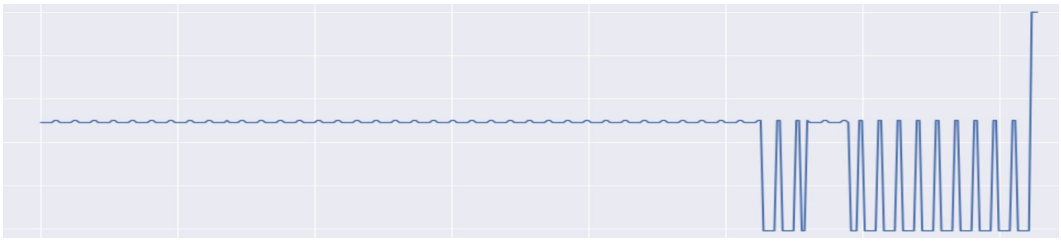
Capstone Project 1: In Depth Analysis

1. Time series patterns:

    A. Weekend spike:

    The weekend has a higher rate than weekdays, hence we see the spikes.

    B. Summer rate:

    This listing has a higher rate set for the summer.

    C. One rate:

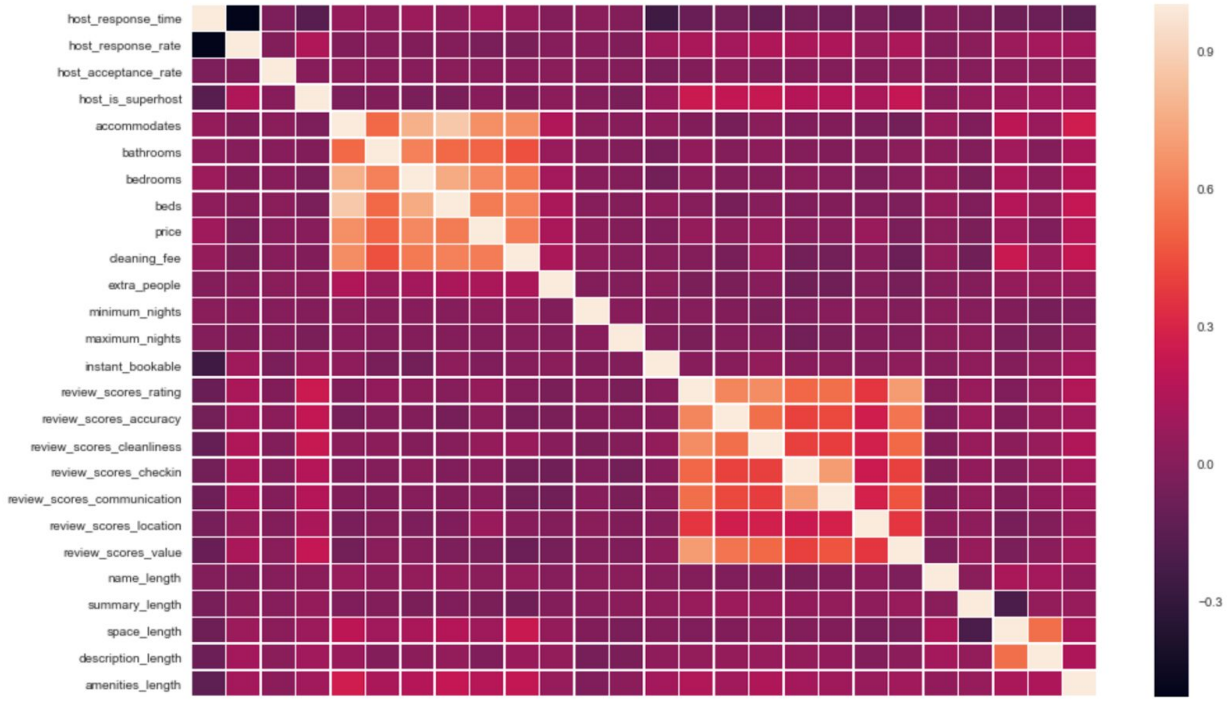
    

    This listing sets at a fixed rate all year round.

    D. Holiday rate: Peaks on Christmas and high rate during summer

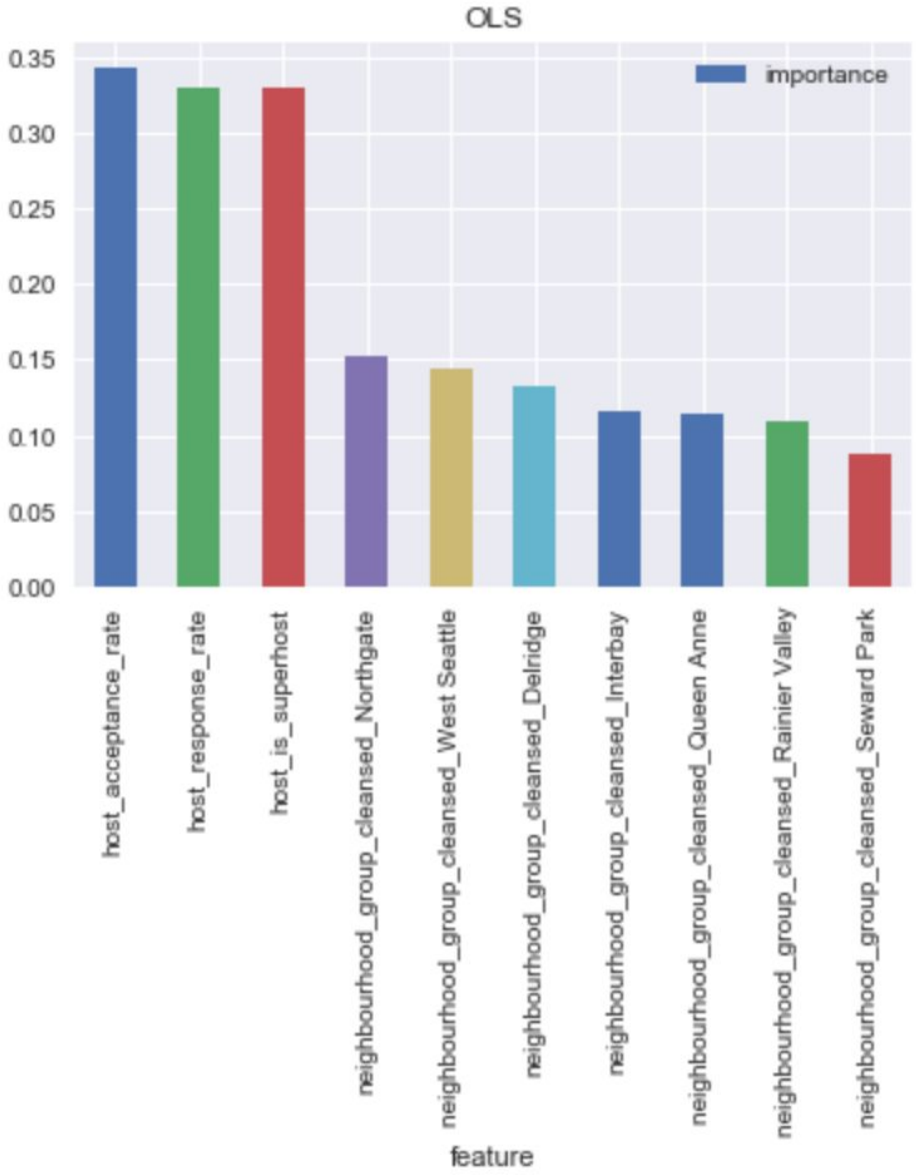
    

2. Preprocessing
    A. Get lengths of description elements
    B. Convert host response time to spectrum 1-4, host response time from percentage to integer, superhost to binary 0 and 1, price from $ to numbers.
    C. Fill missing values
    D. Convert categorical features to numbers.
    E. Omit all listings that miss our target: review_scores_value
    F. Matrix of correlations shows 3 groups of correlations:
        - Description length and summary length: hosts spend about the same amount of time to write descriptions and summaries.
        - Rating correlation: overall ratings and individual scores are highly correlated.
        - Price, beds, accommodates, bedrooms, bathrooms, cleaning fee are highly correlated: people use these features to price their listing. But is it the right way?

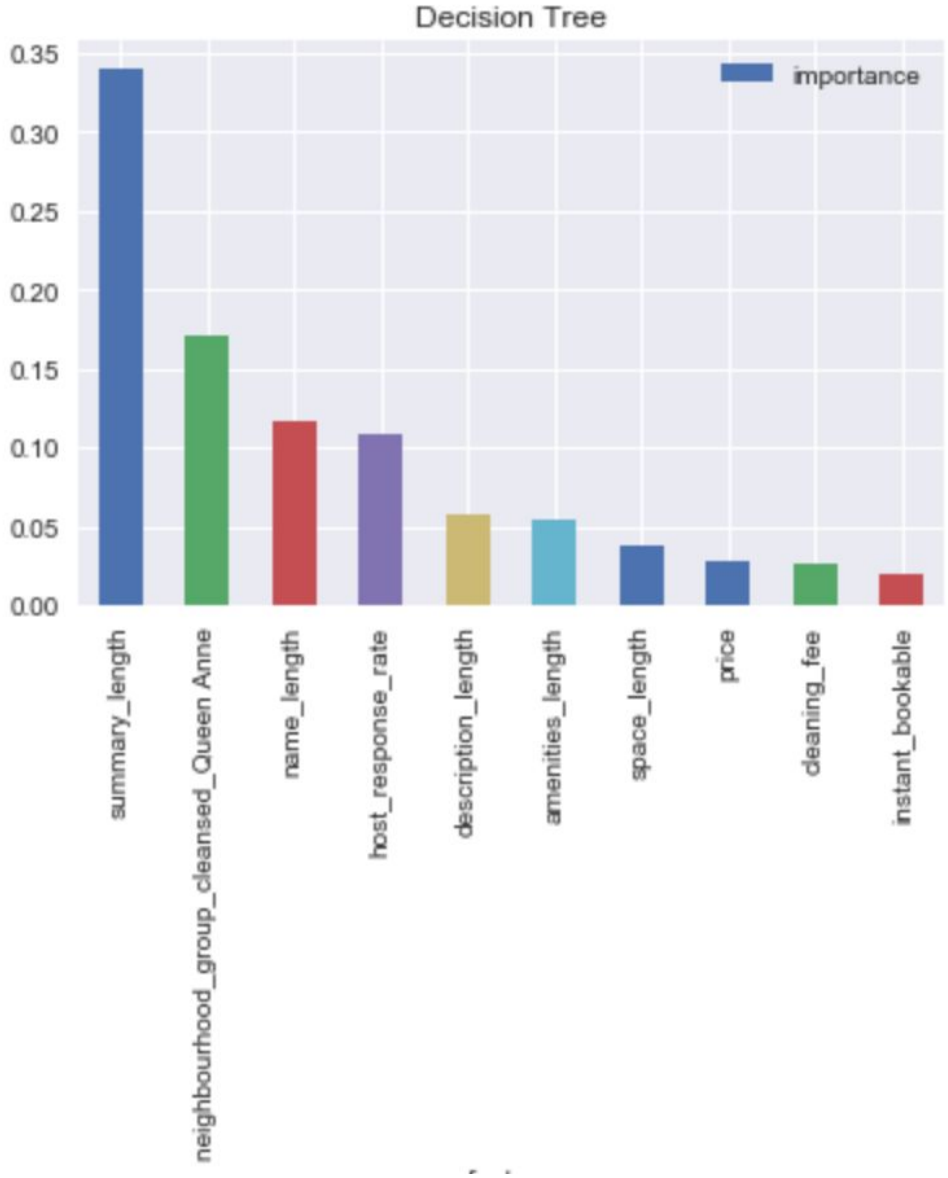


3. Modeling: Logistics Regression, Decision Tree, Random Forest, and Gradient Boosting

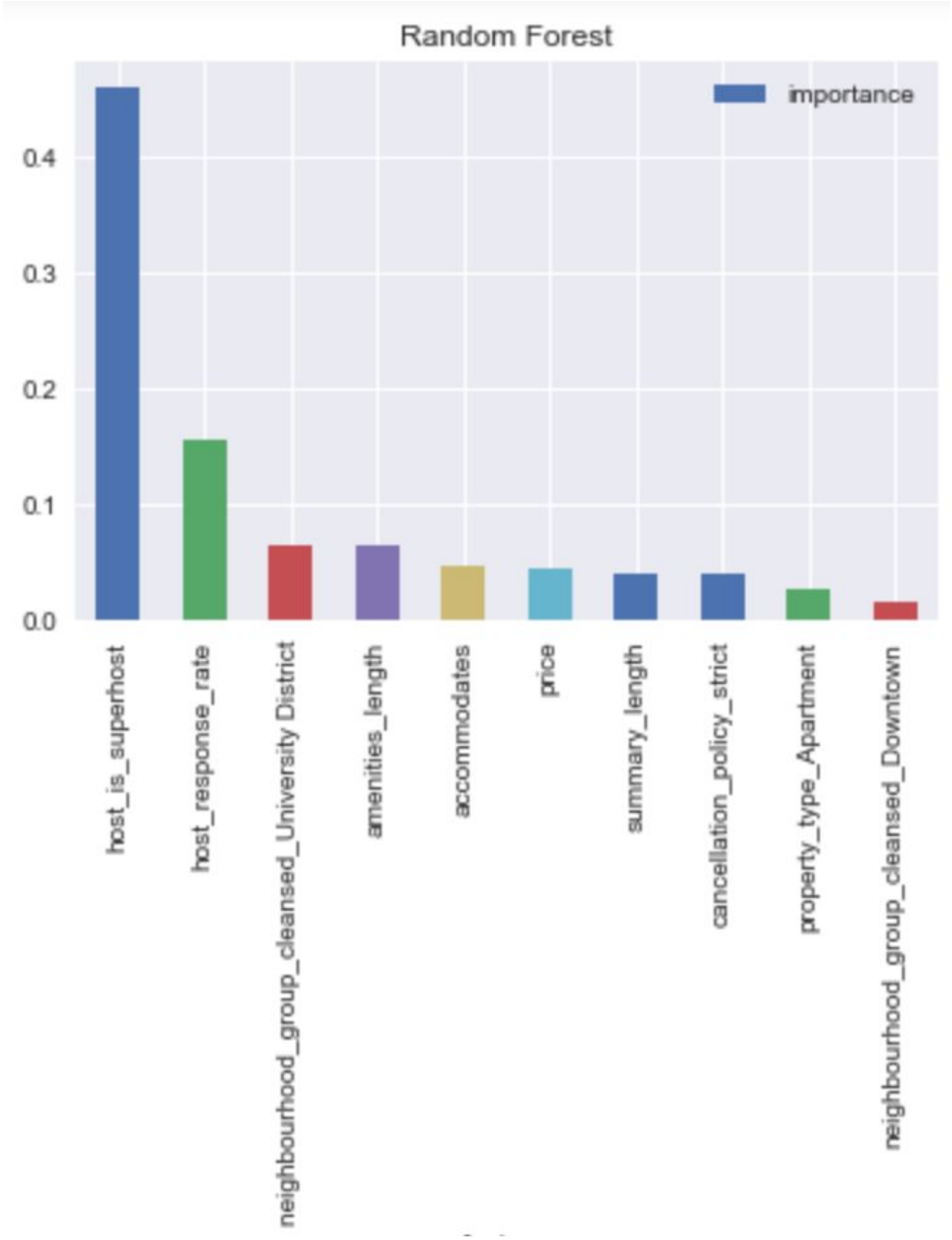
A. Logistics Regression: With a mean square of errors of 0.6977697963551963 10 features that determines the value score are host related and location related. But location score has the weakest correlation with value score. Is that an indicator to the difference between locations and neighborhoods? Or a preference in neighborhood or what guests prefer as individuals are difference than a location as convenience to move around? One thing we learn from this is that what can determine the value score, does not necessary be the most important thing or preferred things. Later in other models we can see that the neighborhoods group changes quite interesting.



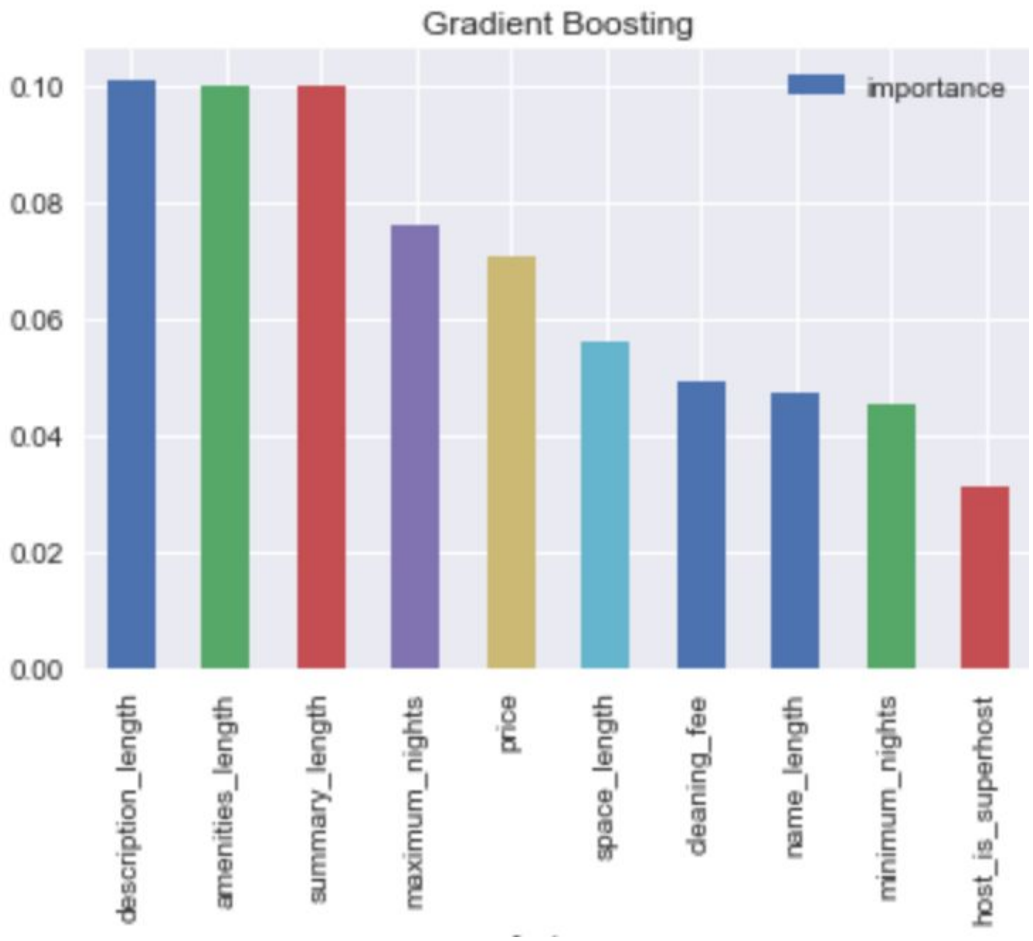
B. Decision Tree: with a mean square of errors of 1.8637822325921867
Decision Tree has the highest mean square of errors or the least accurate model (even though we don't use accuracy score here as our evaluation)
10 most important features are completely different from OLS, but at the same time makes more sense as explanations as to what are most important features in determine values of a listing.



C. Random Forest: with a mean square of errors of 0.7058711712461408, close enough to OLS, with a diverse group of 10 most important features in which 9 of the top 10 features weigh almost evenly, with the indicator that superhost is the determining factor in a high value score.
Notice how the neighborhoods changed from Queen Anne in Decision Tree model to University District and Downtown in Random Forest? While other top features are almost the same in both models: price, summary length, host response rate.



D. Gradient Boosting: with a mean square of errors of 0.8669365301845464. In this model, we do not see any neighborhood group in the top 10 important features, and all top 10 features weigh evenly. Gradient Boosting is the closest to the correlation matrix.




It is hard to decide which is the best model depending on the mean square of errors. OLS might have the lowest MSE, also the fastest model, Gradient Boosting makes more sense to me as what determines values.

However, if numbers speak the truth, there is an underlying relation of value score and what guests value subjectively. As different neighborhoods of Seattle fit different styles and personalities, location might not determine one's value score, but a fitting neighborhood could determine that.

The most common trend of vacation: people are looking for an old vintage, bohemian look which fit Queen Anne, and University District. There is a certainty that styles play an important factor to value these days.