Exploratory Data Analysis and Price Prediction of Airbnb in New York City

Jiani Lyu, Chelsea Liu, Yuewei Shi, Yanchi Jin, Yajie Zeng

NYU TANDON

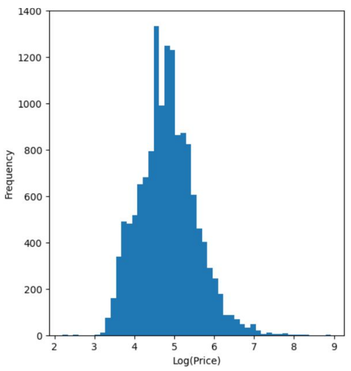
MLC - Spring 2024

Project Introduction

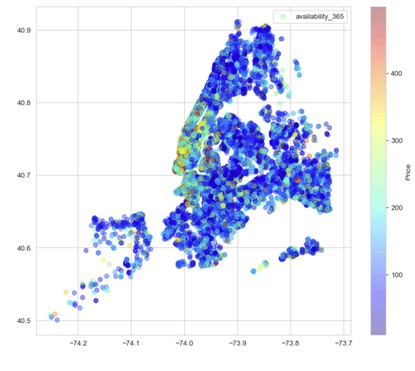
- **Background:** Airbnb has become a pivotal element of the sharing economy, especially in urban settings like New York City. As such, understanding the determinants of Airbnb pricing is crucial for stakeholders to make informed decisions that balance benefits and disruptions within the urban housing market. We focus on analyzing Airbnb pricing dynamics in New York City.
- **Research Question:** Identify key factors in pricing, evaluate the impact of transportation accessibility and develope price prediction models.
- **Method:** Gradient Boosting, Lasso and Ridge Regression, Support Vector Regression

Data Overview

- **Dataset Overview:** NYC airbnb listings information(Airbnb official website); NYC transportation Data
- **Data Processing:** Log (price), missing values, outliers, collinearity and encoding.
- **EDA:** boxen plot of price, price of each listing on map, correlation with price and important variables, correlation heatmap.



Distribution of the Log(Price)

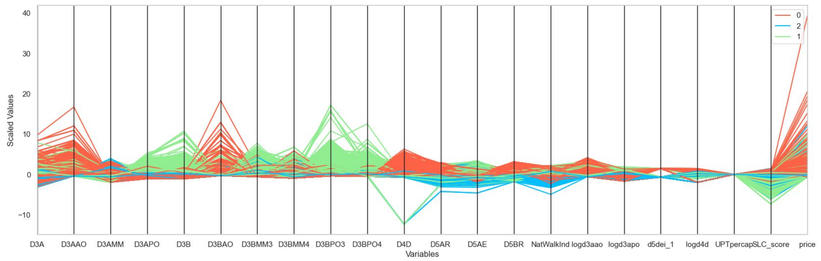


The Price of Each Listing on Map

Methodology

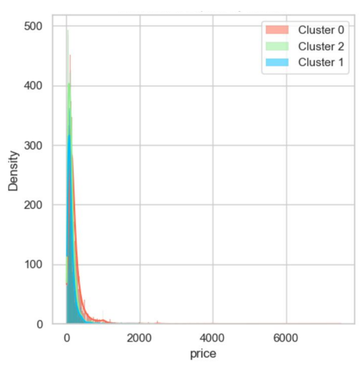
- Use **K-means** on standardized data to explore the impact of traffic factors on house prices, selecting the optimal number of clusters (k=3) using the Elbow Method and Silhouette Score. Analyze cluster differences in characteristics through visualization.
- Build a **Linear Regression** Model by transforming the prices to log scale and standardizing the features. Pick the important features and train the optimized decision tree. Selected features from the decision tree are trained in a random forest model. Train the random forest model by picking out the most important feature and making the predictions more accurate.
- Utilize **Gradient Boosting** for its adeptness at handling non-linear relationships and complex data through an ensemble of decision trees that iteratively enhance prediction accuracy. Integrate **Lasso and Ridge** Regression for their regularization capabilities to counter overfitting. Employ **Support Vector Regression** for its proficiency in high-dimensional spaces and outlier resistance.

Results

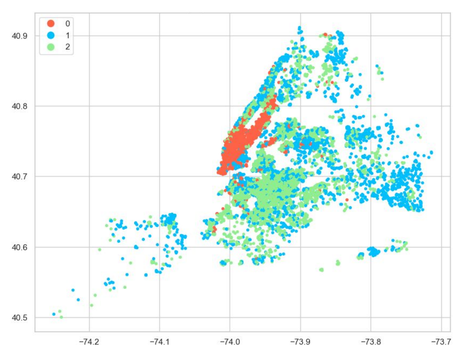


Parallel Coordinates Plot for Clusters

- **D3A*** - the density of the traffic network density:
  D3AAO - Network density of auto-oriented links per square mile
  D3APO - Network density of pedestrian-oriented links per square mile
  D3BPO* - Intersection density of pedestrian-oriented intersections per square mile
- **D4*** - the frequency of transit service
- **D5*** - the accessibility of jobs and workers:
  D5AE - Working age population within 45 minutes auto travel time
  D5BR - Jobs within 45-minute transit commute

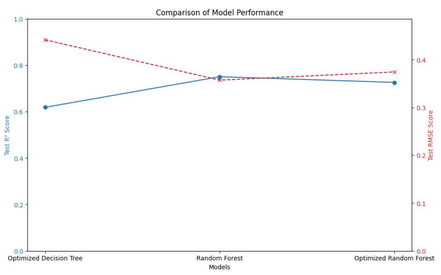


Distribution of price by Cluster

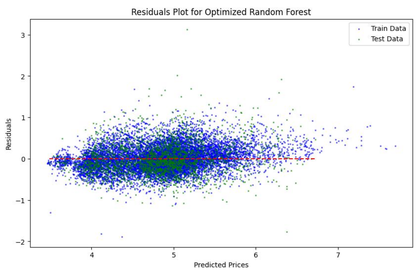


Geographical Distribution of Cluster

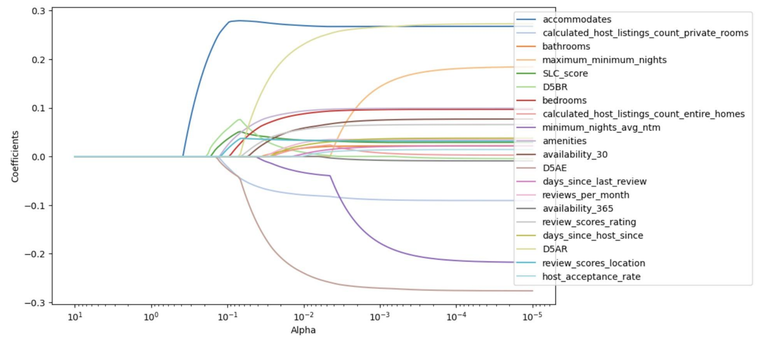
- In the Optimize Decision Tree, the R^2 Value is 0.618, and the RMSE is 0.44. Pushed the R^2 value on the Random Forest test set up to 0.750, RMSE is 0.36. The Optimized Random Forest with an R^2 value stable at 0.726 and RMSE of 0.37.
- The residual plot shows that the residuals of the optimized random forest model are relatively concentrated and distributed, and the data point is around 0, indicating that the consistency and accuracy of model predictions are relatively good and the error is small.



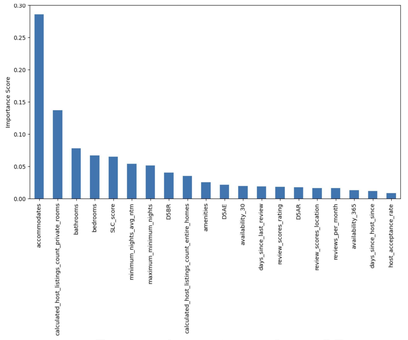
Comparison of Model Performance



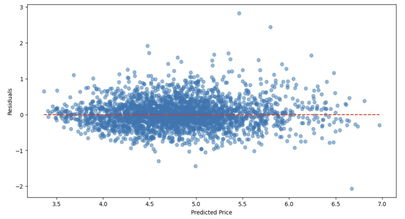
Residuals Plot for Optimized Random Forest



Lasso coefficients as a function of the regularization



Feature Importance from GB



Residual Plot - Test Set

```
Gradient Boosting Training RMSE: 0.29
Gradient Boosting Test RMSE: 0.36
Gradient Boosting Training R²: 0.824
Gradient Boosting Test R²: 0.751
```

- **Gradient Boosting** emerged as the top model due to its superior ability to handle complex datasets and achieve high predictive accuracy, which has the impressive R^2 scores of 0.824 (training) and 0.751 (testing).
- Lasso and Ridge regression techniques address predictive stability, with Lasso achieving an R^2 of 0.542 on the test set. Support Vector Regression further enriches the analysis by capturing non-linear relationships, achieving R^2 scores of 0.762 (training) and 0.695 (testing). The residual plots from SVR demonstrated a generally good fit, with most errors clustering around zero, although some discrepancies pointed towards potential model improvements.

Conclusions

- Transit Proximity: Closer access to public transit increases Airbnb prices.
- Property Attributes: Amenities and room types significantly influence pricing.
- Gradient Boosting Model: High predictive accuracy in pricing analysis.

Recommendations

- Hosts: Use insights to optimize pricing and improve profitability.
- Policymakers: Incorporate findings into urban planning and short-term rental regulations.

Future Research

- Explore seasonal pricing effects and long-term housing market impacts.

References

- Harper, D., & Lim, C. (2021). The impact of environmental quality on Airbnb pricing. Journal of Environmental Economics and Management, 106, 102383.
- Jensen, K., Liu, S., & Yu, J. (2020). Hotel pricing dynamics and Airbnb. Tourism Management, 81, 104145.
- Kim, J., & Park, H. (2021). The impact of technological amenities on Airbnb prices. Hospitality Management, 92, 102719.
- Kisieliauskas, J. (2023). Host-related factors influencing Airbnb prices in rural areas. EconPapers.
- Lee, S., & Nguyen, H. (2019). Safety perceptions and their effects on Airbnb pricing. Crime Science, 8(1), 21.
- Miller, T., & Brown, A. (2019). Determinants of Airbnb prices in European cities: A spatial analysis. Journal of Urban Economics, 87, 85-99.
- Zhang, X., & Gao, J. (2019). Transportation accessibility and Airbnb: A geospatial analysis. Transportation Research Part D: Transport and Environment, 71, 23-36.