Project Report

Airbnb Data Analysis by SAS

Data source: <https://insideairbnb.com/get-the-data/>

Data Details: Albany, New York, United States (05 September 2024)

Submitted By: Bidya Bhattarai

# **Comparing Models and Interpretation:**

## **Examine the regression assumption.**

What is the value of RMSE (Root Mean Square Error)? Assess RMSE for both training and testing samples?

Ans: Here, we are comparing the original price and log transformed price for the better Airbnb price strategy. For the original Price model, the Root Mean Square Error (RMSE) in the training set is 38.62, meaning the model’s average error is about $38 per prediction. In the testing set, the Residual Sum of Squares (RSS) is 307,424, which translates to a testing RMSE of $62.28. It indicates that model is somewhat accurate, but the predictions is not stable mostly for the higher price listings.

In comparison, the Price Log model has a much lower RMSE of 0.3057 in the training set, reflecting a good fit due to the log transformation, which reduces the impact of high-priced outliers. For the testing set, the RSS is 9.1312, leading to a testing RMSE of 0.36 on the log-transformed scale. These value exhibits that model is quite consistent on price prediction even on different price ranges listings.

However, comparing mean residuals of price and price log, mean residual of original price data is 3891.44 which suggests a better fit in comparison to the log-transformed model’s residual mean of 22104.63. This lower mean indicates that the model is closer to accurately predicting prices in their original scale than in the transformed scale.

## **Examine Decision trees results.**

Assess RMSE and compare it against the statistic obtained with the linear regression model.

Ans: To predict Airbnb pricing, decision trees applied to both original price and log transformed price. For the price model, RMSE is calculated from ASE value which is approximately 35.07. The mean residual for this model is 4464.90. From this value, we can there is a significant error in prediction.

In contrast, the price log model has a RMSE approximately 0.24 which was calculated from ASE. The mean residual for this model is 0.1042 which is much lower than price model. This suggest that price log model is more accurate which indicate better predictive accuracy. The log transformation effectively enhances model performance by addressing variance and reducing the effect of high outliers in price predictions.

**Comparing decision trees models with regression models results:**

The Decision Tree model for the original price had an RMSE of 35.07 and a mean residual of 4464.90 which indicates significant prediction errors and challenges in handling raw price variability. In contrast, the price log in Decision Tree model performed substantially better with an RMSE of 0.2486 and a mean residual of 0.1042, showing improved accuracy and stability due to the log transformation.

On the other hand, the Regression model for the raw price had a higher RMSE of 38.62 and a mean residual of 3891.44, while price log Regression model improved with an RMSE of 0.3057 and a mean residual of 22940.94. While comparing these results, the Decision Tree model with log transformation price demonstrated the better fit and most accurate predictions which we can recommend for best reliable pricing for Airbnb reducing prediction errors.

## **Examine Random forests results:**

Ans: In random forest model, for the original price. with ASE of 3230.5, RMSE is approximately 35.07 which indicates the training error due to variability. The Mean residual is 5013.19 over 79 observations, showing considerable error on unseen data due to price variance and outliers. However, for the price log model, with ASE of 0.200, RMSE is approximately 0.1411 which is lower as compared to the raw data while the mean residual is 0.0949. This improved highlight suggests that log transformed price is recommended with the stabilize variance, less impact of outliers leading to accurate and reliable predictions.

**Recommendation:**

In this step, price log model is suggested due to the smaller mean and ASE residuals. The log transformation seems to improve predictive performance, making the model more accurate and dependable by addressing the variation issues in the original price values.

## **Three models interpretation:**

After analyzing the logistic regression, decision tree, and random forest models on the Airbnb dataset, it seems that the Random Forest model with price log transformation is suggested as the best fit for forecasting the target variable price. The log transformation provided a little improvement for both the regression and decision tree models, they still showed greater residual errors than the Random Forest. By handling data variability and outliers, the Random Forest model with price\_log provides great prediction accuracy with the lowest mean residual. Since it constantly produces predictions that are closer to actual values, this model is the best option for business applications and offers the most dependable basis for setting competitive prices. In business point of view, it will improve profitability with the matching market price by minimizing the risk of underpricing and overpricing properties.