Airbnb Homestay Prices Predictions

Jiaqin Wu

04.27.2023

Outline

- Research Questions and Background
- Data Source, Target, and Important Features
- Parametric and Non-parametric Techniques Used
- Prediction Performance
- Conclusion and limitations

Research Questions

- Predict Airbnb homestay prices according to the potential factors including both institutional characteristics and demographic information.

Research Background

Airbnb has become one of the most popular online marketplaces, providing people with a diverse range of options for short-term homestays and experiences. By understanding these factors which may influence Airbnb homestay prices, policymakers can make informed decisions to promote sustainability and equity in the tourism industry, including the development of targeted policies to maximize the benefits of Airbnb homestays for both hosts and guests. Additionally, policymakers can gain insights into the potential impact of Airbnb on the local housing market, allowing them to address any negative consequences, such as housing shortages or rising rents.

Data Source

- **20 csv Data Sets** (from Kaggle)
  - The data sets were divided by 10 different countries and week times (weekends/weekdays)
  - Contains institutional characteristics of homestay
  - Concatenate all 20 CSV files into a final dataset, and add two new columns: 'city' and 'week_time'
  - Website: https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities
- **Housing_price.csv Data Set** (from Numbeo)
  - Contains the average price per square meter to buy apartments in each city
  - Website: https://www.numbeo.com/cost-of-living/region_prices_by_city?itemId=100&region=150
- **imf-df-export.csv Data Set** (from IMF)
  - Contains GDP per capita in each country
  - Website: https://www.imf.org/external/datamapper/NGDPDPC@WEO/UVK/EURO/EU

Target and Important Features

Transform the categorical variables into dummy variables and use them in the model. Also, include 'price_m2' and 'GDP_capita' as continuous variables to represent the city difference.

- Target variable: 'realSum'
- Important feature variables: the maximum number of guests ('person_capacity'), ownership of the listing by hosts ('multi', 'biz'), cleanliness rating ('cleanliness_rating'), guest satisfaction rate ('guest_satisfaction_overall'), number of bedrooms ('bedrooms'), distance from the city center ('dist'), distance from the metro station ('metro_dist'), attraction index ('attr_index_norm'), restaurant index ('rest_index_norm'), housing price per squared meter ('price_m2'), GDP per capita ('GDP_capita'), room type ('room_type_Private room', 'room_type_Shared room'), type of the host ('host_is_superhost_True'), and week time ('week_time_weekends')

Parametric and Non-parametric Techniques Used

- Parametric Techniques:
  - Linear Regression Model
  - Ridge
  - Lasso
- Non-parametric Techniques:
  - XGBoost Model
  - Decision Tree Model
  - Random Forest Model

Performance of different models

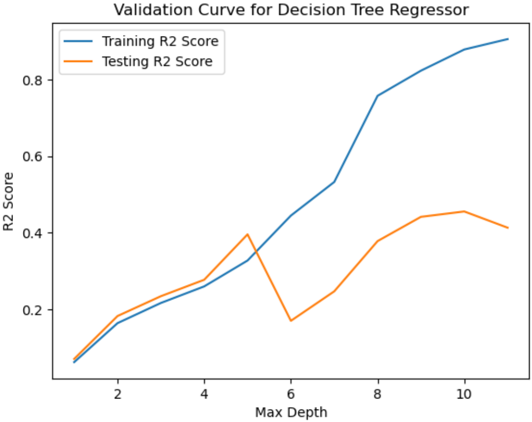
Model	Training RMSE	Training R2 Score	Test RMSE	Test R2 Score	RMSE (Cross-Validation)
XGBoost	196.49	0.65	197.52	0.60	286.14
Random Forest	256.24	0.40	250.27	0.36	286.47
Decision Tree	271.77	0.33	244.06	0.40	294.98
Ridge	292.79	0.22	275.19	0.23	321.30
Lasso	292.79	0.22	275.19	0.23	325.87
LinearRegression	292.79	0.22	275.19	0.23	325.87

The choice of alpha in Ridge and Lasso Model

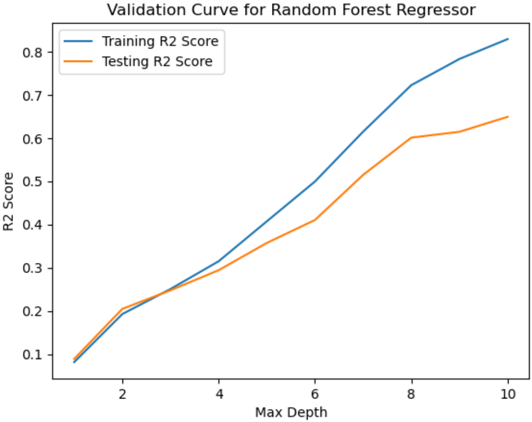
- Choose alpha value from 0.0001 to 100000 to explore the best performance of the linear regression model
- For Ridge, the best alpha is 100
- For Lasso, the best alpha is 0.0001

There are nearly no differences between the linear regression model, Ridge and Lasso models.

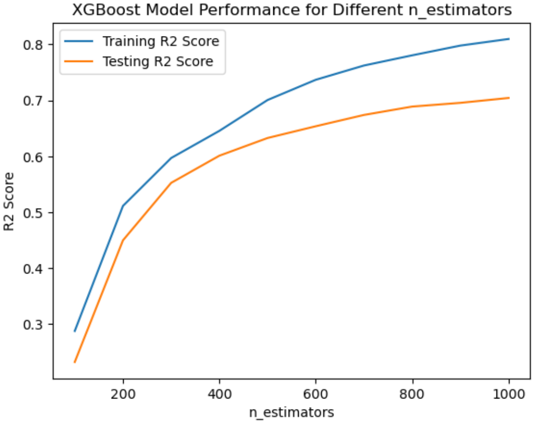
The choice of max depth in Decision Tree model



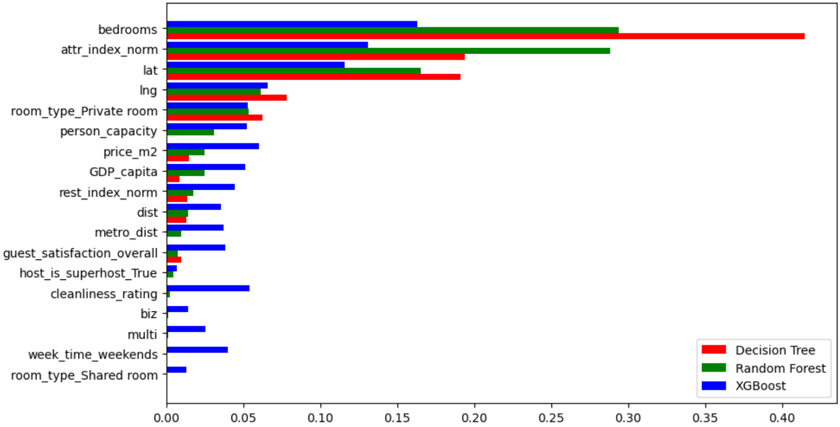
The choice of max depth in Random Forest model



The choice of estimators in XGBoost model



The comparison of feature importance



Conclusion

- My current model achieved around 60% prediction accuracy, which means the complex nature of homestay pricing requires further refinement of the models and the original datasets.

Limitations

- More factors need to take into consideration, such as the impact of COVID-19, the inflation of the Euro, and the criminal rate in each city.
    - The lack of more institutional variables in the original dataset such as the room area and facility rate of the homestay, as these factors also have an impact on homestay prices.
    - The lack of more specific demographic information in each city. For instance, the GDP and housing price conditions can vary in different zones within the same city, and this information could help provide more insights into the relationship between cities and homestay prices.
- More models could be taken into consideration, such as SVM, and neural network models. We can also try to use grid search to further refine the models.