AIRBNB HOST RATING OPTIMIZATION MODEL

Instructor: Prof. Daniel Acuna

Group 4: Yesaswi Avula, Yi Zhong, Woojin Park, Sijie Li

Problems

- The sharing economy is taking American households and vehicles by storm as the expansion of start-up and smart-phone driven sharing services such as Airbnb, Uber, and Lyft expand across the country. Our project examines the implications of the sharing economy, specifically aiming to find the best regression model for Airbnb host's review ratings.
- Currently, the hosts have too many inputs to manage such as prices, facilities, policies, response rate and so on that they don't have a clear idea on what should be the focus.

Objectives

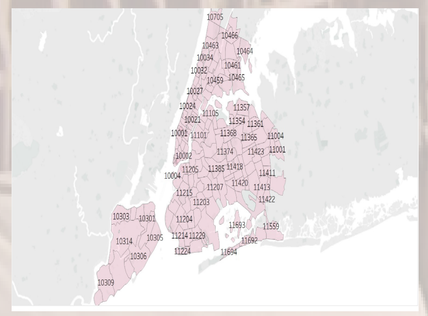
- Perform descriptive statistics to have basic summaries of the data and graphics analysis to form the basis to virtually analyze the quantitative aspects of data.
- Use different machine learning models to find out the key factors influencing the target variable and compare their performance.

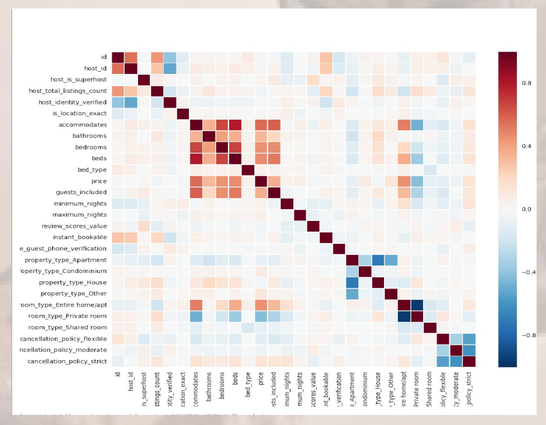
Lorem ipsum

Data Description and Preprocessing

- The dataset: 25 *49057
- Target loation: New York City
- Compiled on: 6th December 2018.
- Target variable: review_scores_value.
- Missing Values: Dropped the NA's.
- Outliers: Winsorized the data above 99% and 1%.
- Created dummy variables for some variables and regrouped a few categorical variables that are required to.

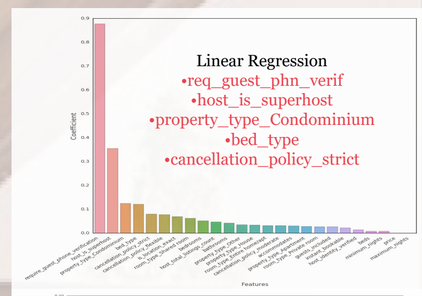
Data Visualization

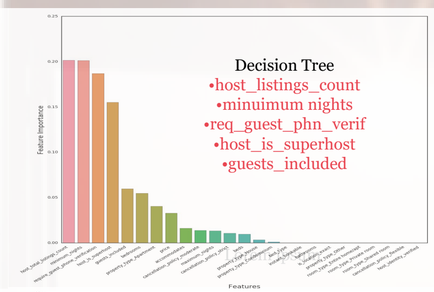


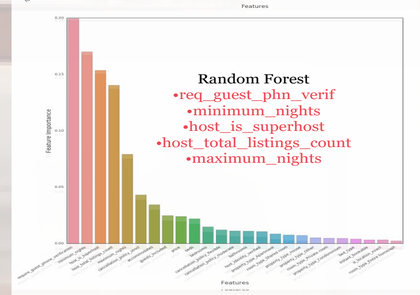


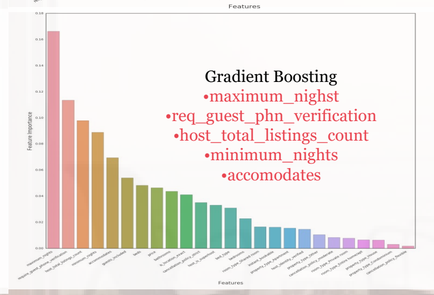
Model Description & Feature Importance

- Regression: Learning a functional relationship between input features and an output target using training data where specific functional form learned based on the choice of model
- Model: Linear regression, Decision Tree regression , Gradient Boosting regression, and Random Forest regression
- Feature Importance: By utilizing feature importance increase our model's performance and know the impact of a certain feature on the model's performance -> Choosing features with the greater importance to do the dimension reduction.





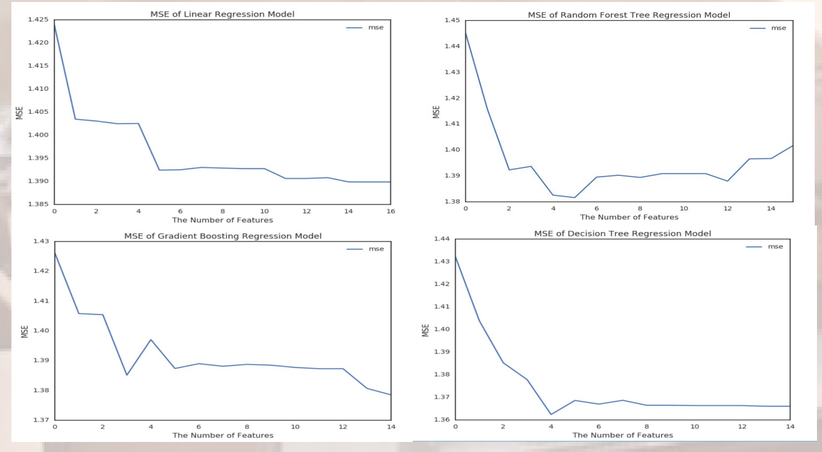




Model Comparison Metrics

- We split the entire dataset into training, validation and testing in 60%, 30% and 10% respectively
- We use MSE(Mean Squared Error) for the generalization performances measure of models & our goal is to minimize sum of squared distances.

Feature Selection



Parameter Tuning

Linear Moedel:
Elasticnetparam = [0.1,0.2,0.3,0.5,0.6], Regparam = [0.01,0.02,0.03,0.04]

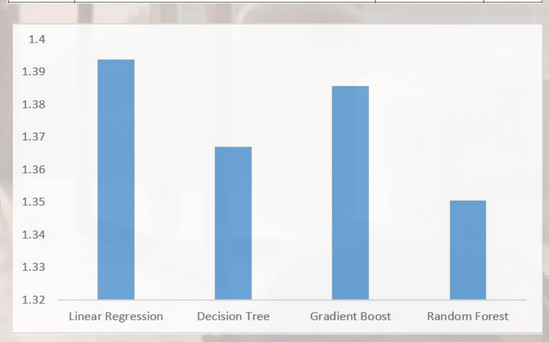
Decision Tree:
MaxBins = (2,10,3), MaxDepth = (1,10,2)

Gradient Boosting:
MaxIter = (2,11,2), MaxDepth = (2,10,2)

Random Forest:
NumOfTrees = (1,200,20), MaxDepth = (2,10,2)

Results

Model	Features	Parameter	MSE
Linear Regression	'require_guest_phone_verification','host_is_superhost','property_type_Condominium','bed_type','room_type_Shared room','cancellation_policy_strict','cancellation_policy_flexible','is_location_exact','bedrooms','room_type_Entire home/apt','room_type_Privateroom','host_total_listings_count','cancellation_policy_moderate','bathrooms'	ElasticNetParam (0.1),RegParam(0.01)	1.393787
Decision Tree	'minimum_nights','require_guest_phone_verification','host_is_superhost','maximum_nights','host_total_listings_count','cancellation_policy_strict'	MaxDepth(8), MaxBins(5)	1.366983
Gradient Boosting Tree	'maximum_nights','host_total_listings_count','accommodates','require_guest_phone_verification','bathrooms','minimum_nights'	MaxDepth(8), MaxIter(8)	1.385786
Random Forest	'host_total_listings_count','require_guest_phone_verification','host_is_superhost','minimum_nights','maximum_nights','cancellation_policy_strict'	NumTrees(181), MaxDepth(8)	1.350456



Conclusion

- The random forest regression model has the lowest MSE which is 1.350456 and then is the decision tree regression model.
- Based on our best regression model's output, we suggest hosts pay more attention on the following aspects like 'whether phone verification is needed', 'whether they are super host or not', ' what is their minimum nights' which are crucial for their review score

Challenges & Future Work

- Most of the review scores are very high in the data. **Hence, we** collect data with low review scores and try to improve the performance of the model for balanced data with all types of reviews.
  Use the text data in few of the variables and implement sentiment analysis to understand ideas of the people and improve the ratings.
- Apply more complex algorithms like Deep Neural Network.

Data Source: http://insideairbnb.com/get-the-data.html