Airbnb House Price Prediction

Team 2
MSBA Cohort B

Agenda

1. Defining Business Problem
2. Exploratory Data Analysis
3. Initial Price Model
4. Review Analysis NLP
5. Model Ensembling
6. Summary
7. Q&A
8. Appendix

I. Defining Business Problem

BUSINESS OBJECTIVE AND METHOD

Goals:

- To build a predictive model for price
- To integrate review information into pricing model

Challenges:

- Some of the listings are not intended for “short-term vocational booking”
- Review scores are not reflective and not consistent

Solutions:

- Using insights generated from review text mining, calculating true negative reviews
- Directly apply text mining results for listing scores
- Ensemble listing score & negative reviews labels in final price model

DATA & RESOURCES

Data Description

Data Source:	InsideAirbnb.com (3rd party scraping)
Data Info:	Listings & Reviews
Data Dimension:	Listings 50,599 * 106 Reviews 1,255,322 * 6
Key Variables:	Price, Neighborhood, Availability, Property type, Number of beds, Number of bedrooms, Accommodates, Review scores

Libraries

Cleaning:	lubridate, varhandle, caret
Visualizating:	rpart,dygraphs,wordcloud xgboost Explained, DiagrammeR, coefplot
Modeling:	glmnet, xgboost, lightgbm
Text Analysis:	tidytext, quanteda,cld3

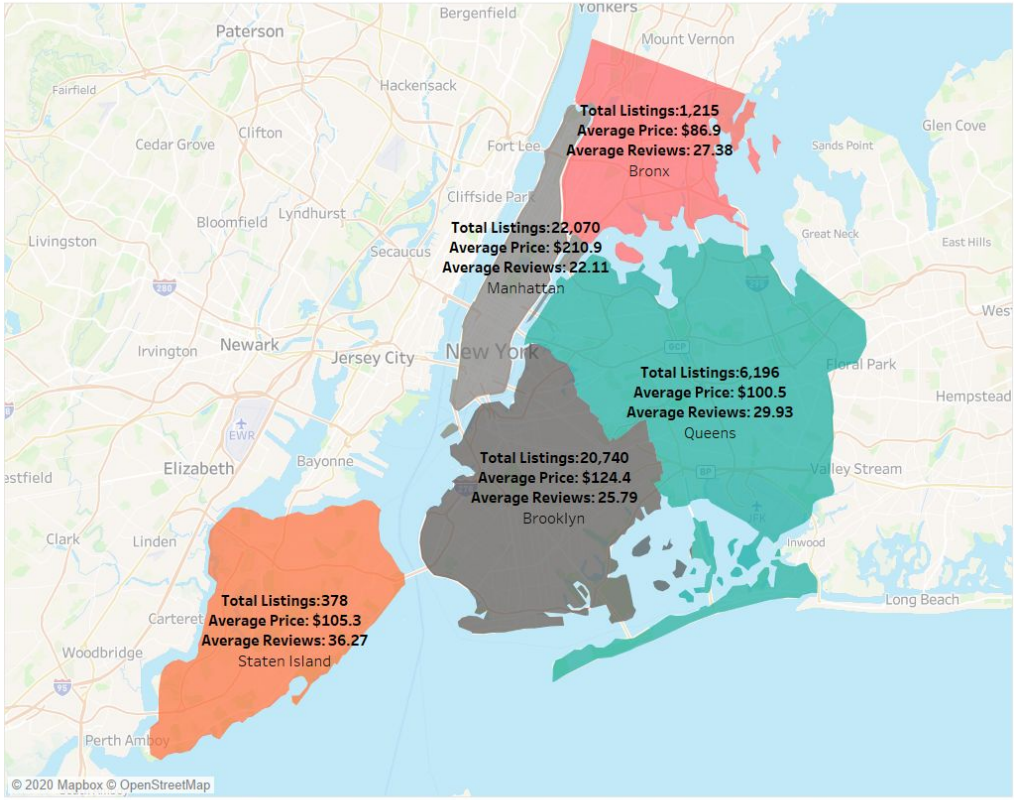
Languages & Tools

Cleaning & Storage:	Google Translation Google BigQuery
Visualizing:	R, Tableau, Google BigQuery
Modeling:	R, Python

II. Exploratory Data Analysis

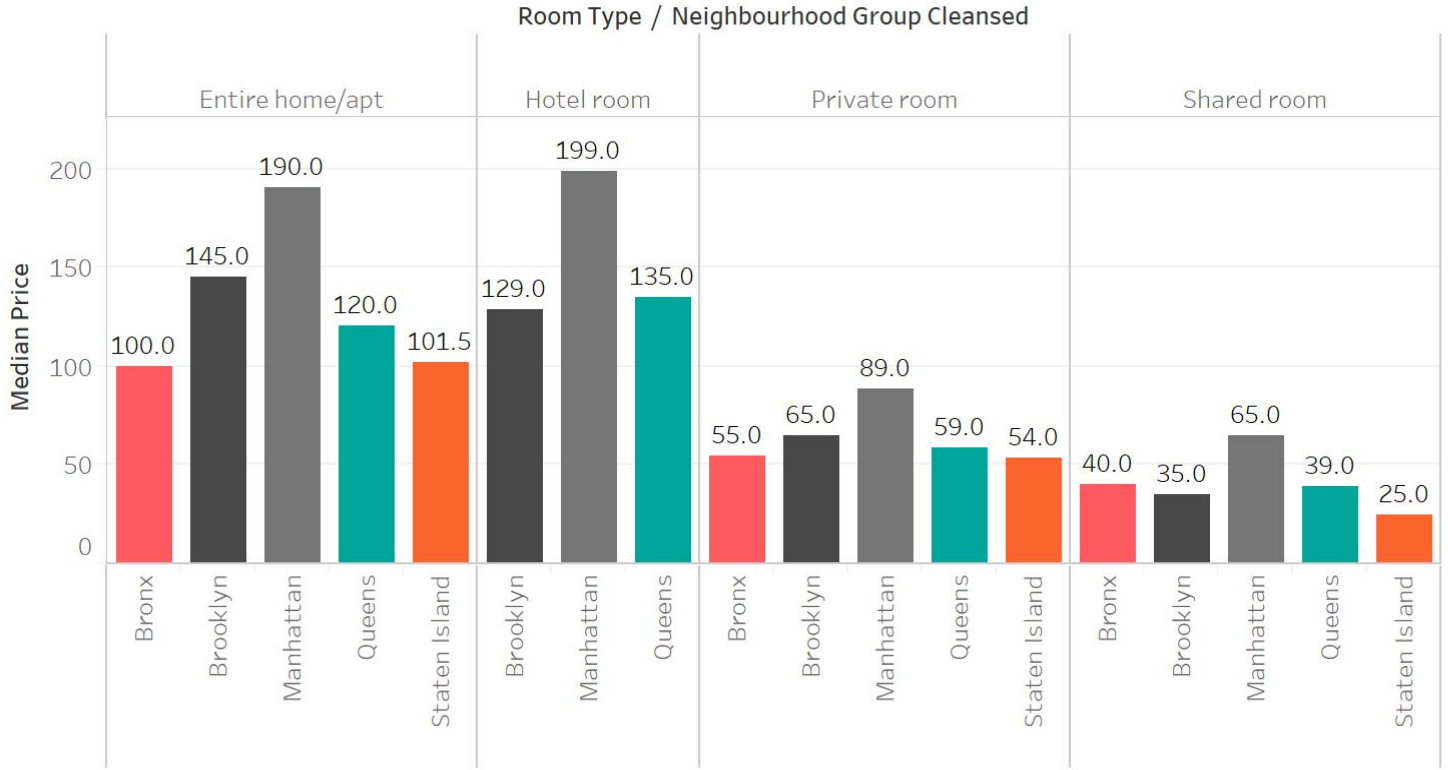
Overall Prices and Listings Distribution of New York

Listings Differ by Neighborhood & Manhattan Tops Listing Counts and Price



Map based on Longitude (generated) and Latitude (generated). Color shows details about Neighbourhood. The marks are labeled by sum of Number of Records, average of Price, average of Number Of Reviews and Neighbourhood.

Manhattan Has Highest Price on Average & Hotel Room and Entire Home Lead Prices in Room Types

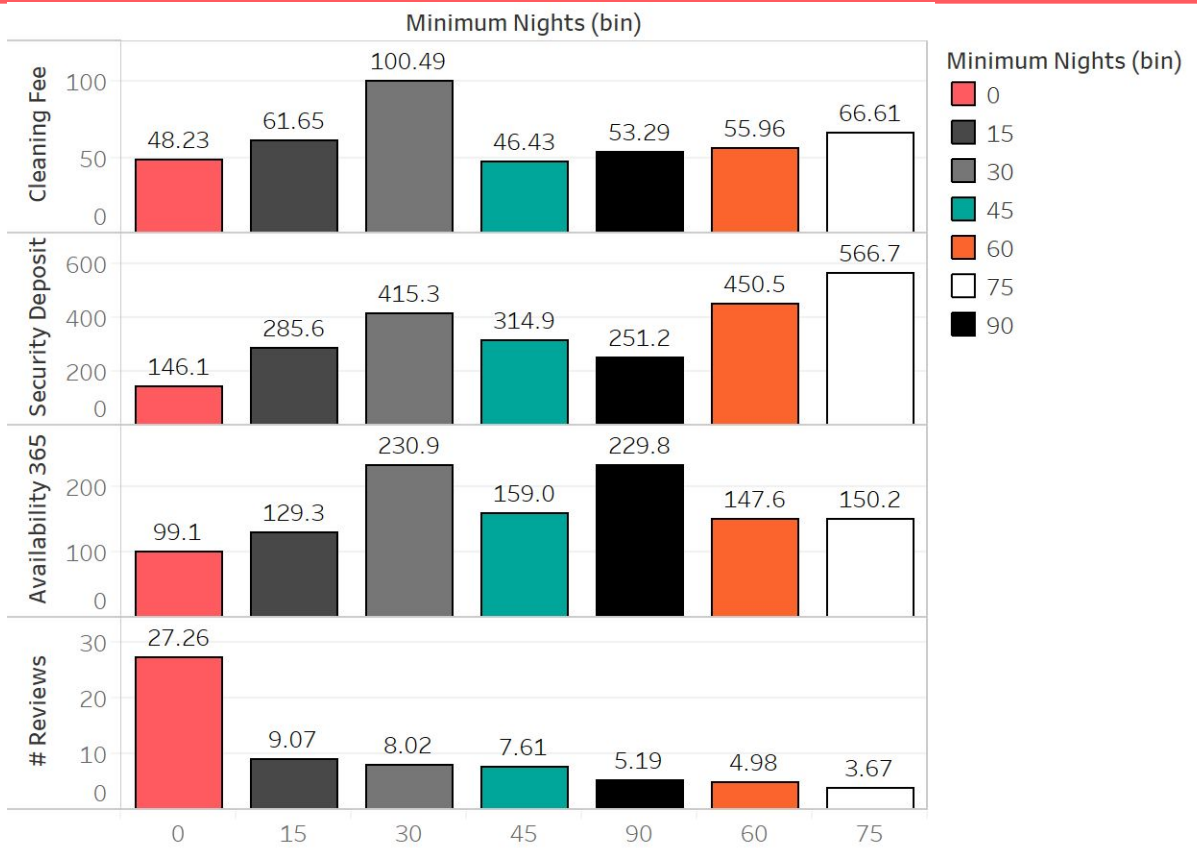


Median of Price for each Neighbourhood Group Cleansed broken down by Room Type. Color shows details about Neighbourhood Group Cleansed. The marks are labeled by median of Price.

Hotels Top Price but There Aren't Too Many of Them

Average of Price for each Property Type. The marks are labeled by sum of Number of Records. The view is filtered on Property Type, which keeps 16 of 35 members.

Higher Fees, Availability but Fewer Reviews for Rentals



89.88%

listings are “**Short-term**” listings

that have a minimum-night less than 30

We **excludes** the 10.1 % “Long-term” rental

in this project

III. Initial Price Model

LASSO REGRESSION

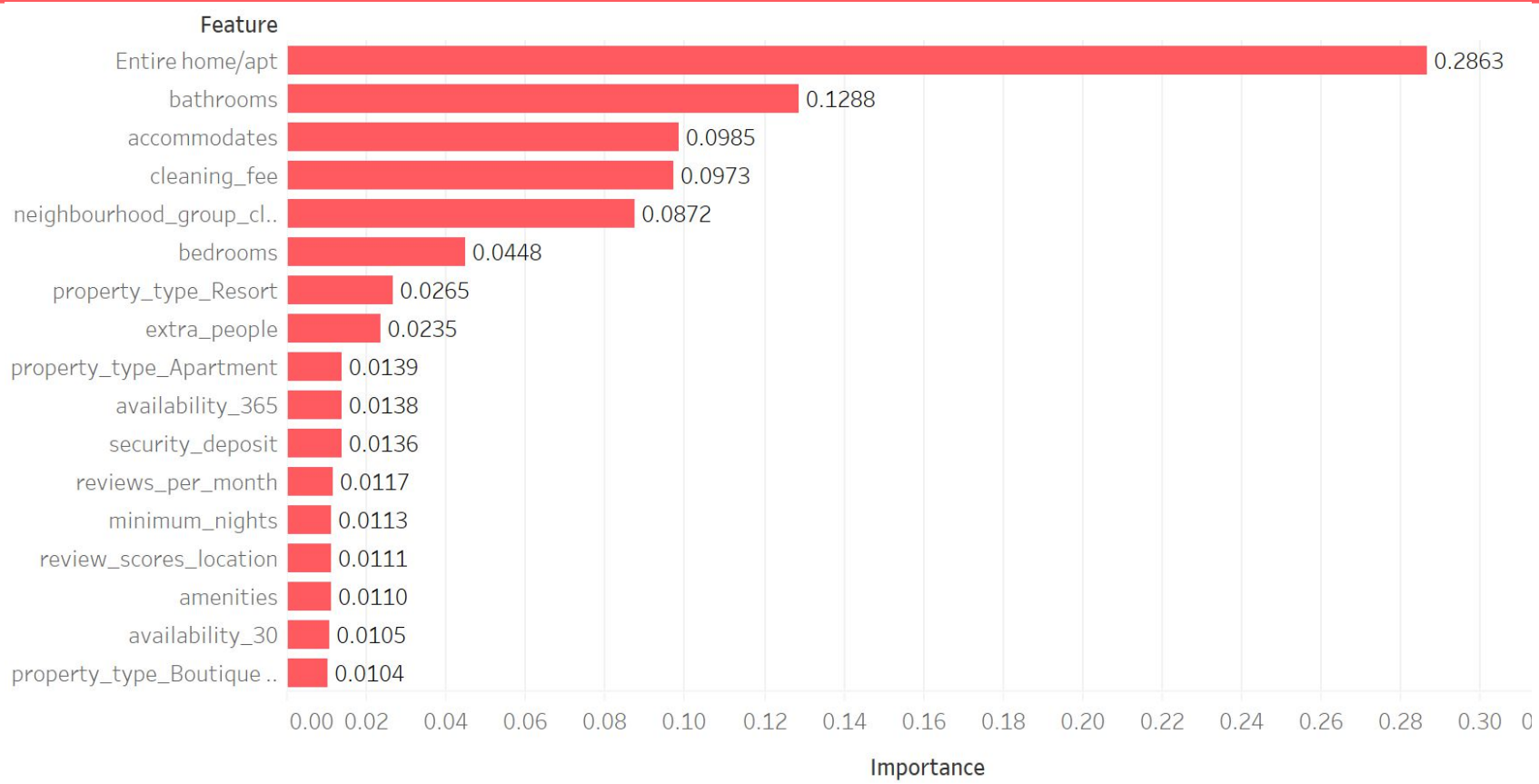
factor	coef
property_type_Resort	233.65
property_type_Boutique hotel	50.84
neighbourhood_group_cleansed_Manhattan	43.47
room_type_Entire home/apt	41.17
bathrooms	23.37
Intercept	15.30
bedrooms	13.76
accommodates	12.42
room_type_Hotel room	9.59
property_type_Loft	9.05

factor	coef
property_type_Condominium	6.16
reviews_per_month	0.86
review_scores_checkin	0.38
cleaning_fee	0.31
availability_30	0.03
security_deposit	0.00
number_of_reviews	-0.01
review_scores_value	-0.62
room_type_Private	-7.36

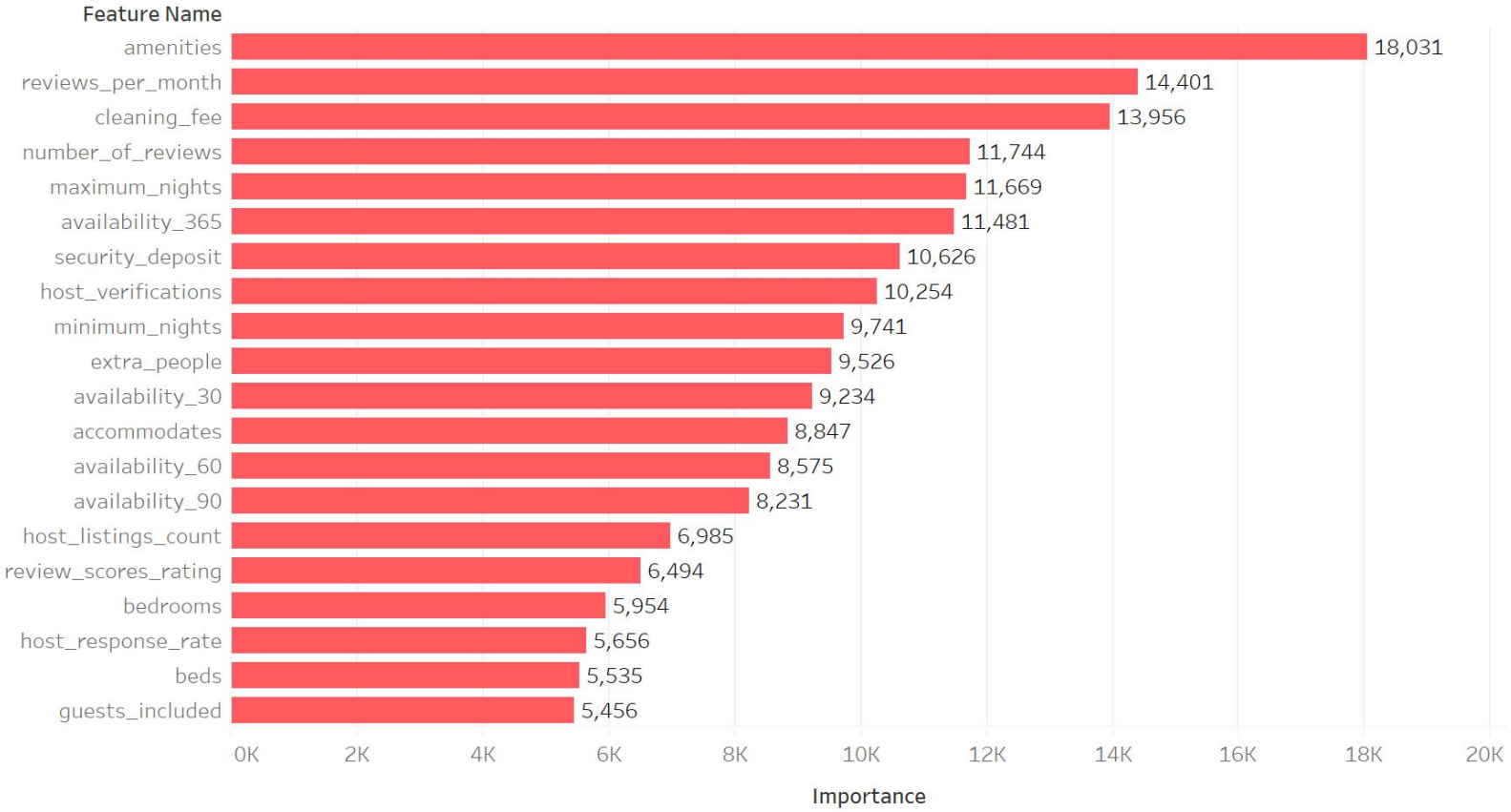
What Matters?

- Location (Manhattan +)
- Apartment Layout (accommodates +, bedrooms + , bathrooms +)
- Property Type (Resort +, Boutique hotel +)
- Room Type (Entire home/apartment +, Private Room -)

TOP FEATURES IN BEST XGBoost MODEL

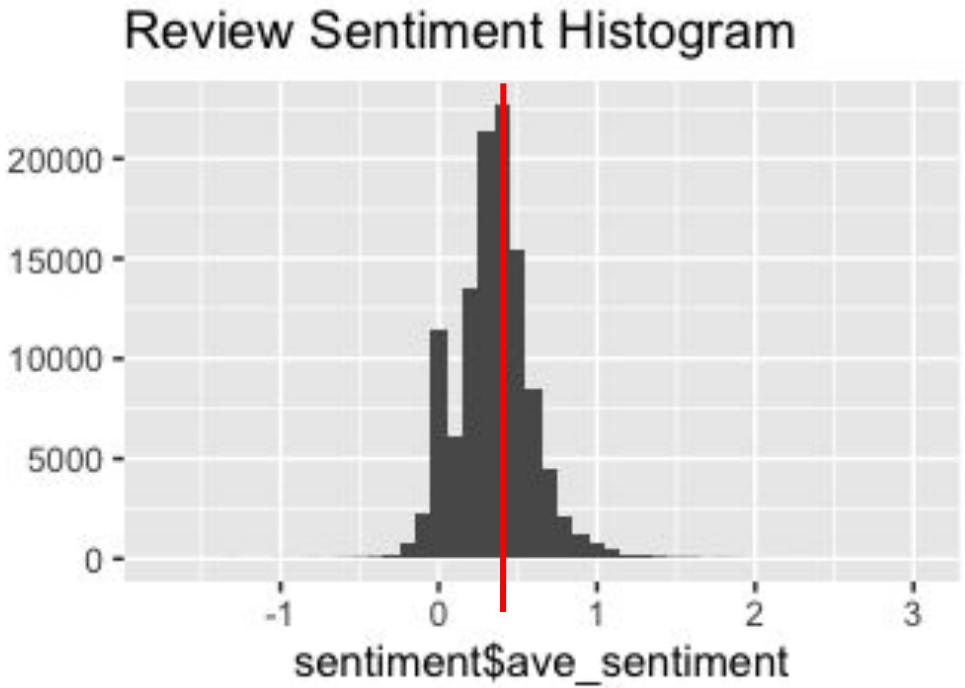


TOP FEATURES IN LIGHTGBM MODEL



IV. Review Text Analysis

EDA: Most of the reviews have positive sentiment scores per sentence



Using Reviews Score Proxy: Bad Review Predictor

Goal:

- Use textual analysis to pick out the reviews that contain negative sentiments due to problems with the listings

Issues:

- People score subjectively with inconsistent standards;
- People may feel “being nice” and not giving out poor scores for reviews, the average rating on Airbnb is 4.7 out of 5 stars;
- Airbnb did not have a review system that prevents retaliation review until 2014.

Solution for
Misstatement & Retaliation

Airbnb launched new review policy in mid-2014, **preventing hosts retaliating guests** by writing a malicious review.

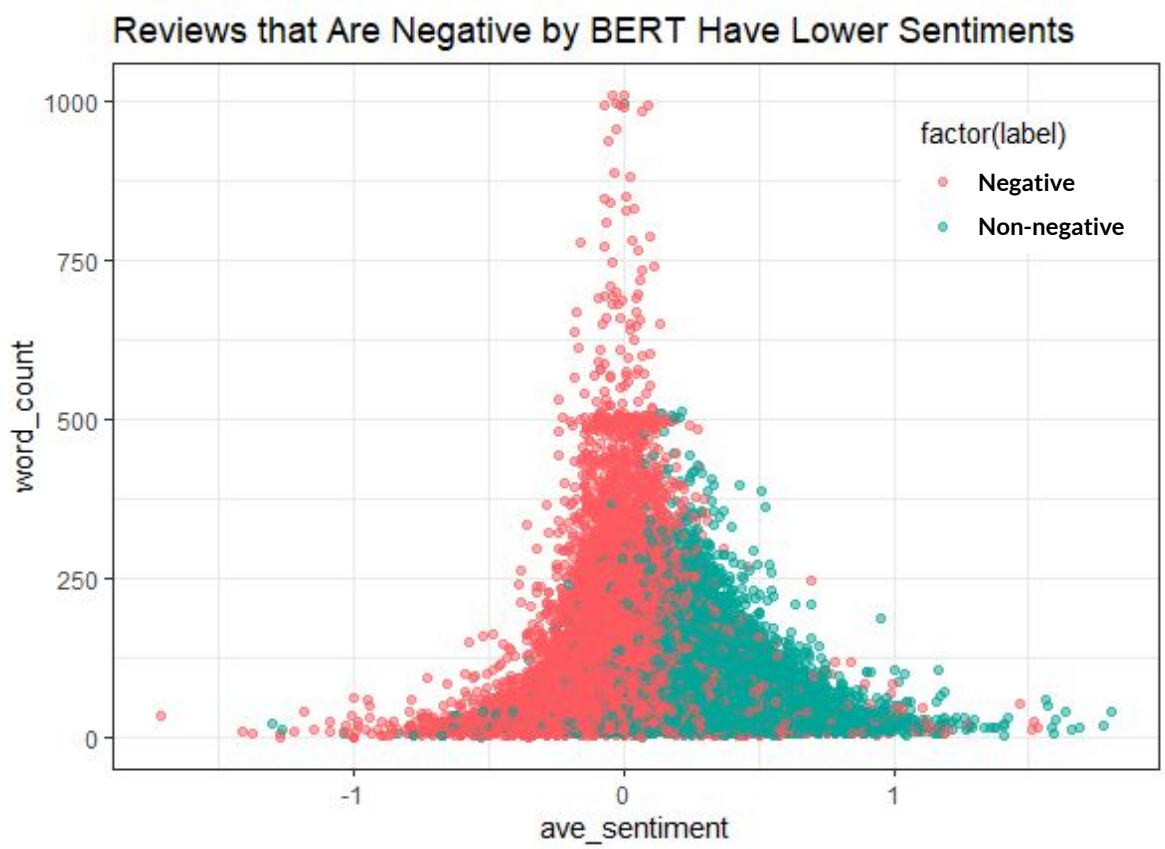
We are not going to consider reviews before that timestamp for this reason.

Using Reviews Score Proxy: Bad Review Predictor

Solution:

- Step 1: Make **dictionary** for the negative words related to each review score topic such as hospitality and cleanliness to detect reviews with negative sentiments
- Step 2: **Filter out potential bad reviews with SQL** on Google Cloud Console
- Step 3: Use 1600 hand labelled reviews to train **DistilBERT + Logistic regression model.**
    - **Label: 0 for reviews that reflects unsatisfying experience; 1 with no unsatisfying experience**
- Final Goal: Predict which potential bad reviews actually reflect negative sentiments due to problems with listings

BERT Labelling Sentiments



DistilBERT + Logistic Regression Accuracy

Sample Data Set Used for Training and Testing	Prediction Accuracy
1 (#obs=500)	0.808
2 (#obs=500)	0.812
3 (#obs=500)	0.864
4 (#obs=500)	0.824
5 (#obs<500)	0.789

CHALLENGES & SOLUTIONS

Challenges: SIZE & LANGUAGES

- The review data contains 1.2 M records and is very demanding for memory and computing time.
- The review data contains languages other than English, but most language processing packages in R cannot process multiple languages in a single dataset.

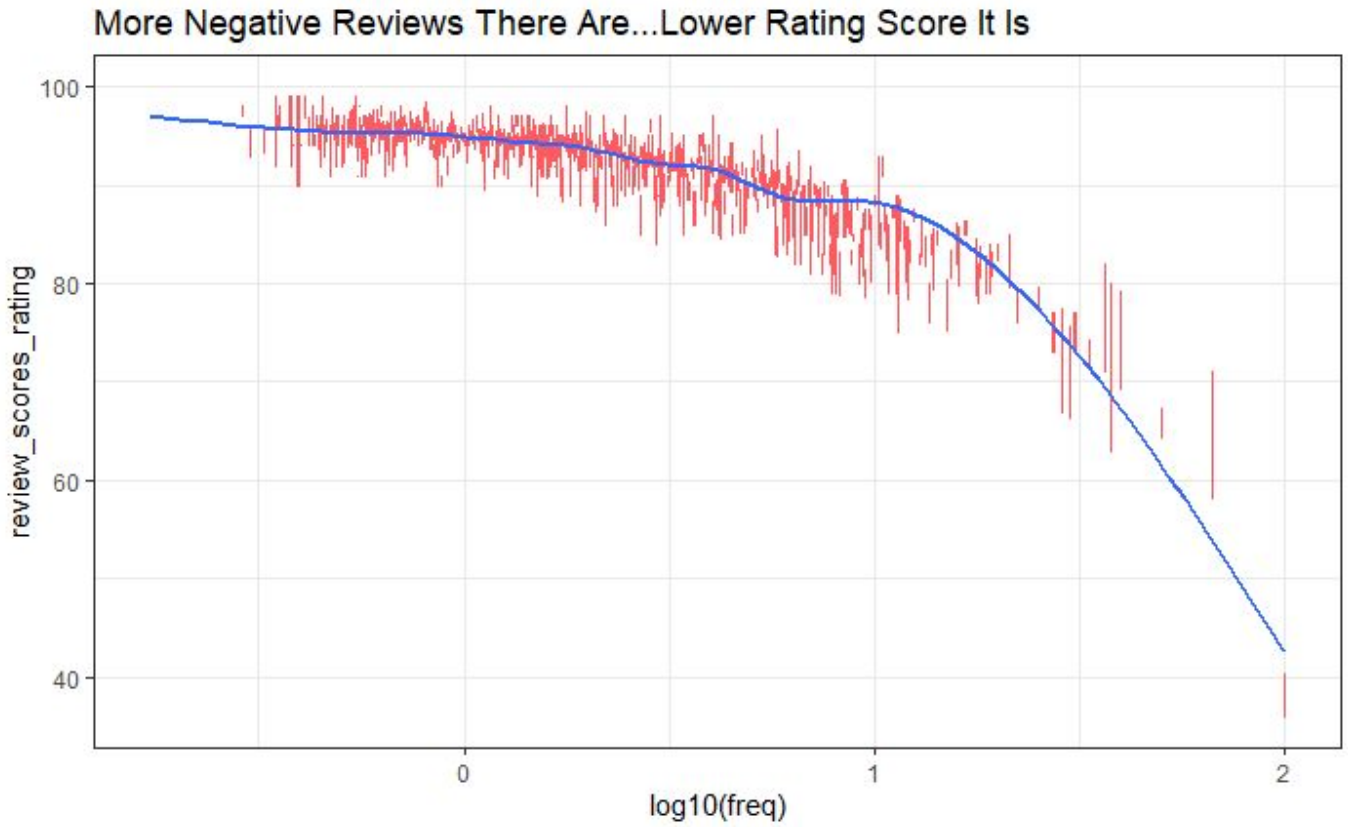
Solutions:

- Splitting the data randomly into 5 pieces and run codes on smaller samples before applying to whole data set.
- Use Google Translate API to sort out the reviews in english

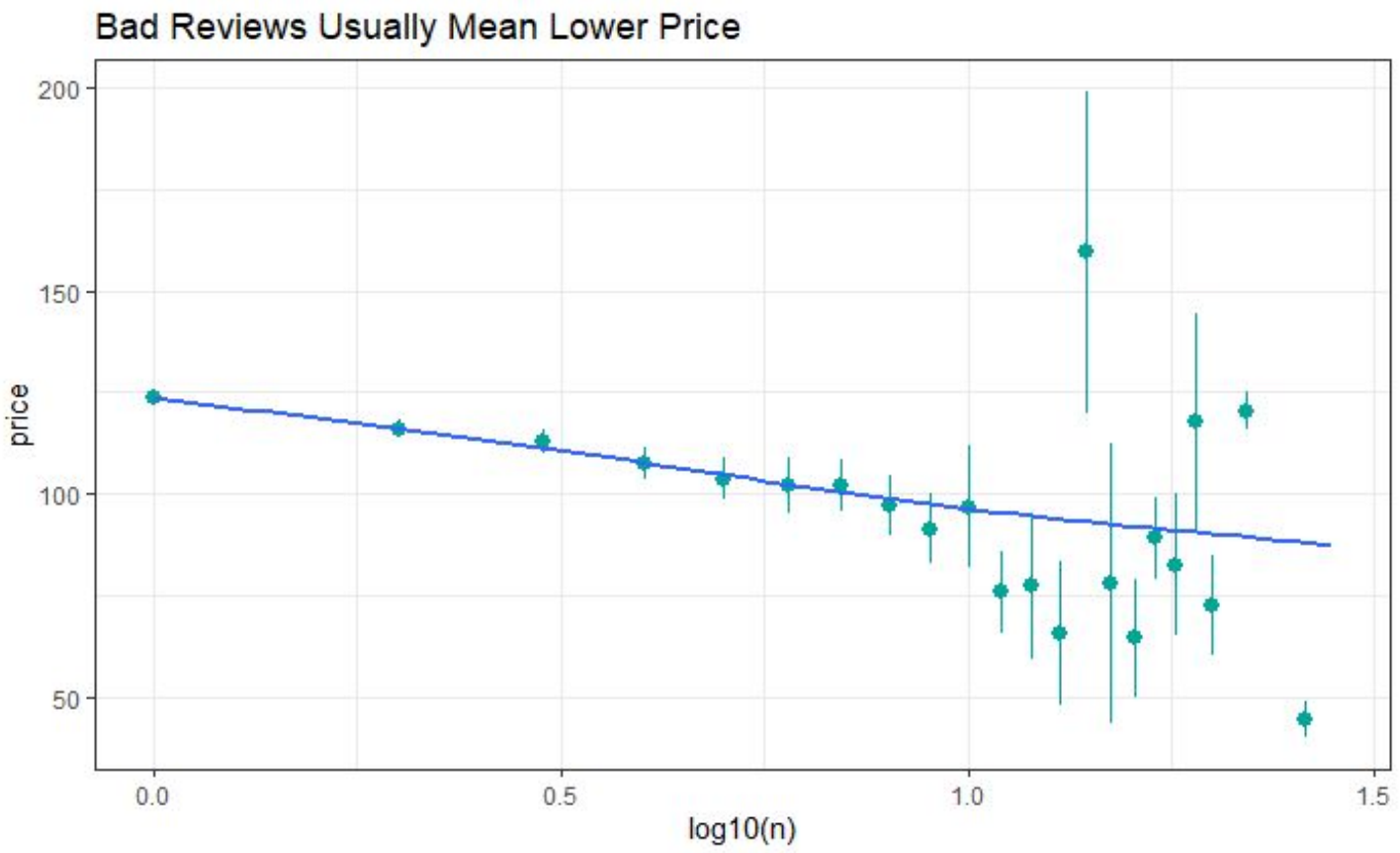
BERT: Good VS Bad

V. Model Ensemble

Negative Reviews Correlates with Lower Score



Negative Reviews Correlates with Lower Price



LASSO: Negative Reviews + , Price -

factor	coef
property_type_Resort	233.50
property_type_Boutique hotel	50.55
neighbourhood_Manhattan	43.48
room_type_Entire home/apt	41.13
bathrooms	23.46
Intercept	15.19
bedrooms	13.72
accommodates	12.47
property_type_Loft	8.99
property_type_Condominium	6.05
cleaning_fee	0.31
availability_30	0.04
review_scores_checkin	-0.17
reviews_per_month	-0.60
review_scores_value	-0.84
n(negative_review_count)	**-1.32**
room_type_Private room	-7.28

Most Consistent Model: LightGBM

Model	Accuracy(RMSE) Training	Accuracy(RMSE) Testing
Linear Regression	69.47	69.75
LASSO	69.38	85.37
XGBoost	57.55	83.19
LightGBM	61.95	60.72

VI. Summary

What We Learn

- Provide better amenities
- Update calendar frequently (for better availability)
- Respond to message promptly (for smooth booking)
- Verify your account (for smooth booking)
- Encourage your guests to write reviews - you may need to impress!
- Review scores matter, but read reviews carefully for improvements
- If you can't do something -- don't promise such things in your description. Many guests are annoyed when they feel hosts are providing misleading information.

VII. Q & A

Q&A

Thank you for listening!

VIII. Appendix

Literature & Acknowledgement

- We want to thank Professor Fradkin for answering some of our questions regarding our project.

- Zervas, Georgios, et al. “The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry.” *Journal of Marketing Research*, vol. 54, no. 5, 2017, pp. 687–705.

1. Pricing model
    a. Use Google Translate API to translate all reviews to English
    b. Use text analysis explore features that could potentially affect price: quantify the details in reviews that are not shown in ratings
        i. Create dictionaries that are similar to those in sentiment analysis. Instead of sentiments, they are going to feature cleaness, location.
        ii. Assign scores/weights to words used to describe the features
    c. Compare our ratings with the original ratings in predicting prices, and then decide with variable to include
2. Fake listing detection
    a. Use the result of our pricing model to detect fake/unrealistic listings
    b. Problem:
        i. No labels available