# Predicting Airbnb prices in US Cities

Riyaz Ali Mohammad

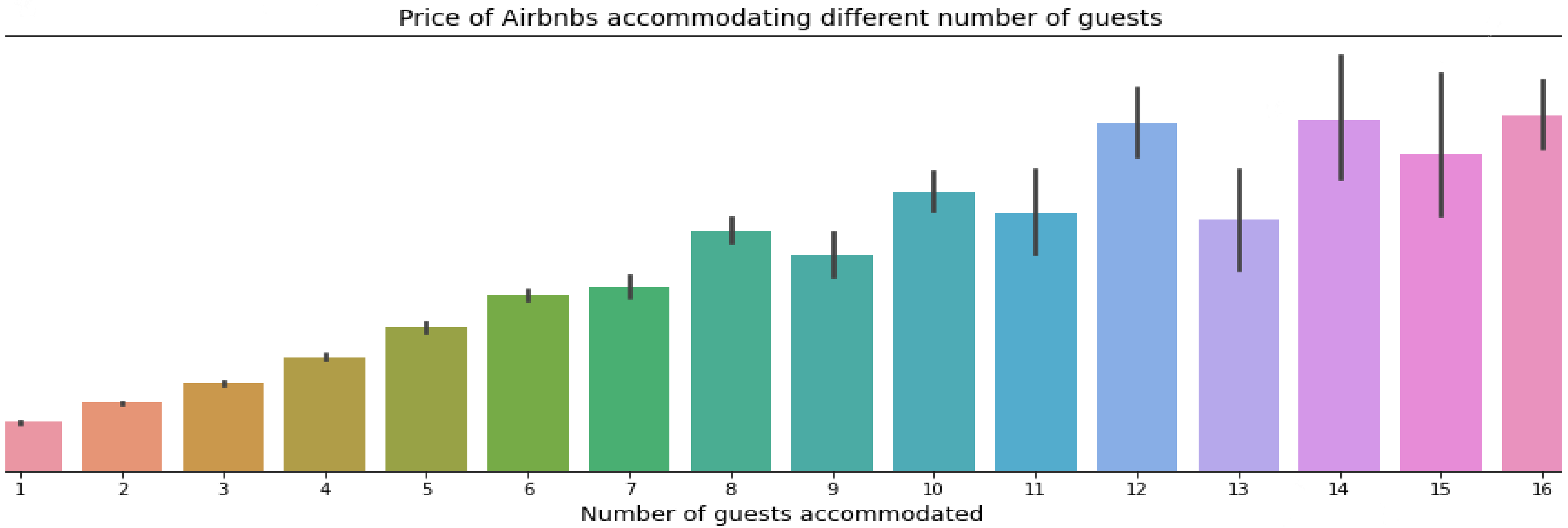
Nov 2020

# Every Airbnb host wants to know the price at which he wants to list his property.

But there is no way to guess the prices that are suitable to both hosts and guests, unless you look deep into the numbers.

# Is there a solution to this problem?

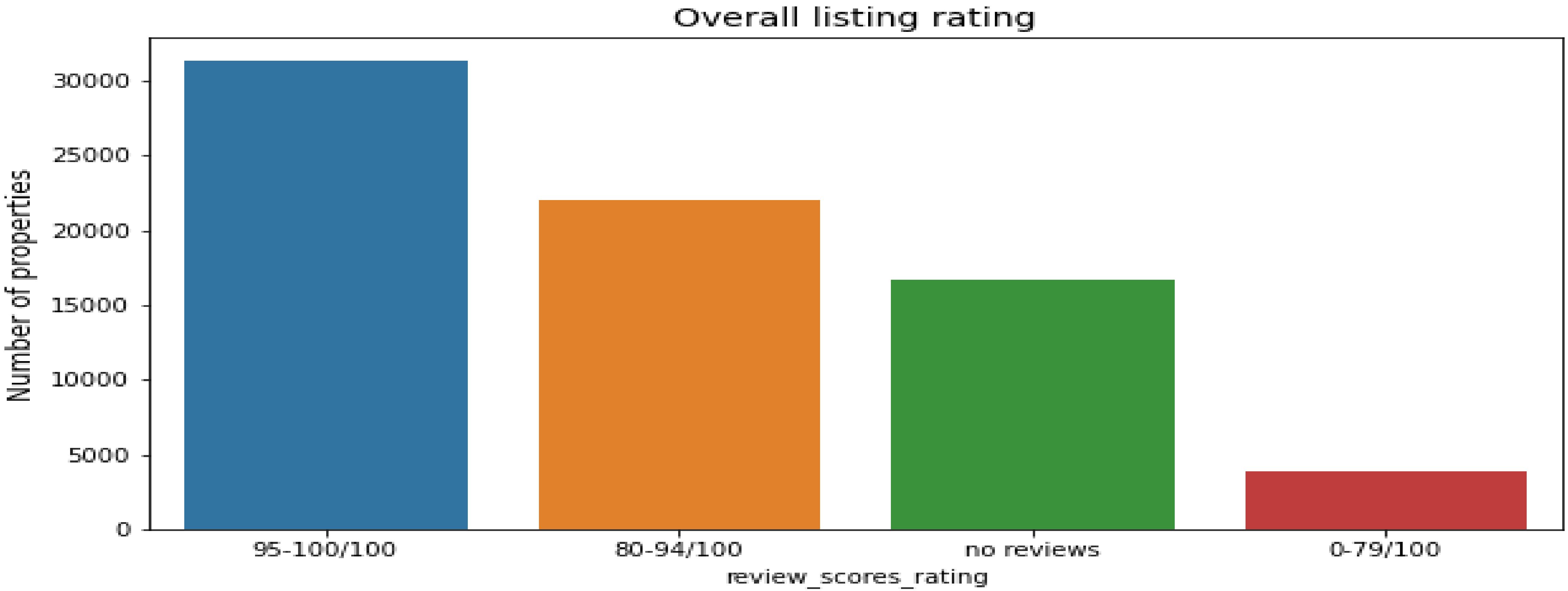
By utilizing the data available about the listings and their corresponding prices, a Machine learning model could be built that could predict the prices that are suitable for hosts and guests.

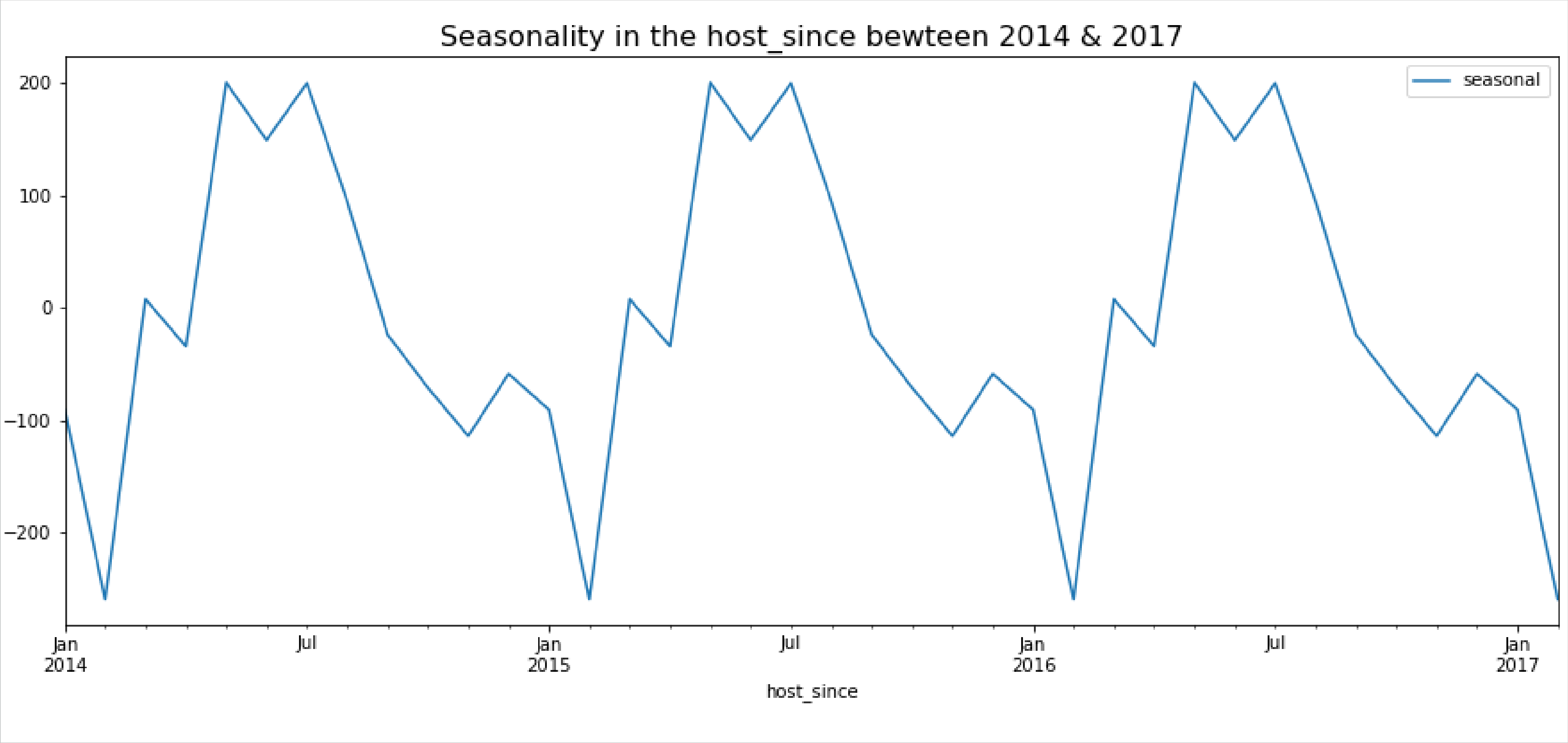


# The more people you accommodate, better rates for your listing!

It’s suggested to allow up to 8 guests. Inviting more members than that, would be futile.

# See how the listings are rated. Make sure the guests love your lisitng too!





Look at the period of the year when the listings are high in demand. Make sure you’re ready by the time summer holidays arrive.

# From the description of amenities available in the data, we have created multiple features to better predict the price.

## List of features created from amenities:

- check_in_24h
- air_conditioning
- high_end_electronics
- privacy
- breakfast
- essential_electronics
- Security
- white_goods
- Elevator
- child_friendly
- Facilities
- hot_tub_sauna_or_pool
- pets_allowed

# We’ve removed some variables from the data that aren’t contributing to the model built.

- Description
- Neighbourhood
- zipcode
- city
- first_review
- last_review
- host_since
- Beds
- High_end_electronics

# The list of models we’ve fit the data to

- Linear Regression
- Ridge Regression
- Decision Trees
- Random Forests
- Gradient Boosting Regressor
- Extreme Gradient Boosing Regressor

# Model that has been selected based upon metrics

Extreme Gradient Boosting Regressor is selected to predict the prices.

Hyper Parameters used in the model are:

- 'n_estimators': [50,100,200,500,700,1000],
- 'max_depth': [2,3,4,5,6],
- 'min_samples_split':[2,5,10,15,20,25],
- 'learning_rate': [0.01,0.1,0.5,1],
- 'loss': ['ls','lad','huber']

# The model performance could be improved by selecting the best parameters available to boost the model

Implemented Randomized search CV to tune the hyper parameters

The best parameters we deduced from the cross validation is as follows:

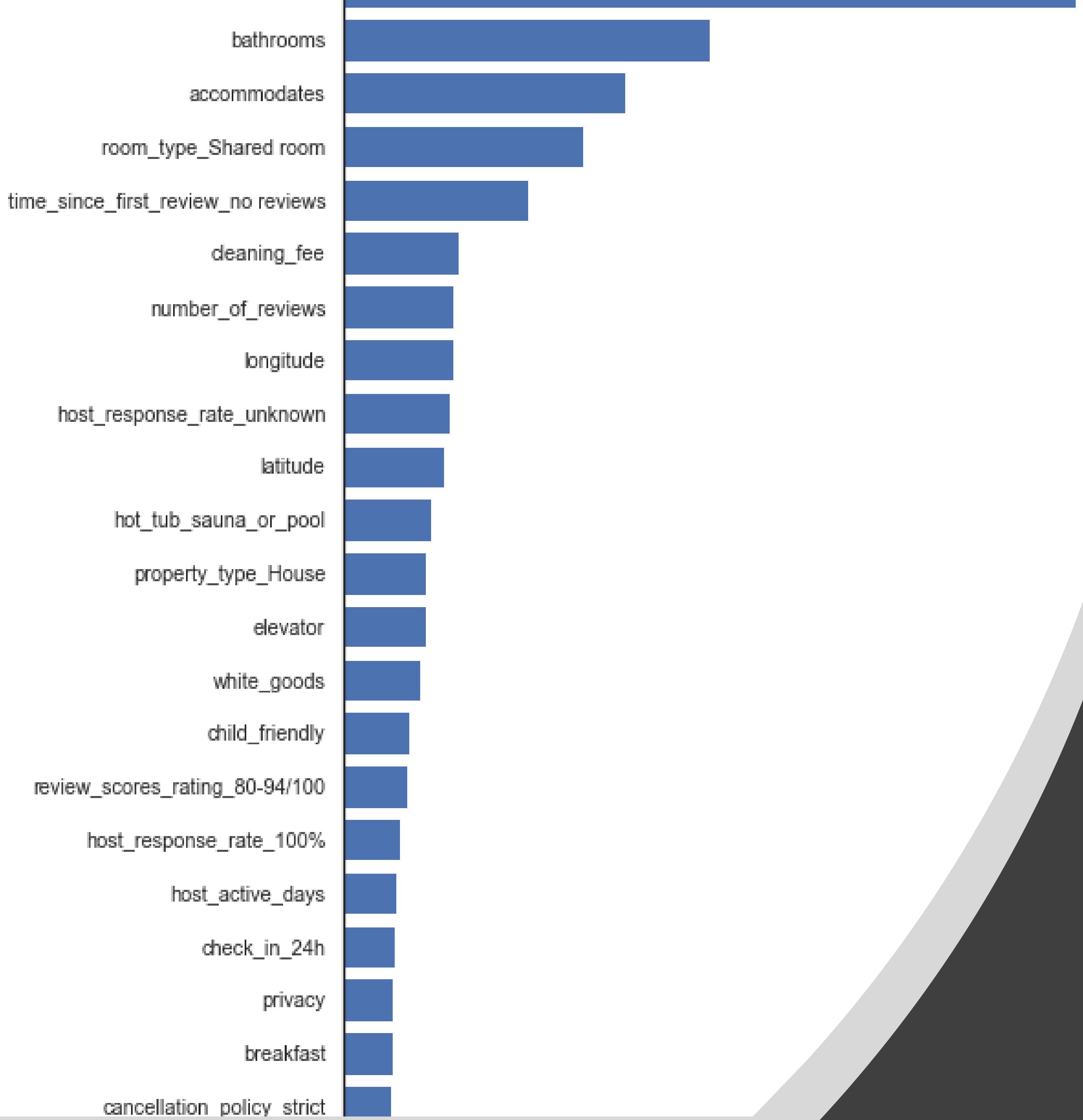
subsample: 0.7

n_estimators: 700

Max_depth: 6

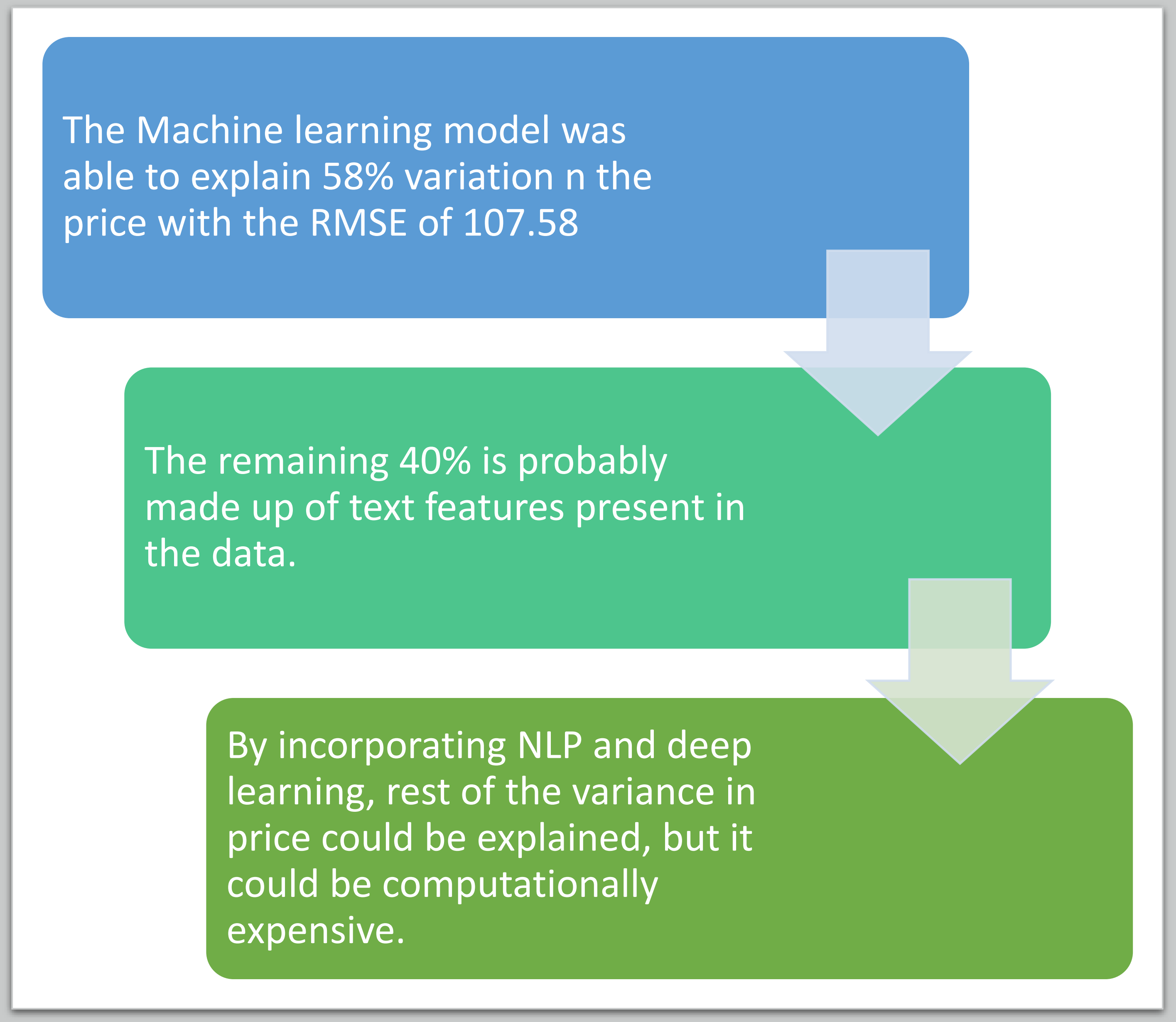
Gamma: 0.7

Col_sample_bytree: 0.7



# Features that contribute to learning are

# Conclusions and recommendations



The Machine learning model was able to explain 58% variation n the price with the RMSE of 107.58

The remaining 40% is probably made up of text features present in the data.

By incorporating NLP and deep learning, rest of the variance in price could be explained, but it could be computationally expensive.