Brett Gustin

IST 407

Final

Air BnB Lodging Quality Analysis through Machine Learning Techniques

The use of mobile applications to deliver accessible and useful lodging has changed the way that humans book and go about trips outside of their home. The app AirBnB provides a wide range of lodging options across the world in order to deliver fast and easy plans for those travelling. Using data that is collected across major cities in the US and respective lodging information, machine learning techniques can be used to analyze and draw from lodging reviews in order to enhance the experience in the time to come.

Introduction

In each of the major cities in America, information is tracked regarding the logistics of the lodge and also data on the host themselves. The goal of analyzing this data centers around first understanding the influence of the host description of the lodge on the overall experience guests had with their stay. This is done through sentiment analysis to understand the description of the AirBnB and how it relates to the overall score or rating given to the lodge. Providing this analysis allows for detection of where description's might not be exactly true in their indication of the experience of the area that they own. This can provide guests with a larger grasp of what to look for when trying to book an AirBnB. Secondly with logistic information about the lodge itself and its surrounding area, machine learning techniques can help determine the most significant variables that influence the price range of your stay along with the overall customer satisfaction. These tools allow for expertise on the subject matter for what is important and valued when preparing for your travel.

Methodology & Data Preparation

The collected information on AirBnB rented spaces and it's logistics come from the Kaggle data source. Some of the main columns include:

- Price(numerical in log form)
- Property Type(categorical form of the lodge)
- Bedrooms, Bathrooms, & Accomodations(separate numerical values)

- City/location(categorical city name)
- Description(text written by the host)
- Host Identity & Response Rate

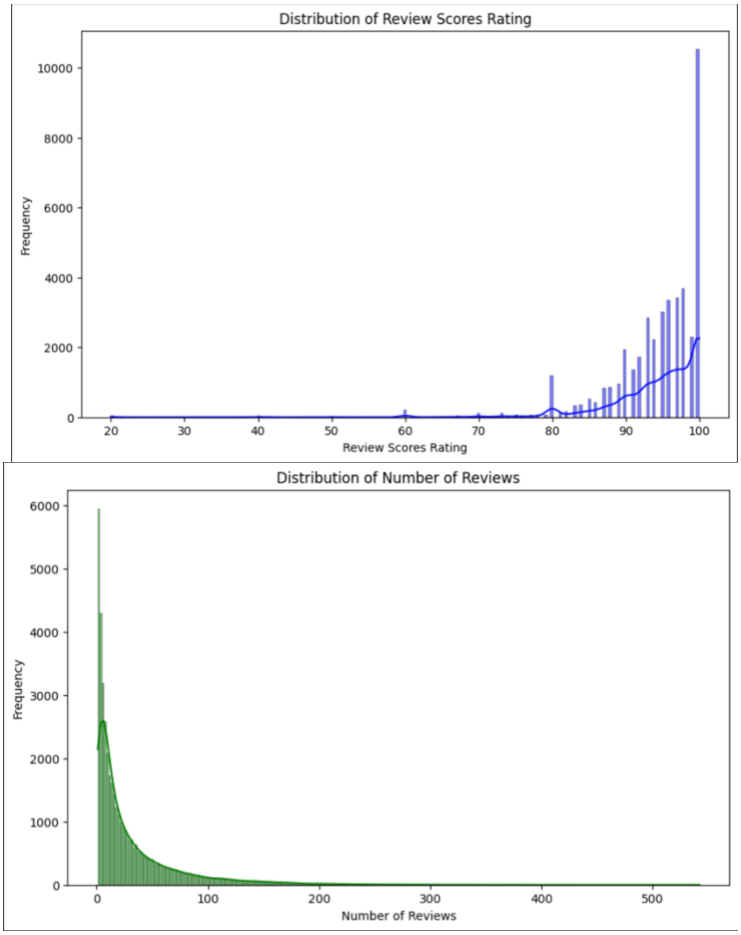
While there are other variables that will be discussed, these are some of the most important throughout the research. The first step was to drop NA rows that were mainly seen in the host_response_rate column and review_score, which is the combined average score of all the reviews left(scale of 1-5 displayed as a whole number 1-100). Next, irrelevant columns that are summed up in other columns are dropped for analysis(latitude, longitude, and zip code are all far too specific and we just want relation to the location as the city). After this, there are over 33,000 rows to be analyzed within the problem set.

Figure 1

While investigating pre-model trends, it was clearly discovered in this bar graph indication of location's influence on price and overall rating. While this will be further analyzed with model results, at first glimpse it is obvious that while the city of San Francisco has the priciest AirBnB options, they receive the highest average review score in relation to other cities. While at the same time, Boston has some of the second highest average price for their residencies, but the worst easily the worst average review scores. Further analysis provides explanations to these surface level results.

The first step in sentiment analysis of the reviews written by the host of each lodge is to consider what determines a good or bad review. The column review_scores_rating provides the average score given by the customer and their experience, however there is a large class imbalance that needed to be addressed prior to modelling.

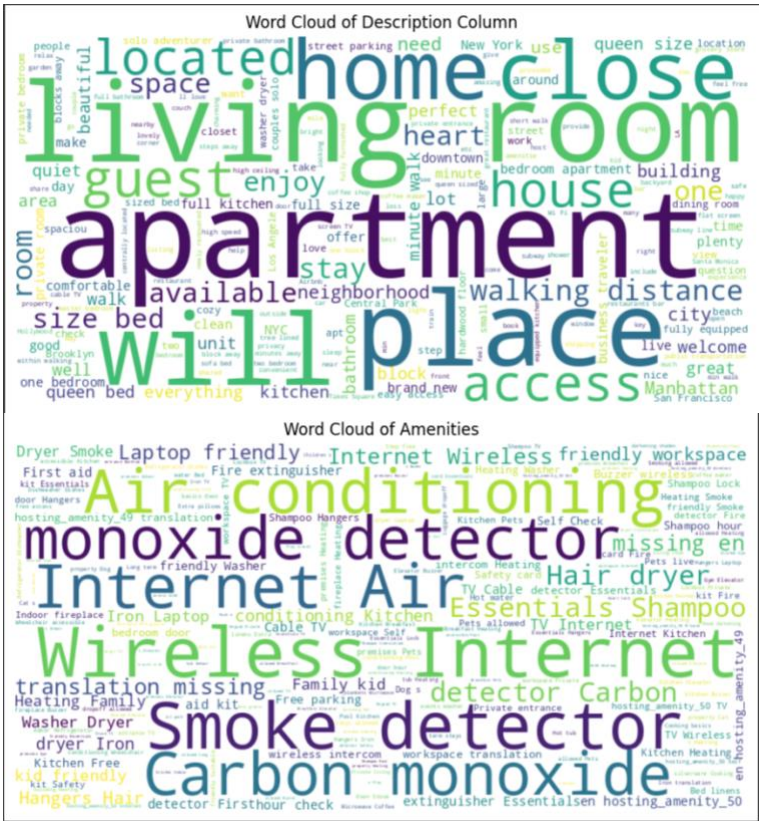
Figure 2



These two variable distributions represent the relationship between one another. Review scores rating is heavily right skewed, while the number of reviews is heavily left skewed. It can be concluded that a lot of the high review scores are inflated due to the lack of total reviews for that AirBnB. In other words, an AirBnB needs to have more than a small number of reviews in order for it to be considered a reasonably unbiased review. For this, the data is filtered to only allow for review scores with at least fifteen reviews, to ensure that it is a valid score. This eliminates a majority of the class imbalance in the target variable of review score. Still, it is necessary to go forward with an indication that the review score ratings are still right skewed after dropping

invalid results. This means that customers still score lodges higher than they would lower as it relates to an 100 point scale. Techniques Multinomial Naïve Bayes and Standard Vector Machine are used in the first part of modelling sentiment analysis and are tuned in order to account for the smaller class imbalance, as seen later. The target variable of review score is created into a binary classification target that represents if the review is higher or lower than 92. At this score, it is necessary to note that there are still more reviews greater than this value than lower. Therefore, this target variable is adjusted to not define just good or bad review, **but is considered the threshold of an outstanding review compared to what is expected as a score**. Keeping this definition allows for proper sentiment analysis while understanding that a lot of reviews that range from 85-92 are not necessarily bad reviews, but do not stand out as much as the others that were scored higher. After creating our enhanced target variable, the text description that is used to identify sentiment is analyzed through word clouds.

Figure 3



At first glance some of the most common words associated with both the description of the lodge and amenities are the most frequent in the cloud. Apartment, living room, and house all seem reasonable in relation to city lodges for AirBnB, while internet and air conditioning are common

amenities. However, there are underlying descriptive words used in high reviews that indicate what hosts are trying to sell to the customers. Words in the description such as access, home, close location, walking distance, available, and heart all specify having easy access to the options surrounding the lodge. Within cities, this makes sense as ideas that customers care the most about. While for amenities it may seem clear, but people care a lot about smoke and carbon monoxide detectors, especially in high rise apartments in the city. Since we know the area and location of these lodges, inferences can be made about what matters most to customers, which is seen also in further modelling.

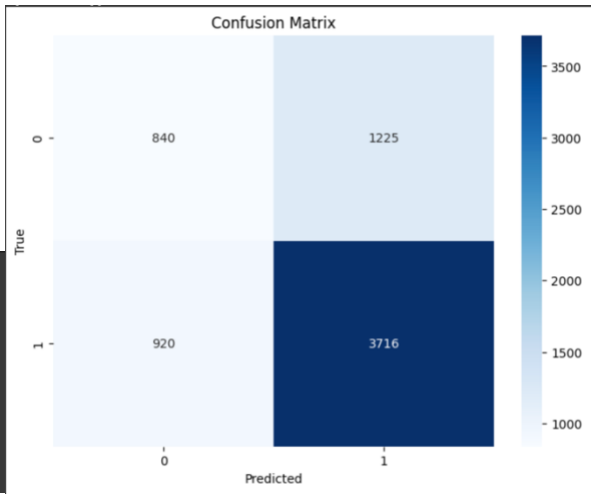
For host text description analysis, SVM and MNB models are compared to provide the most efficient algorithm for understanding differences in an overperforming lodge. Vectorization methods are compared to provide the greatest insight between Boolean and TF-IDF vectorization. TF-IDF usually can have more accurate results when translating word text into frequencies, since it takes into account document length and word frequency in relation to a corpus. However for our modelling it was found that simple Boolean vectorization was the best implementation for MNB and SVM. For the second part of analysis which accounts for more of the logistic variables of the AirBnB in relation to customer experience and pricing dependencies, some variables have to be converted from categorical to numerical. This includes one hot encoding variables such as city, property type, and cancellation policy into binary and numerical columns of their own. This allows for these influential variables to be included in machine learning algorithms to be further analyzed including regression, decision trees, and XGboost.

Results- Detecting Review Sentiment

Multinomial Naïve Bayes allows for text mining of word frequencies with the assumption of the text frequencies being independent of one another. Once again, Boolean vectorization is used and the trained model of the description text is compared to test data to determine binary value for if the review is over exceeding expectations or not in relation to “what is advertised”. The test data can provide predictions that are used to measure the accuracy of the model through precision, recall, and F1-score.

Figure 4- MNB Sentiment Detection Model

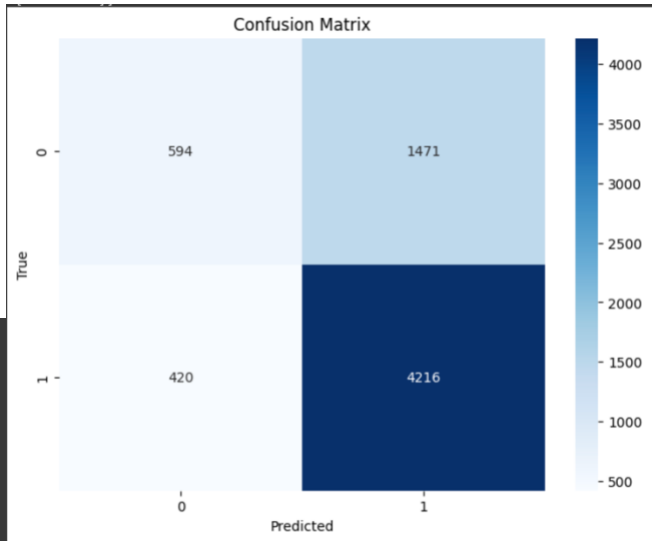
```
Classification Report:
              precision    recall  f1-score   support

       False       0.48      0.41      0.44      2065
        True       0.75      0.80      0.78      4636

    accuracy                           0.68      6701
   macro avg       0.61      0.60      0.61      6701
weighted avg       0.67      0.68      0.67      6701
```



As seen in this first model detecting an overwhelming review compared to average scores, MNB does a good job in predicting the true positive cases. This means the model does well predicting an overwhelming review over the average review more often than not. However as seen with the recall and precision values for false predictions, the model is struggling with false positives. This means a review that is not above an average review is getting predicted with an overwhelming review. The F1 score weighs these two metrics and determines that the false predictions are not nearly as accurate as the positive predictions. SVM modelling is compared to this technique in order to try and associate negative reviews with a different approach.

Figure 5- SVM SMOTE Tuned Sentiment Detection Model

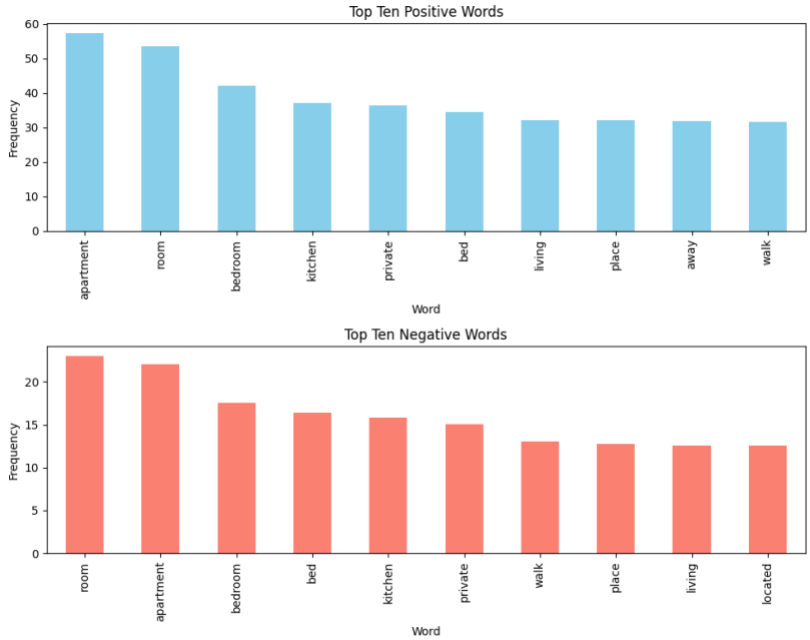
```
Classification Report:
              precision    recall  f1-score   support

       False       0.59      0.29      0.39      2065
        True       0.74      0.91      0.82      4636

    accuracy                           0.72      6701
   macro avg       0.66      0.60      0.60      6701
weighted avg       0.69      0.72      0.68      6701
```



In this model, SVM turns text into numerical feature values, that are separated by an optimal hyperplane to distinguish classes. SMOTE is a method used to tune the model to oversample the

minority class of non over exceeding reviews. As seen in the model results, the overall accuracy has increased however this does not indicate that SVM still did a better job in predictions. As seen with the lower recall and f1-score with false predictions, the model still struggles with over predicting overwhelming reviews. After creating a higher threshold for an overwhelming review to 95, the model struggled even more to decipher between descriptions of lodges that scored in the 90 range. This confirms that the best threshold is at the 92 rating, which the model still struggles to connect overwhelming reviews to. This is because of the similarity of text that hosts use to advertise their lodge for the customer experience. This can be seen in the most frequent words associated with overwhelming reviews compared to frequency in reviews below what is most highly desired.

Figure 6



Even after stopwords are removed from the text analysis it is clear why the model is struggling to predict a majority of false positives from the overwhelming reviews. Most of the frequent words are similar in the top ten, indicating that hosts learn from top descriptions in what best to advertise. Still, words such as private, walk, located all indicate similar concepts that location and access to other aspects of the city beyond the lodge is what is most desired in these descriptions.

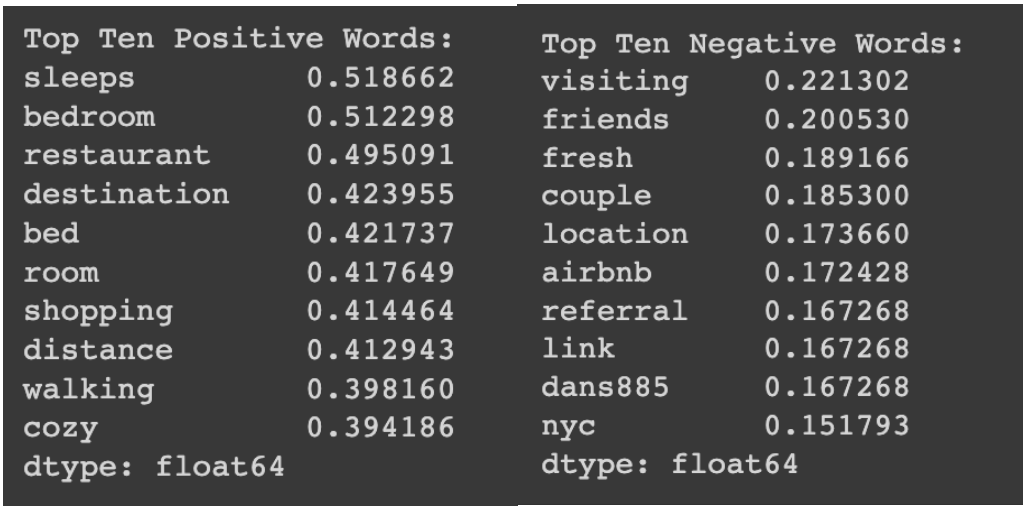
True Positive: "This apartment opens up into the bright and spacious open living room/dining room area. The living room has a 55 inch TV with ROKU streaming."

False Positive: "The common space is huge, bright, with high ceilings and shiny hardwood floors."

True Negative: "A turn of the century Victorian - close to the 101 and 280 freeways."

These examples of cases from predictions in the test data also help determine the struggles in the model for distinguishing reviews. Both of the top two statements are similar with their definitions including information about living room areas and their brightness, so the model associates the two. However, one is a below overwhelming review and not picked up as one where as the true negative that seems dull in its statement is predicted with how it was reviewed. After investigating these analysis, the SVM model is sampled with just one thousand rows for the most frequent words for the review score split.

Figure 7



```
Top Ten Positive Words:
sleeps          0.518662
bedroom         0.512298
restaurant      0.495091
destination     0.423955
bed             0.421737
room            0.417649
shopping        0.414464
distance        0.412943
walking         0.398160
cozy            0.394186
dtype: float64
```

```
Top Ten Negative Words:
visiting    0.221302
friends     0.200530
fresh       0.189166
couple      0.185300
location    0.173660
airbnb      0.172428
referral    0.167268
link        0.167268
dans885     0.167268
nyc         0.151793
dtype: float64
```

This sample indicates that the top overall words are not necessarily associated because of their frequency, but because of their similarity in that class. However, since they are similar words this is where the model can get confused for the reviews that are close to the threshold. Still the most influential takeaway from sentiment analysis lies with observations from the smaller sample size. Words such as walking, distance, cozy, and shopping all indicate that the surrounding area of the AirBnB matters. If areas are easily accessible around the lodge, this leads to higher reviews. Meanwhile, visiting, referral, or friends might be descriptions that are trying to overlook some downfalls of the lodge. These conclusions make sense in relation to the location of most of the

highest reviews(San Francisco, LA). Cities where it is not only warmer out but easier access to walking and activities around the lodge in the description helps lead to a higher review.

Results- Lodge logistics Economic Influence on Customer Satisfaction

Through classification machine learning techniques, models can determine the influence of information on the AirBnB and the host performance on both pricing and reviews of the lodges. First, linear regression is run on the continuous target variable of price.

Figure 8

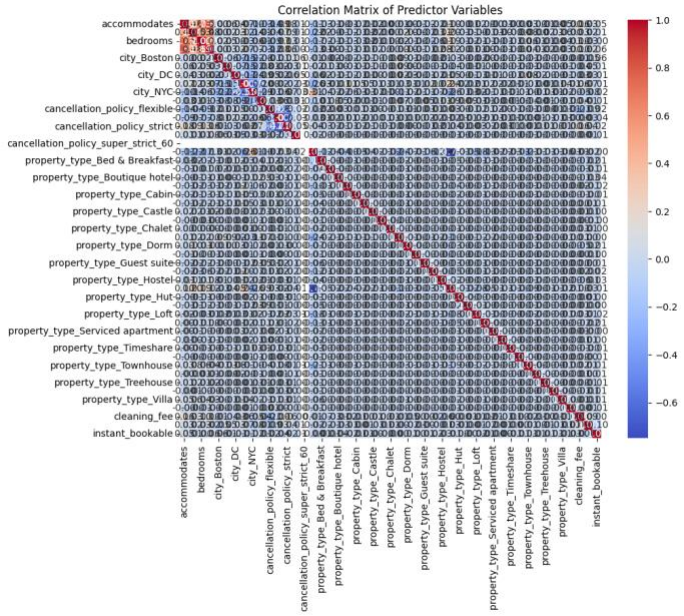
OLS Regression Results

Dep. Variable:	log_price	R-squared:	0.444
Model:	OLS	Adj. R-squared:	0.443
Method:	Least Squares	F-statistic:	475.6
Date:	Wed, 24 Apr 2024	Prob (F-statistic):	0.00
Time:	15:27:18	Log-Likelihood:	-18511.
No. Observations:	26802	AIC:	3.711e+04
Df Residuals:	26756	BIC:	3.749e+04
Df Model:	45		
Covariance Type:	nonrobust		

Feature	Estimate Coefficient	P>|t|
property_type_cabin	-.13	.175
property_type_vacation_home	.52	0.056
property_type_house	-0.05	.140

Interpreting the R squared of the model shows that 44% of the dependency in variables on price of a lodge is identified and accounted for in the model. Some of the most important features on pricing were in the type of the property. All of these were in the p-level significance of 5-10%, to indicate statistical significance. A one unit increase of the lodge being a cabin or house have a negative influence on price, while a vacation home associated to a positive influence. This makes sense for house to have a small negative influence since a lot of the pricier lodges in these big cities are in high rise apartments, and houses would be located farther outside the city and for cheaper.

Figure 9



This correlation matrix indicates that there is some above expected correlation outside of the diagonal of features that would suggest multicollinearity. To account for this, a linear regression was once again run with property type excluded and bedrooms and beds columns combined.

Figure 10

OLS Regression Results

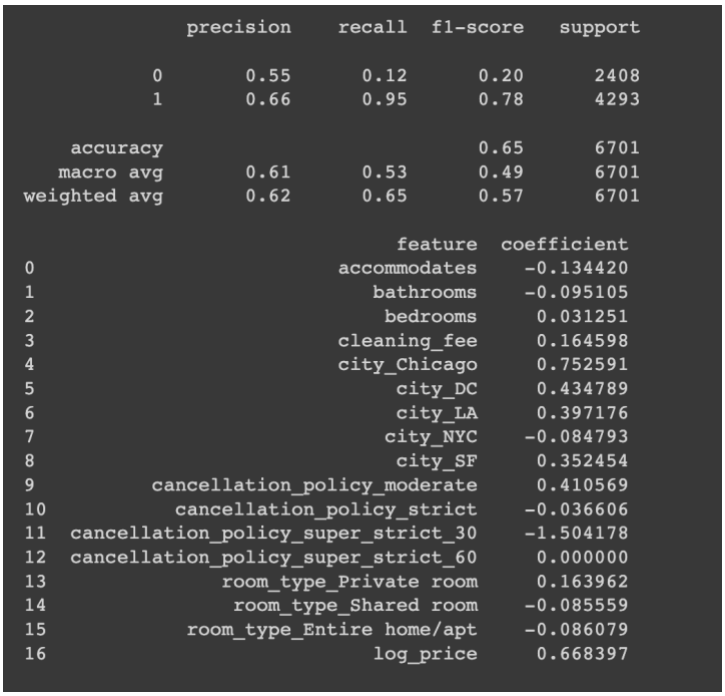
Dep. Variable:	log_price	R-squared:	0.604
Model:	OLS	Adj. R-squared:	0.604
Method:	Least Squares	F-statistic:	2556.
Date:	Wed, 24 Apr 2024	Prob (F-statistic):	0.00
Time:	16:06:15	Log-Likelihood:	-13964.
No. Observations:	26802	AIC:	2.796e+04
Df Residuals:	26785	BIC:	2.810e+04
Df Model:	16		
Covariance Type:	nonrobust		

	Coefficient	Error	t	P>|t|
city_Boston	0.4403	0.009	47.518	0.000
city_Chicago	0.1342	0.009	14.385	0.000
city_DC	0.3151	0.009	36.209	0.000
city_LA	0.2913	0.005	53.358	0.000
city_NYC	0.4168	0.005	88.273	0.000
city_SF	0.6886	0.008	90.282	0.000

These regression results show that the model's accuracy increased when tuning the model, shown with the increased R-squared. The city variables that were one-hot encoded have extremely low p values, indicating that there is collinearity in the ability to predict the price with the location of the lodge. Still, the coefficients demonstrate that the cities such as Boston and San Francisco that were seen to have the highest priced lodges, have the largest positive influence on our price model.

Next, logistic regression is ran with the target variable switched to the binary overwhelming review column.

Figure 11



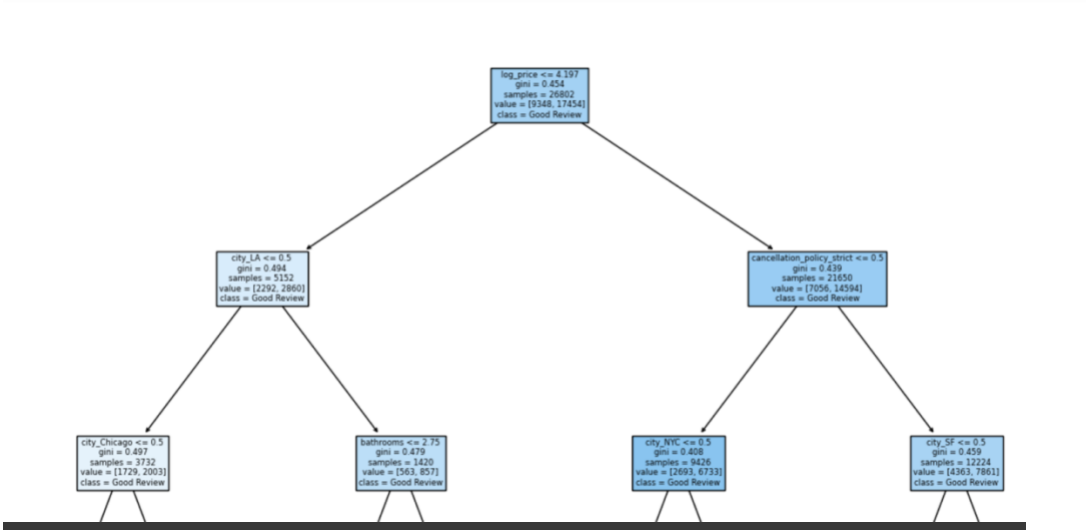
	precision	recall	f1-score	support
0	0.55	0.12	0.20	2408
1	0.66	0.95	0.78	4293
accuracy			0.65	6701
macro avg	0.61	0.53	0.49	6701
weighted avg	0.62	0.65	0.57	6701

	feature	coefficient
0	accommodates	-0.134420
1	bathrooms	-0.095105
2	bedrooms	0.031251
3	cleaning_fee	0.164598
4	city_Chicago	0.752591
5	city_DC	0.434789
6	city_LA	0.397176
7	city_NYC	-0.084793
8	city_SF	0.352454
9	cancellation_policy_moderate	0.410569
10	cancellation_policy_strict	-0.036606
11	cancellation_policy_super_strict_30	-1.504178
12	cancellation_policy_super_strict_60	0.000000
13	room_type_Private room	0.163962
14	room_type_Shared room	-0.085559
15	room_type_Entire home/apt	-0.086079
16	log_price	0.668397

These logistic regression results once again correlate to strong true positive predictions of overwhelming reviews, but struggling with false positives since the reviews can be so similar. Coefficients indicate that strict policy and the entire home or shared room compared to private room have a negative unit influence on a better review. Since the city variable is a dummy, it is analyzed in relation to the city of Boston. This shows that all cities have a positive unit increase on a better review other than NYC.

Figure 12

```
Best Parameters: {'max_depth': 7, 'max_features': 'auto', 'min_samples_leaf': 15, 'min_samples_split': 12}
```

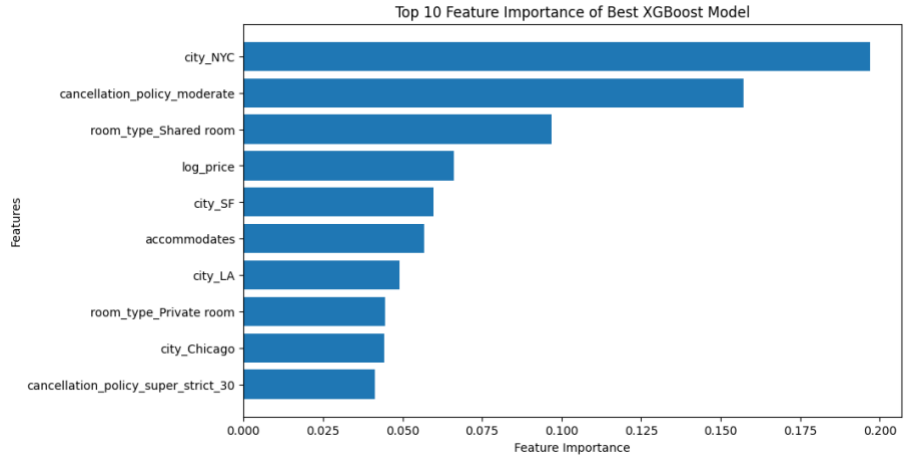


	precision	recall	f1-score	support
0	0.52	0.07	0.12	2408
1	0.65	0.96	0.78	4293
accuracy			0.64	6701
macro avg	0.58	0.52	0.45	6701
weighted avg	0.60	0.64	0.54	6701

A decision tree is modelled and tuned with the hyperparameters listed to try and identify the highest information gain from features that relate to a better review, our categorical variable. The price is the root node with the highest information gain, therefore inferring the most influential feature for this model. The cancellation policy and city are other variables in the top of our decision tree, having a higher impact on our target. Almost all of the cities are in the top of the tree, growing on the conclusion earlier of city having high correlation to what the review or price will suggest. The predictions created by the decision tree towards analyzing the best reviews or not struggles with the middle 90 level reviews as they have common feature values to the reviews that are just above the threshold. The recall indicates the false positives once again overtake the models performance when it is very good at predicting the necessary positive cases.

Figure 13- XGBoost

	precision	recall	f1-score	support
0	0.61	0.61	0.61	3438
1	0.59	0.58	0.58	3263
accuracy			0.60	6701
macro avg	0.60	0.60	0.60	6701
weighted avg	0.60	0.60	0.60	6701



From our decision tree, XGBoost is applied to have the model learn from weaker trees and its respective errors, creating randomized sampling of the features that create a random forest. Tuning this model for parameters including estimators, learning rate, amount of trees and their depth all create the most accurate model seen at making predictions toward a better review. The Recall and F1 score of lower reviews show a much higher capability of recognition within predictions. Some of the most important feature variables include location of cities such as NYC, SF, LA, and Chicago, along with the cancellation policy, price, and room type. Having a moderate cancellation policy had large importance in separating the target variable.

Conclusions

Overall, the classification models were compared in order to build the most accurate prediction models on lie detection in host descriptions of lodges, as well as the review of the lodge with respect to its makeup and logistics. Despite pre processing to drop inflated reviews, the models performed well in identifying the highest reviews in relation to both the description and logistics, however struggled to identify some of the variation in reviews that were very close to being overwhelming. This indicates that hosts are not trying to “reinvent the wheel” when it comes to advertising their lodge. Very similar concepts and text are used when defining the lodge to customers, and in turn creates a cycle of expectations and higher reviews. Some of the most

influential aspects of higher review scores include the location, and having access to walkable activities outside the lodge. Customers want easy access to comfortable measurements of safety if they are in an apartment, and the ability to freely walk around the area and participate in activities without travelling too far. This is why cities such as San Francisco stand out in comparison to Boston, which is much more spread out. Air BnB's are not only seeing the reviews and prices higher in warmer areas, but also where spatially for the city the lodge makes everything easier to travel around. How a host treats cancellation policy and the setup of the room also have relation to how customers feel after their stay. These models help create business decisions on the safety of the lodge and what customers expect and desire out of their stay and surrounding area.