Estimating Feature Importance in Airbnb-listings Clusters

Darshan Panse
Computer Science
Northeastern University
Boston MA USA
panse.d@husky.neu.edu

Shantanu Kawlekar
Computer Science
Northeastern University
Boston MA USA
kawlekar.s@husky.neu.edu

Abstract

With rents on the rise in many of America's cities it has become increasingly popular for some residents looking to earn some side money to invest in short-term rental market. Nowadays people are doing rental business using Airbnb by hosting guests at their properties. These hosts fall into two categories: the ones who are looking for places across U.S. to list new properties and want to know what features of the listings are important to those locations and thereby maximize profit and others who already own a property at a certain location and want to improve the chances of getting picked by guests.

The Airbnb-listings data is skewed, i.e. the review scores for most of the houses are similar and high. The problem got reduced to finding values of the features which would yield low review scores and flagging those feature values.

We clustered listings based on their features, leaving the review scores aside and found that there was always one cluster which would have lower review scores as compared to others. In this paper we have mentioned our approach in detail and have summarized our findings.

KEYWORDS

Soft clustering, hard clustering, PCA, GMM, agglomerative clustering, k-means, Airbnb, listings, cluster analysis, cluster profiling, feature importance.

1 Introduction

Online peer-to-peer marketplaces have been proliferating in recent years [3]. These marketplaces match sellers who want to share underutilized goods or services with buyers who need them [5, 11]. Some of the examples include Uber, Lyft, TaskRabbit. One such peer-to-peer application for accommodations, Airbnb, has grown rapidly in the last few years. Airbnb began in 2007 and has rapidly grown to include listings over three million available properties in more than 65,000 cities around the world.

Householders can become small business owners and take some of the rental burden off their shoulders. Lately, a lot of people have been doing mainstream business on Airbnb. They buy apartments and houses across U.S. to put up as a listing on Airbnb. These sellers are known as hosts and the people interested in staying in these apartments are known as guests.

One important question that hosts need an answer for is what kind of features their listings must possess in order to be picked by guests with a higher probability. There are a ton of services out their which recommend the prices of the listings in an area or based on the features of their home. Although listing price has a substantial impact on the probability of a listing getting picked but it is not the only factor affecting it.

The review scores data is highly biased towards high scores. It is even more biased than the Yelp dataset [8]. That said, instead of finding feature values which lead to high scores, we find feature values which lead to low scores and flag the values of those features.

To this end, we present a data driven study on Airbnb listings data, focusing on the factors which lead to a good review score by guests rather than only price. We paint a more complete picture by first looking at places across U.S. where most of the listings are concentrated geographically so that geographical areas can be recommended to hosts to buy new houses. We then pick up a single concentrated geographical area and work with listings falling in that area. We put groups of features of these listings under the lens, one group at a time and summarize what feature values lead to low review scores.

Our analysis reveals that when we cluster the listings based only on the listing features, leaving the review scores apart, using Gaussian Mixture Models, we always get a cluster with low review scores compared to other clusters. This result suggests that the feature values for the dominant features of that cluster affect the review score for that cluster.

We believe that our work significantly advances the current understanding of how to get good review scores, in other words, how to avoid bad review scores on Airbnb. This would ultimately, lead to a listing being picked by the guests with a higher probability. Our key contributions in this work are:

- We find geographical concentrations of Airbnb listings.
- We cluster listings based on different groups of features like continuous, categorical, amenities, property features, host features etc.
- We investigate if certain clusters have low or high review scores compared to other clusters and find features dominating a clustering.
- We perform cluster profiling and analyze what feature values lead to low review scores for the cluster.

2 Related Work

Most of the work regarding analyzing Airbnb data revolves around prices of the listings and presenting statistical facts. Ruchi [1] has performed two analyses on Airbnb data finding what causes difference in prices of listings and where to invest a property in Boston to get maximum return on Airbnb but her results are driven by the prices of the houses. Tameemi [2] has answered questions like what factors affect the price the most and can the price be accurately predicted given other information about listing. Most of the work out their neglects the impact that review scores have on the probability of a listing getting picked. Our analyses reveal that certain feature values directly or indirectly affect review scores.

3 Background Information

We form the clusters based on listing features. Once the clusters are formed, we profile those clusters based on approaches proposed by Inna [4]:

- **Empirical Approach-** We join cluster label to the features we are interested in and plot each cluster's average scaled value. The feature of interest is on x-axis with the average scaled value of that feature on y-axis.
- **Centroid Approach-** We figure out which data point is closest to each cluster centroid. These data points are considered as the representative points of each cluster. We then look at the values of the features of interest for the cluster which has low or high review scores.

4 Proposed Approach

To find out what feature values affect the review scores, we propose the following approach:

1. Pick latitudes and longitudes and plot them on the map of U.S. This will yield the geographical concentrations of the listings across U.S. We are interested mainly in densely concentrated areas.
2. Pick listings falling in a certain densely concentrated geographical area. Analyzing the listings in a certain area yields better results than analyzing them across entire U.S. possibly due to changing trends across states.
3. Group features based on the following classification: continuous, categorical, representing amenities and about property. This is done to put set of features separately under the lens and find the importance of those features.
4. Preprocessing: perform exploratory analysis, standardization etc. Bringing the features down to same scale yields better results and ensures the convergence of the clustering algorithm.
5. We use Principal Component Analysis for dimensionality reduction. This yields us features which capture the maximum variance and drops the rest of the features.
6. For each group of features, perform clustering (k-means, GMM, agglomerative, spectral whichever yields better results for step 5) with and without dimensionality reduction. The review scores are not added to the feature set for clustering.
7. Once the clusters are formed, compare the average review scores for value, location, communication, rating, cleanliness and accuracy for each cluster. See if a certain cluster has a high or low average score for any of the above review scores.
8. Look at the projections of features onto the principal components which preserves more than 80% of the variance.
9. Pick relevant features from them and apply the empirical approach to profile the cluster with low average review score.

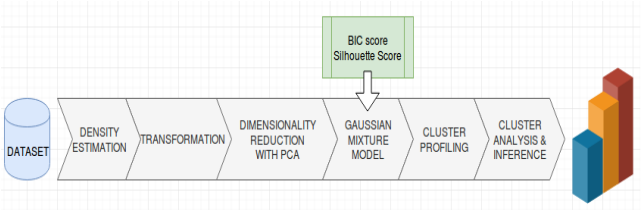

Figure 1: Explains the flow of the experiment.

5 Experiments

Our dataset has about 89 features and most of the features are right or left skewed depending on their significance. A dataset like this needs careful attention so that features which are important to this domain are preserved along with their skewness. To tackle the skewness and to scale them onto same scale we normalized the data by trying various transformations like box cox transform, z score, power transform, min max and a combination of some as box-cox with minmax to get normalized data with values between 0 and 1 [7]. We used histograms to plot the individual columns data and observed its distribution to plan accordingly. The review score data is skewed.

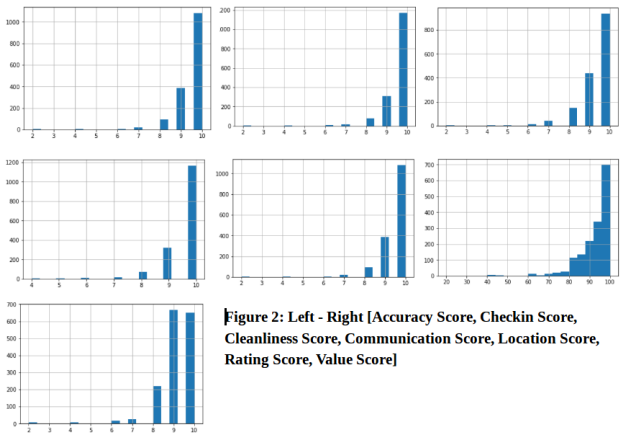

Figure 2: Left - Right [Accuracy Score, Checkin Score, Cleanliness Score, Communication Score, Location Score, Rating Score, Value Score]

Preprocessing is very important because it is sometimes the difference between good results and bad results from clustering algorithms. Eventually we performed log transform followed by minmax transformation for scaling the continuous features. For

categorical features we used one hot encoding and we had to handle two of the categorical features (host verification and amenities) separately as it took too much time with the library which helped us in getting the other features encoded. We used string manipulation with regex to clean the data. We imputed the nans with median values for the continuous features to preserve the distribution of those features. Our US based Airbnb listings dataset has dimensions as 134000*89 and our major focus was on clustering within a local cluster that is the region wise clusters. We worked on the Boston Airbnb Listings dataset (with dimensions 2059*85) for our experiments which can be scaled to other cities and countries. We performed feature selection based on the domain knowledge and on the quality of the features (features which had meaningful and complete data) by removing unwanted features and features with most values missing.

5.1 Geographical Concentrations

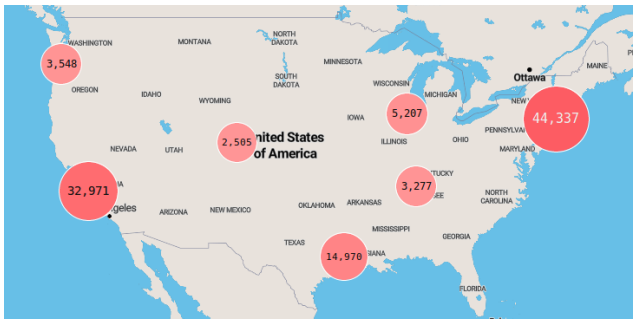

Figure 3: Plot of latitudes and longitudes of listings on U.S. map

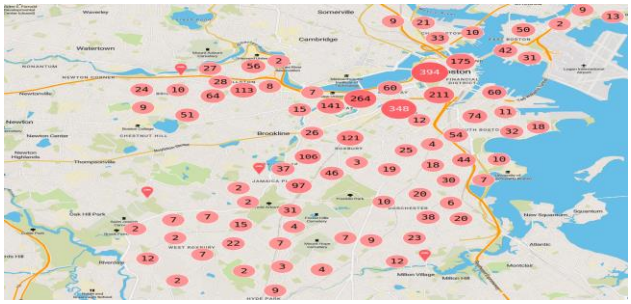

Figure 4: Plot of latitudes and longitudes of listings on Boston map

Then we picked up the listings falling under the Boston area for clustering. Our expectation was that the clusters should have different mean review scores.

We performed PCA for dimensionality reduction and made sure that we preserve 80% energy (variance) during various runs of clustering algorithms with different groups of features.

We started with K Means clustering initially with k=5 obtained from the elbow method. The clusters formed using K-Means clustering algorithm were not distinctive as each cluster had similar and high average review scores. A possible reason for this could be that most of the listings had high review scores and a hard-clustering algorithm like K-Means would assign a listing to a cluster but won't consider it as a mixture. As many features are skewed a soft clustering algorithm would perform better. We explored Spectral Clustering to observe how it would perform and the results were not as per our expectations. All the clusters consistently had similar and high review scores. Agglomerative clustering was performed on the categorical features with cosine similarity measure as its parameter along with single, complete and average linkages. However, with single linkage the clusters were imbalanced and with average and complete linkages the clusters had similar and high ratings. Our experiment with Gaussian Mixture Model gave us better results and formed clusters that were distinctive enough to make some judgements about the listings in them. We consistently got one cluster with lower review scores compared to other clusters. We went ahead with 5 mixture components with the help of Bayesian Information Criterion score. We sometimes had low BIC scores for other values of k but we got similar and high review scores for all clusters. k=5 yielded us consistent results in most cases. This was a tradeoff that we had to make. Our objective now was to find features that would contribute to the cluster formation. We implemented the empirical approach for cluster profiling and selected a set of features, taking the top three features from each of the principal components that yielded us more than 80% variance and based on domain knowledge. After selection of the features we calculated scaled average scores for each feature from this set of features across the different clusters and plotted the values together on a bar graph. With these visualizations for different groups of features, we could infer what features were most influential and how they relate to the different review scores for the listings. We have mentioned our findings below:

Graph Legends common for all bar graphs: [Cluster 0-orange, 1-turquoise, 2-green, 3-navy, 4-red]

5.2 Categorical Features

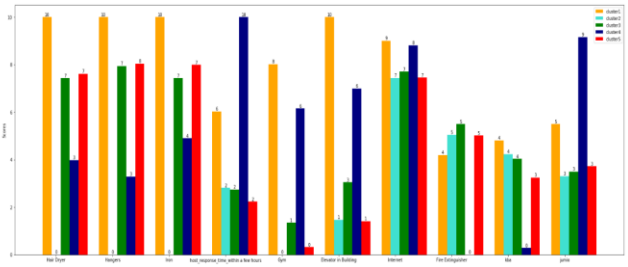

Figure 5: Bar graph showing average scaled values of categorical features for each cluster after PCA and GMM. x-axis has features, y-axis has scaled average feature values. [Features selected L-R – hair dryer, hangers, iron, host response time within a few hours, gym, elevator in the building, internet, fire extinguisher, kba, jumio]

	cluster0	cluster1	cluster2	cluster3	cluster4
review_scores_rating	91.0816	91.7892	93.8935	88.4743	92.1156
review_scores_accuracy	9.28571	9.40811	9.62102	8.93103	9.47335
review_scores_cleanliness	9.54082	9.21081	9.48089	9.25287	9.38558
review_scores_checkin	9.22449	9.64054	9.76752	9.14943	9.73041
review_scores_communication	9.32653	9.62703	9.80255	9.09195	9.70219
review_scores_location	9.58163	9.53388	9.61404	9.57471	9.54545
review_scores_value	8.90816	9.14324	9.34873	8.54023	9.24138

Figure 6: Mean review scores of each type for each cluster.

From figure 5 we observe that cluster 3 (navy) has high scaled average values for features representing amenities like internet, gym, elevator in the building. However, figure 6 shows that cluster 3 has low accuracy and value for money score. This implies that listings in this cluster have incorrect information about the amenities provided and are expensive which led to the low scores.

5.3 Continuous-Categorical Combination

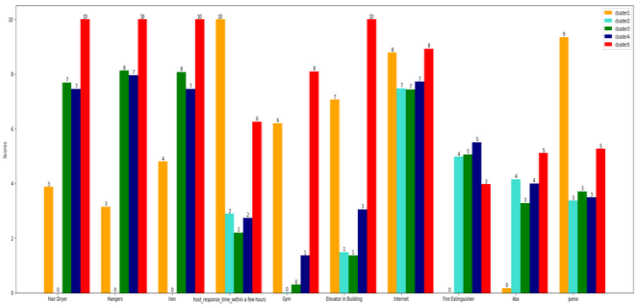

Figure 7: Bar graph showing average scaled values of a combination of continuous and categorical features for each cluster after PCA and GMM. x-axis has features, y-axis has scaled average feature values. [Features selected L-R – hair dryer, hangers, iron, host response time within a few hours, gym, elevator in building, internet, fire extinguisher, kba, jumio]

	cluster0	cluster1	cluster2	cluster3	cluster4
review_scores_rating	88.0482	91.8245	92.1546	93.9088	90.9897
review_scores_accuracy	8.88485	9.41489	9.47468	9.62362	9.25773
review_scores_cleanliness	9.21212	9.21809	9.38924	9.48346	9.54639
review_scores_checkin	9.10303	9.64096	9.73418	9.7685	9.2268
review_scores_communication	9.04848	9.63298	9.70253	9.80157	9.31959
review_scores_location	9.55152	9.53333	9.55063	9.6183	9.57732
review_scores_value	8.49697	9.14362	9.24051	9.35118	8.89691

Figure 8: Mean review scores of each type for each cluster.

From figure 7 we can observe that cluster 1 (orange) has low scaled average value for amenities like hair dryer, hangers and iron which explain the low review score value (value for money) from figure 8. On the other hand, cluster 1 has high scaled average value for gym, elevator in building and internet but at the same time it has low accuracy score. This tells us that the listing information was incorrect and that the amenities were not actually present. This explains the low overall value for money score.

5.4 Continuous Features

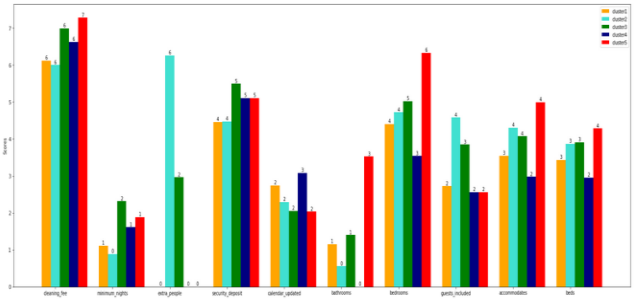

Figure 9: Bar graph showing average scaled values of continuous features for each cluster after PCA and GMM. x-axis has features, y-axis has scaled average feature values. [Features selected L-R – cleaning fee, minimum nights, extra, security deposit, calendar updated, bathroom, bedrooms, guests included, accommodates, beds]

	cluster0	cluster1	cluster2	cluster3	cluster4
review_scores_rating	93.3322	92.8092	92.1135	91.5609	90.4922
review_scores_accuracy	9.58305	9.56214	9.40541	9.35371	9.03906
review_scores_cleanliness	9.45424	9.45698	9.32973	9.26201	9.35938
review_scores_checkin	9.7322	9.71128	9.53514	9.5917	9.32812
review_scores_communication	9.6339	9.7782	9.64865	9.56987	9.25781
review_scores_location	9.68814	9.47228	9.49189	9.63158	9.65625
review_scores_value	9.32542	9.23136	9.15135	9.06987	8.86719

Figure 10: Mean review scores of each type for each cluster.

We can observe from figure 10 that Cluster 4 has a low value for money score, but figure 9 shows that its scaled average feature values are high when compared to other clusters. Although this observation is a bit conflicting with other results, the result can be explained with the help of the average price of the listings falling in cluster 4. The average price is high compared to other clusters and from figure 9, we can observe that the security deposit and cleaning fee for the cluster are high.

This experiment with continuous features can be marked as a conflicting observation.

5.5 Centroid based approach failures

When we performed clustering without PCA, we did cluster profiling using the centroid approach. We found the representative point for the clusters by gathering the listings per cluster and finding the euclidean distances between the listings and the GMM means for clusters formed. Then we found the listing with minimum distance from the cluster center and claimed it as the representative point for this cluster. We then studied the feature values for the representative points across the clusters and observed that many of the representative points had nan values for some features. We observed that the representative points were not a good measure to find influential features and thus went ahead with features obtained from principal components and domain knowledge. We evaluated our clusters using the silhouette score and we got values between 0 and 1 which signifies mixtures in the clusters formed.

6 Conclusion

In this work we found geographical areas where most of the Airbnb listings are concentrated. We picked one of the areas and separated features into groups. None of the groups included review score features. We performed Principal Component Analysis on the data to get features which are most variant. We performed k-means, GMM, agglomerative and spectral clustering and separated the listings into clusters. We consistently found a cluster with low review score. This finding indicates that the listings falling in the cluster have similar values for the features and those values lead to a low score. Using PCA we found the features which were dominating the cluster formation and found the mean feature values in the original dataset.

A limitation of our study is that listing occupancy is not available. We have features telling listing availability but a listing is not available does not mean that it is occupied. It could mean that the host did not put up the house on Airbnb for rent for some reason.

Our work is significant as it opens a possibility of implementing recommendation systems which focus not only on price but also on reviews by the guests. Future work can involve temporal data, investigating if the review scores change in the vacation season. Other future work could involve having more granular geographical areas and finding patterns in data on community level rather than city level. This process of finding feature importance can be generalized for other similar problems.

ACKNOWLEDGMENTS

We thank Professor Tina Eliassi-Rad for her guidance and inputs throughout the term of this project. We would also like to thank many anonymous resource authors who have published their valuable findings on the internet. We also thank our peers for their suggestions in taking the project to completion.

REFERENCES

[1] Ruchi Gupta, Boston Airbnb Open Data Analysis, https://github.com/ruchigupta19/Boston-Airbnb-data-analysis.

[2] Faisal Al-Tameemi, Airbnb Listings Data- Toronto, October 2018, https://medium.com/datadriveninvestor/airbnb-listings-analysis-in-toronto-october-2018-2a5358bae007

[3] Liran Einav, Chiara Farronato, and Jonathan D. Levin. 2016. Peer-to-Peer Markets. Annual Review of Economics 8, 1 (2016), 615–635. https://doi.org/10.1146/ annurev-economics-080315-015334.

[4] Inna Kaler, So You Have Some Clusters, Now What?, https://medium.com/square-corner-blog/so-you-have-some-clusters-now-what-abfd297a575b

[5] Eduardo M. Azevedo and E. Glen Weyl. 2016. Matching markets in the digital age. Science 352 (2016), 1056–1057. Issue 6289. https://doi.org/10.1126/science.aaf7781.

[6] Eyal Ert, Aliza Fleischer, and Nathan Magen. 2016. Trust and Reputation in the Sharing Economy: The Role of Personal Photos on Airbnb. Tourism Management 55 (2016), 62–73. https://doi.org/10.1016/j.tourman.2016.01.013

[7] Corey Wade, Transforming Skewed Data, https://towardsdatascience.com/transforming-skewed-data-73da4c2d0d16

[8] Qing Ke, Sharing Means Renting? https://arxiv.org/pdf/1701.01645.pdf