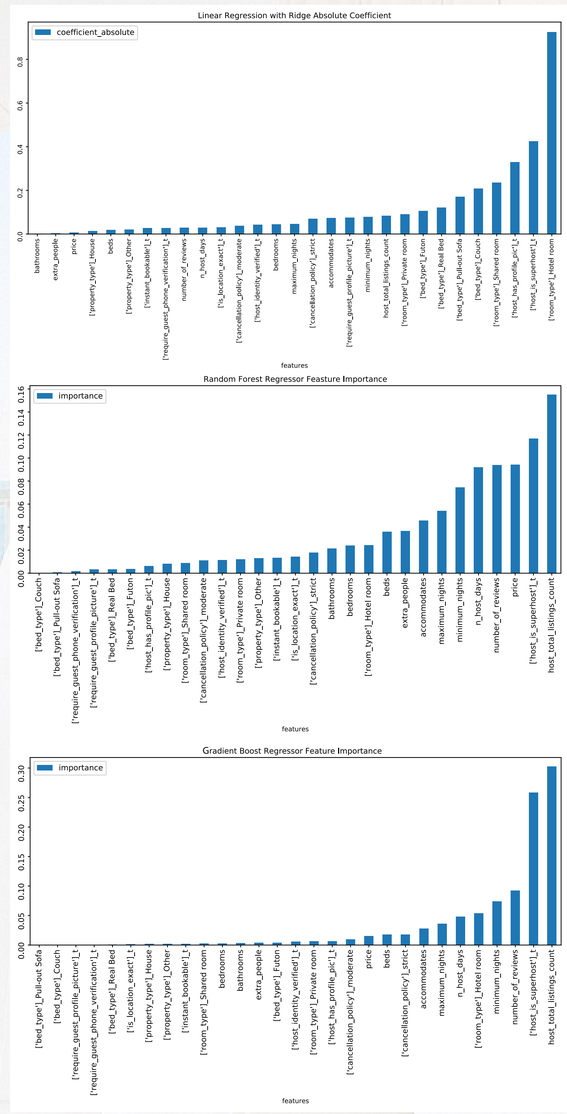
Model	RMSE (before tuning)	RMSE (after tuning)
OLS Regression	0.906100	
Ridge Regression	0.906100	0.905976
Rand Forest Regression	0.904948	**0.892957**
Gradient Boosting Regression	0.895410	0.894496

Model Performance Visualization





Features Importance



Conclusion

- The Random Forest Regression Model with parameters [bootstrap = True, max-_depth = 10, max_features = 5, n_estimators = 200] is the best model which has the lowest RMSE 0.892957.
- Based on the results of feature importance, we suggest that Airbnb hosts should focus on features like 'total listings counts', 'being a superhost', since those features have higher weight.