Regression-Based Artificial Neural Networks to Predict Airbnb Prices in Berlin and Listing Segmentation Through PCA Dimension Reduction and K-Means Clustering

Introduction

2020 was a difficult year for the travel industry. International lockdowns and pandemic fears prevented most from traveling and booking future stays. However, as more people are vaccinated and restrictions are relaxed, the travel industry is preparing for a surge of new reservations. Travel agents are making a resurgence due to "constantly changing rules and restrictions that travelers must navigate" (Sloss). Travel agents have generally avoided booking with Airbnb in part because of questionable quality of the listings (Cogswell).

A dataset ("Berlin Airbnb Ratings") containing information about each Airbnb property in Berlin is analyzed with regression-based artificial neural networks (ANNs) to predict the expected price of a listing based on factors such as property details including the number of guests it can accommodate and room type (shared room, private room, or entire property) and average reviews of the listing in areas such as cleanliness and location. This could provide individuals and travel agents with a better sense of what a listing should cost and whether it is a good value. ANNs are appropriate for this application because they work well for supervised learning regression problems for complicated, non-linear relationships for large datasets. Fourteen predictors and nearly 11,000 listings are included in the final model.

Clustering is an unsupervised learning technique employed to determine if certain categories of listings exist, such as properties which are a good value, which will further help customers book with confidence. Due to the complexity of visualizing and interpreting clusters with a high dimensionality, principal component analysis (PCA) is performed prior to non-hierarchical clustering through K-means. Twenty four predictors are included in the PCA analysis.

Data Preparation

The original dataset includes separate rows for each review for each listing. In order to simplify the dataset, the rows are first grouped by listing ID. The categorical variables are then converted to numerical representations. These include Host.Response.Time, which are given values from 0 to 3 depending on the time taken, Room.Type, which are given values of 0 (shared room), 1 (private room), or 2 (entire home/apt), and Instant.Bookable and Is.Superhost, which are converted to binary (0 or 1) variables from "true" and "false" values. Finally, the data is further cleaned by omitting listings with no reviews, filtering entries with Min.Nights greater than 365, Beds with values of zero, Bathrooms with values greater than the number of people accommodated, and Price less than $0 or greater than $2500, which were found to be unrealistic upon reviewing the listings in this group.

Artificial Neural Network Price Regression

The ANN regression analysis is performed by first splitting the data into three sets (training, validation, and test). The training set is comprised of 80% of the data and the validation and test sets both contain 10% of the remaining data. ANNs with one hidden layer were used for this analysis, via the nnet function in R. Parameter tuning of the decay rate (lambda) and number of nodes in the hidden layer was executed by fitting 90 different models on the training set with lambda values between 0.01 and 0.024 and nodes between 5 and 10. The models were then used to predict the validation set. The model with the lowest mean squared error (MSE) for the validation set was chosen as the best model. This model was then fit on the training and validation set and used to predict the test set to assess the model.

The best model has five nodes in the hidden layer and a lambda of 0.023. This model has an MSE of 0.42 and coefficient of determination R^2 of 0.58 on the training set. On the test set, the MSE is 0.52 and R^2 is 0.48. This indicates the model only explains about half of the variation in price. Addition of predictors such as distance of property to attractions, quality of the listing photos, interior design, unique property features, and view ratings (city view, water view, etc) may increase the predictive ability of the model. In addition, Airbnb prices are subjective as hosts can choose to ignore the suggested price and set their own nightly rates, which decreases the power of the predictors. The final fitted model is a 14-5-1 network with 81 weights is shown in Figure 1 and given by:

```
nnet(Price~Latitude+Longitude+Room.Type+Accomodates+Bathrooms+Bedrooms+Beds+
Guests.Included+Min.Nights+Reviews+Cleanliness.Rating+Communication.Rating+
Location.Rating+Value.Rating, data=Berlin_trainvalid, size = 5, trace = F, linout = T, maxit =
1000, MaxNWts=5000, decay = 0.023)
```

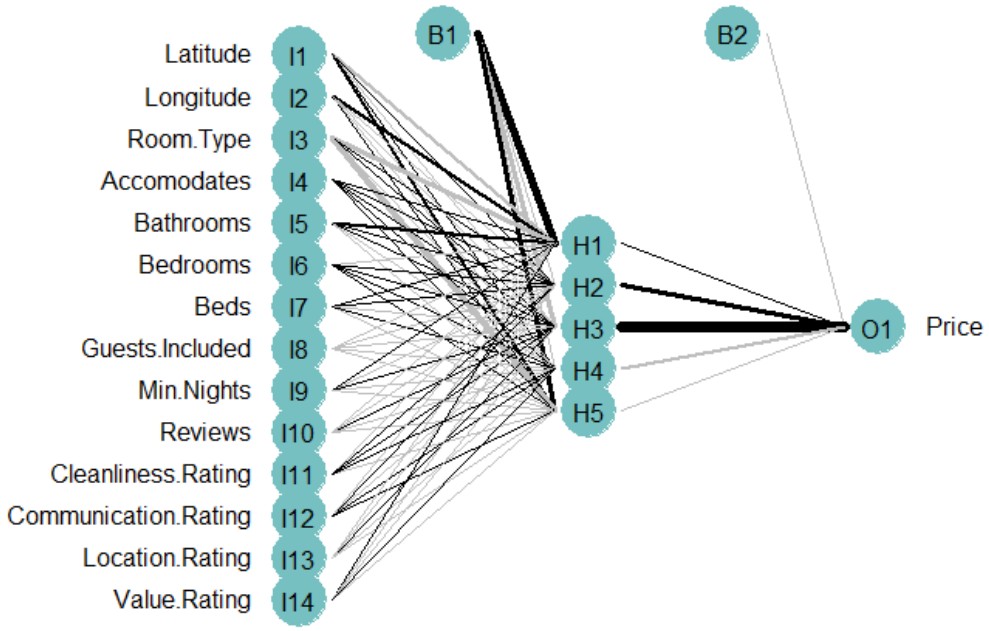

Figure 1: Artificial Neural Network Architecture

The relative importance of the input variables in the final model is evaluated using Garson's algorithm via the garson() function in R and summarized in Figure 2. Room type, latitude/longitude, number of bathrooms, and number accommodated have the largest relative importance in determining listing price. This is reasonable as an entire property would be expected to cost more than a shared room, and a larger property would be expected to cost more as well. Properties in more desirable locations are expected to have a higher price.

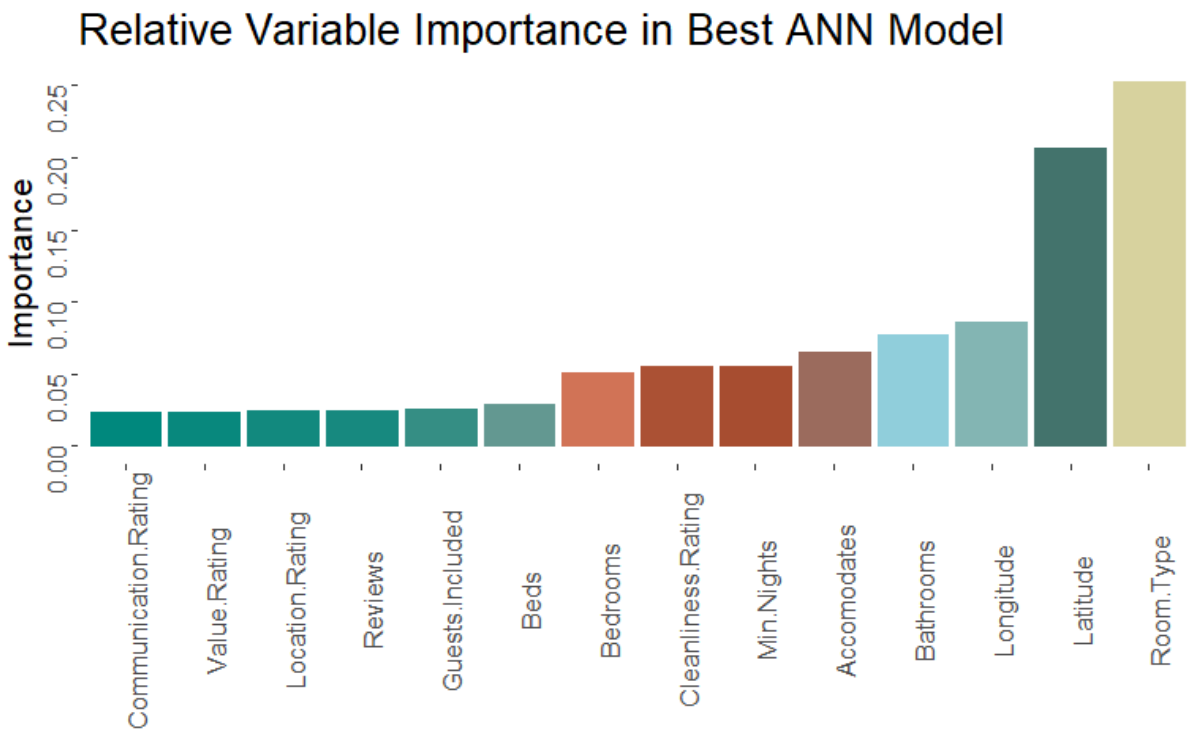

Figure 2: Relative Variable Importance for Predicting Listing Price

Listing Segmentation through PCA and Clustering

Principal components analysis was performed with 24 predictors. Within-cluster sum of squares was calculated for 1 to 10 clusters to determine an appropriate number of clusters to evaluate using the PCA data. Three clusters were determined to be suitable for this analysis. The first two principal components are plotted in Figure 3 along with the assigned cluster for each listing.

The predictors related to property characteristics (beds, bathrooms, room type) are plotted orthogonal to the predictors related to the property rating. This implies that for clustering, the data can likely be split separately along these two groupings. For example, properties with one bed can have high or low ratings, or properties with high ratings can have any number of beds.

A further examination of the listings assigned to each of the clusters reveals three types of properties. Cluster 1 contains highly rated properties that accommodate on average two people, with an average price of about $50 a night. Cluster 2 contains properties with similar ratings to those of cluster 1, but accommodate an average of 6 people for an average price of about $150 a night. Cluster 3 contains properties with similar room traits (beds, price) to those of cluster 1, but with lower ratings than the other two clusters. Most of the shared room listings fall into this cluster. A summary of these predictors for each cluster is included in Figure 4. About 70% of the listings fall into cluster 1. Further clustering of these listings may expose additional subcategories that could further assist potential customers in choosing a property.

Conclusion

Artificial neural networks and clustering can be used by individuals and travel agents to determine whether a listing is a good value, which can assist in choosing a property to book. Addition of predictors and more refined clustering could further improve the predictions and insights about a listing.

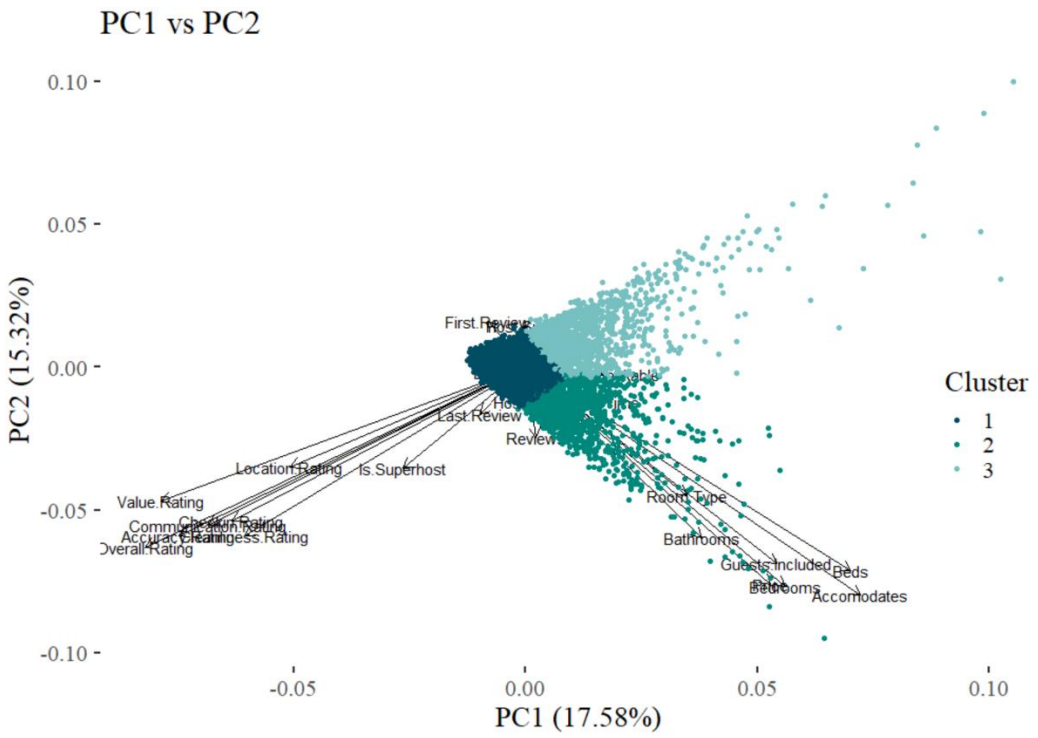

Figure 3: PC1 vs PC2 by Cluster Assignment

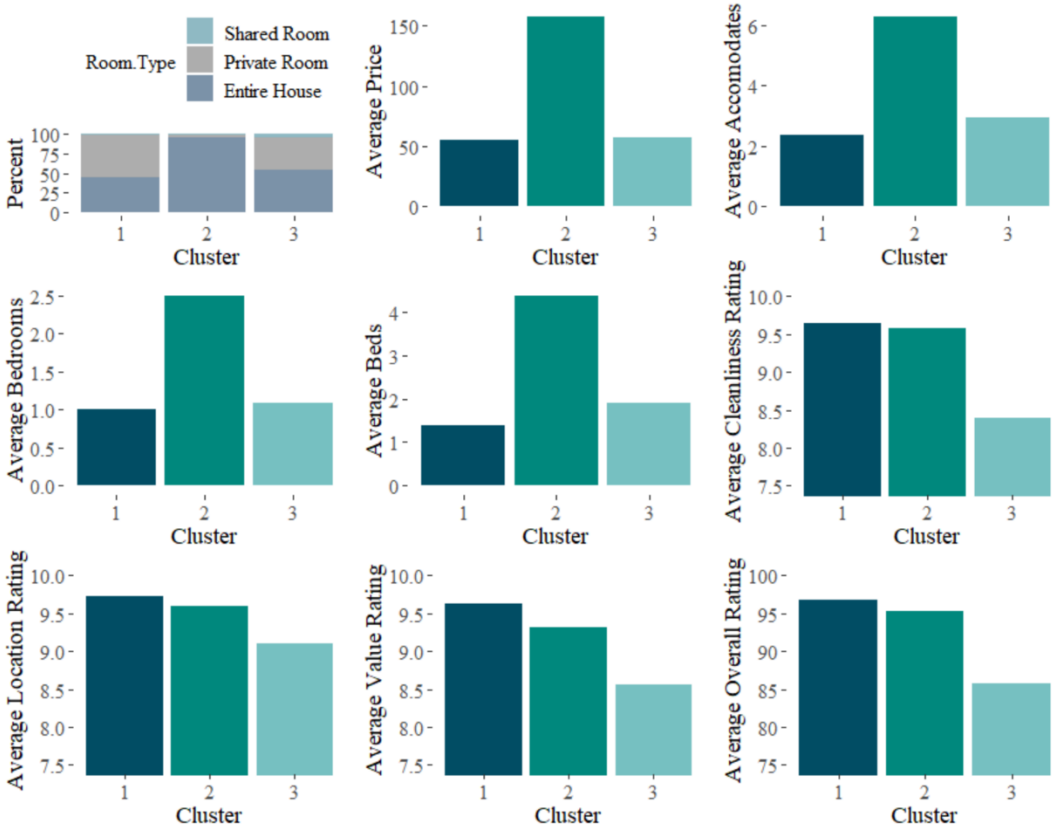

Figure 4: Summary of K-Means Clusters (K=3)

References

1 Sloss, L. (2021, April 14). *Make Way for the Travel Agents. Again.* The New York Times. https://www.nytimes.com/2021/04/14/travel/summer-vacation-travel-agents.html.

2 Cogswell, D. (2018, July 23). *Why one travel agent doesn't use airbnb.* Retrieved April 24, 2021, from https://www.travelmarketreport.com/articles/Why-One-Travel-Agent-Doesnt-Use-Airbnb.

3 Berlin Airbnb Ratings - Dataset by makeovermonday. (2019, June 17). Retrieved April 15, 2021, from https://data.world/makeovermonday/2019w25/workspace/project-summary?agentid=makeovermonday&datasetid=2019w25.