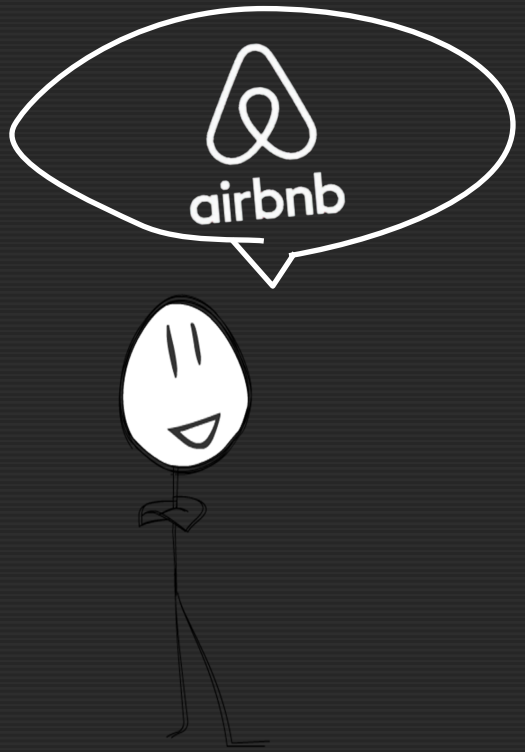


How to Succeed as an Airbnb Host

---Evidence from Text Analysis and Machine Learning Approach

MACS 30250 Research Proposal

Yutian Lai

1 Research Question

What factors would help
Airbnb hosts obtain high ratings?

Importance

- Ratings can be used as an inferential metric to the satisfaction level of guests
- Explore how hosts with high ratings strategically position themselves through text analysis(topic modeling and bigram)
- Find the correlation of home characteristics and hosts description on customers' satisfaction level
- Generate machine learning models to predict ratings

2 Theory

Research has been conducted to investigate the correlation between social features/house characteristics with house sales/ratings (Lee et al., 2015) , and to identify guests' needs from their reviews(Cheng& Jin, 2019)

Text can be invaluable in capturing marketing insights and serve as a new variable in economic/business research(Balducci & Marinova, 2018); **Machine learning** has been widely adopted in marketing research(Jordan& Mitchell, 2015); **Combination** of the two can help accurately identify customers' needs(Glance et al, 2005)

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science, 46*(4), 557-590.

Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, *76*, 58-70.

Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., & Tomokiyo, T. (2005). Deriving marketing intelligence from online discussion. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 419-428).

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260.

Lee, D., Hyun, W., Ryu, J., Lee, W. J., Rhee, W., & Suh, B. (2015). An analysis of social features associated with room sales of Airbnb. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing* (pp. 219-222).

3

Data source

DATA SOURCE:

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

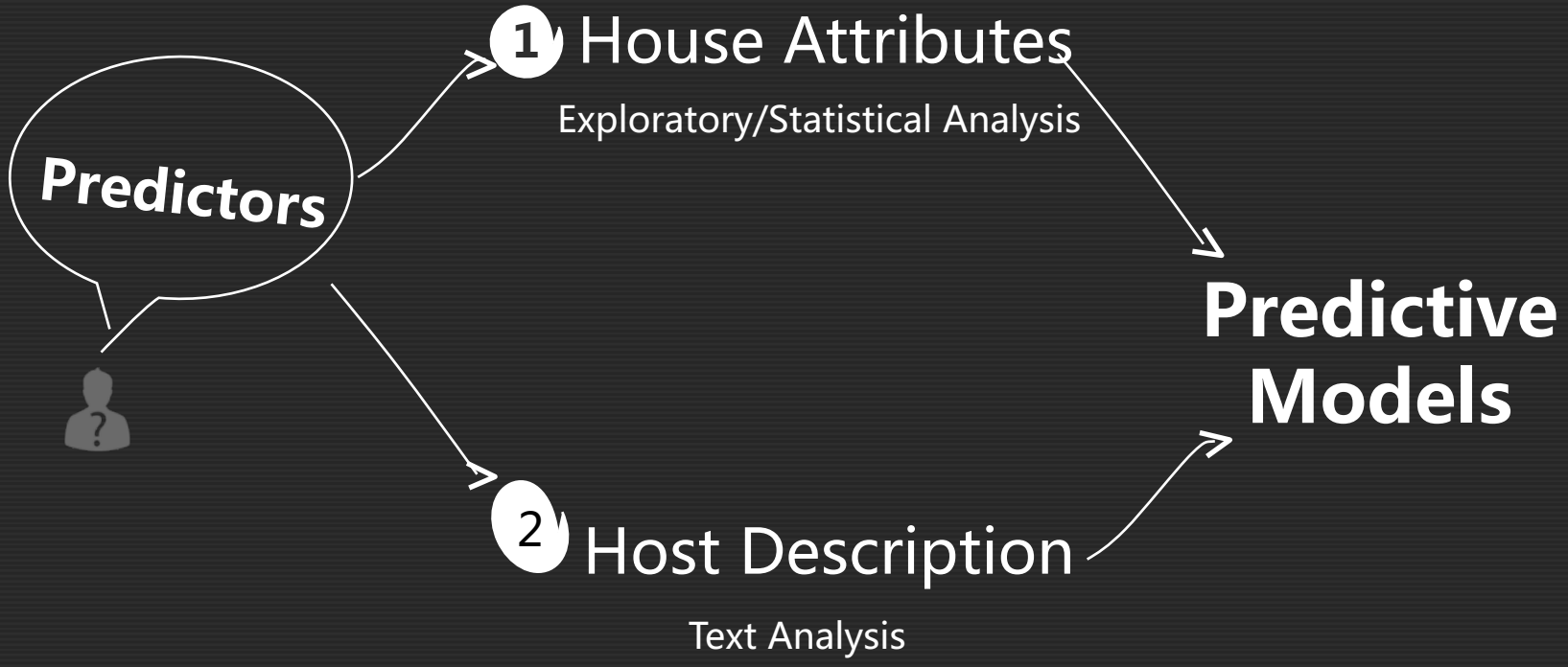
CITIES :

New York
Tokyo
London
50000 houses

Variables(80)

- **Review_scores_rating:** guests' evaluation, ranging from 0 to 100
- **Summary:** general introduction of the house
- **Host_about:** self-introduction of the hosts
- **Neighborhood_overview:** description of the surroundings of the house
- **Bathrooms :** number of bathrooms
- **Bedrooms :** number of bedrooms
- **Room_type:** type of room offered(room or whole house)
- **Price :** nightly price of the listings
- **Long-term stays allowed:** does the listing allow for long term stay

4 Methods


Predictors
1
House Attributes
Exploratory/Statistical Analysis
2
Host Description
Text Analysis
Predictive Models

4.1 Exploratory/Statistical Data Analysis

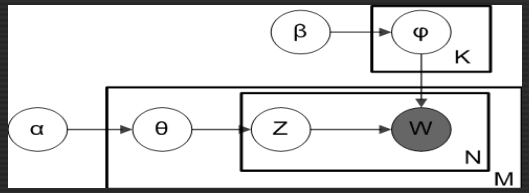
- Distribution of rating
- Price and rating
- Transportation and rating
- Location and Rating
- Bathroom/bedroom number and rating
- Host identification and rating
- Time as a host and rating
- Time of staying and rating
- Host response time and rating
- Room type and rating
- Neighborhood and rating

.....

4.2 Text Analysis

- **Classification:**
- high ratings(review score>=90)
- low ratings(review score<90)

- **Topic Modelling:**
- Latent Dirichlet Allocation



- **Bigram Analysis**:
- more useful than single word counts

- **Similarities and Dissimilarities**

4.3 Predictive Models

- Heatmap of Kendall Correlation between Numerical Features
- Preprocessing

- **Linear regression**

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2$$

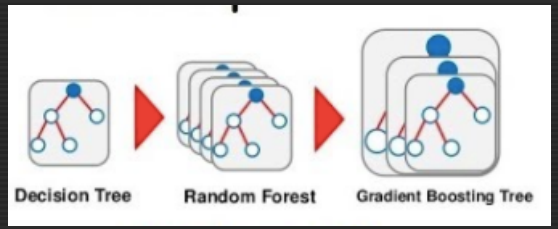
- **Lasso regression(bag of words)**
- **Lasso regression (TFIDF)**

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|$$

- **Ridge regression(bag of words)**
- **Ridge regression(TFIDF)**

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}w_j^2$$

- **Random Forest**
- **Gradient Boosting**



- Find the model with best performance (MSE/Mean Absolute error)
- Draw feature importance plot to find which predictors are impacting the ratings most

4.3 Predictive Models

- Heatmap of Kendall Correlation between Numerical Features
- Preprocessing

- **Linear regression**

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2$$

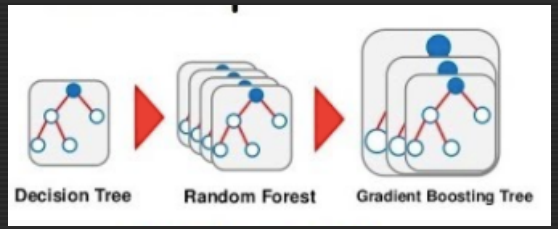
- **Lasso regression(bag of words)**
- **Lasso regression (TFIDF)**

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|$$

- **Ridge regression(bag of words)**
- **Ridge regression(TFIDF)**

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}w_j^2$$

- **Random Forest**
- **Gradient Boosting**



- Find the model with best performance (MSE/Mean Absolute error)
- Draw feature importance plot to find which predictors are impacting the ratings most

5.1 Expected Results

- Gain a better (maybe counterintuitive) understanding of the real value of the rating system
- **Hosts** have a guideline to condition their self-introduction and houses to high ratings, so as succeed in the competition with neighboring Airbnb houses.
- **Airbnb** could motivate hosts to have certain features, thus improving guest lodging/return rates, and boosting business growth.
- Predictive models provide marketing researchers with new metric of customers' satisfactory level

5.2 Alternatives

- **Geographical Difference---** London, New York and Tokyo

- **Sentiment analysis** on customer review as alternative metrics to satisfactory level

- When dealing with feature collinearity, use **PCA** rather than reducing some of the features