Amanda Howarth, Dave DeStefano

Data Science II

Final Report

**Introduction**

Gauging how a potential guest would rate your hotel as well as other hotels could be a pertinent question within marketing. Hotel rating predictors could allow you to tailor experiences for certain guests and suggest additional features/amenities for lower or higher prices that you know the guest would enjoy based on their TripAdvisor history. Knowing and understanding what factors affect a user’s experience at the hotel can provide hotels with valuable insight into improving guest experiences. In this exercise, the researchers intend to simply predict guest ratings of a hotel, but this exercise could be extended to a prescriptive analytic framework, in which guest experiences could be tailored based on their TripAdvisor accounts.

The goal of this analysis is to determine the best model to predict TripAdvisor users’ ratings of hotels in the Las Vegas Strip. We used the Las Vegas Strip Data Set, which include user’s ratings of 21 hotels in 2015. This data was provided by the UCI Machine Learning Repository and includes 504 observations and 20 variables, such as characteristics of the TripAdvisor user (country of origin, number of reviews, number of years as a member of TripAdvisor), period of the user’s hotel stay, hotel amenities (pool, gym, casino, etc.), and hotel characteristics (number of stars, number of rooms). Each hotel in this dataset received ratings (scores ranging from 1 to 5) from 24 random users across one year (2 ratings each month of the year). The data was precleaned; however, we dummified multilevel categorical variables.

**We aimed to answer the following question in our project:**

* Which approach (linear regression lasso regression, random forest model, decision tree model, or support vector regression) best predicts a TripAdvisor user’s rating of their hotel stay in the Las Vegas Strip?
* Which predictor variables play important roles in predicting a hotel’s score (rating)?

There were some issues with this data, including weekdays were sometimes miscoded as the month. When modeling, the number of rooms is perfectly 1 to 1 with the hotel name, so coefficients are NA automatically.

**Exploratory analysis/visualization**

When analyzing users’ ratings by the “traveler type” (business, couples, families, friends, or solo), we found that ratings appear to differ based on the users’ traveler type (Figure 1). For example, users who traveled as a couple provided the largest number of 5-star ratings. Individuals who traveled as friends or solo gave the least number of 1-star ratings (n=1).

**Figure 1. Distribution of Scores by User Traveler Type**



We plotted a correlation matrix in Figure 2, displaying the relationship between five continuous variables and score (rating). In our dataset, there were few numeric variables about the hotel itself. The numeric variables included in the dataset show us more about the reviewers than the hotels, namely, that reviewers that are very active on the site are less likely to leave negative reviews, or vice versa, that non-TripAdvisor users are more likely to leave a negative review/people with no reviews will leave a bad review but not a good review. We found that since scores can only be 1, 2, 3, 4, or 5, scatter plots were not particularly useful for visualization.

**Figure 2. Correlation Matrix: Relationships between Continuous Variables and Score**

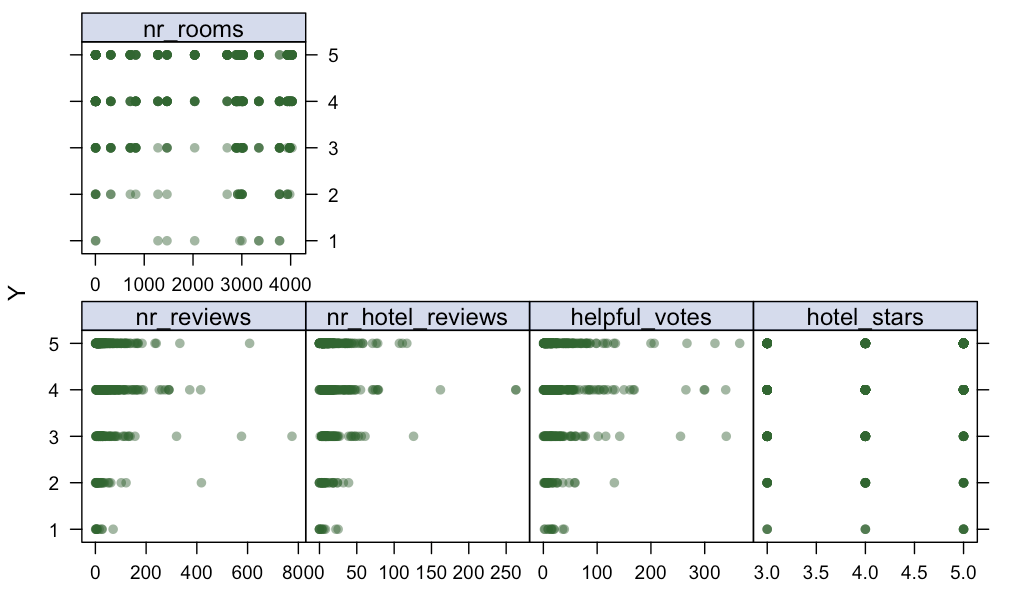
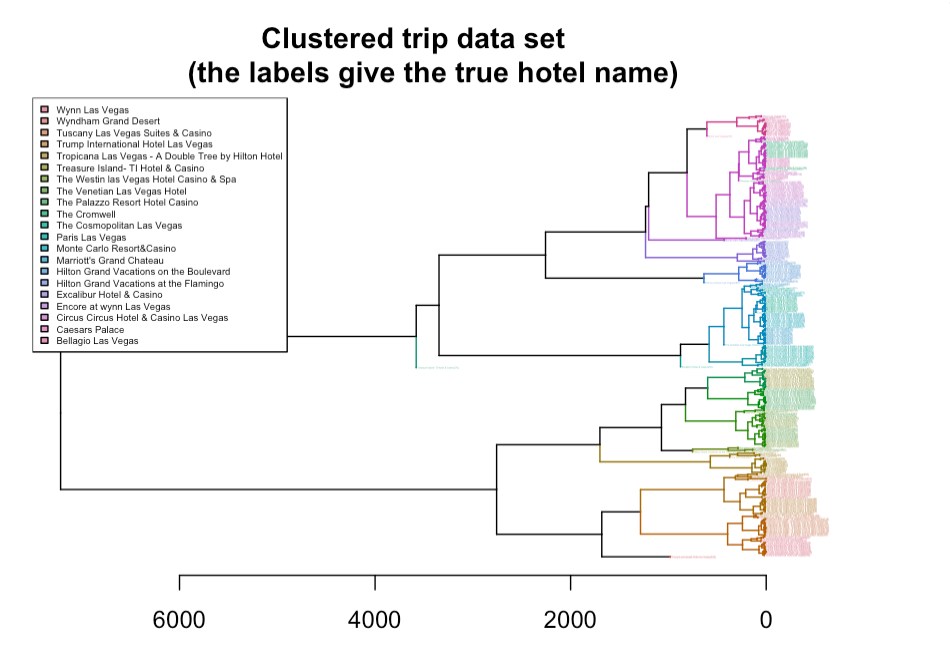


Figure 3 displays hierarchical clustering in a dendrogram so we could see how the 21 hotels clustered together.

**Figure 3. Dendrogram of Hierarchical Clustering Among Hotels**



**Methods and Results**

All predictor variables (19 features) were included in our models to predict score as the response variable. We fit 5 different models using cross validation to determine which model would fit the data best:

* linear regression
* lasso regression
* random forest model
* decision tree model
* support vector regression

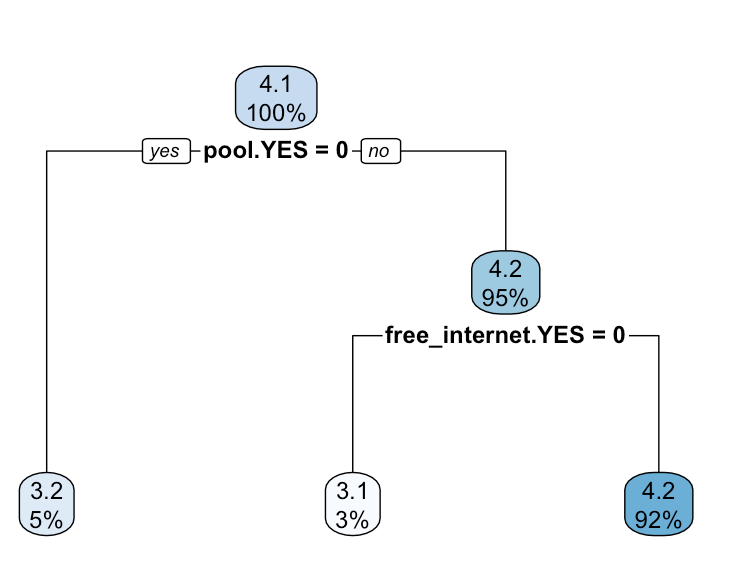
Predictor variables included: 1) user country, 2) number of user reviews, 3) number of hotel reviews, 4) number of user helpful votes, 5) period of stay (months) at hotel, 6) type of traveler, 7) hotel pool (yes/no), 8) gym (yes/no), 9) tennis court (yes/no), 10) spa (yes/no), 11) casino (yes/no), 12) free internet (yes/no), 13) hotel name, 14) hotel stars, 15) number of rooms in hotel, 16) continent of user, 17) number of years individual has been a TripAdvisor user, 18) month of review, and 19) weekday of review. All categorical variables were made into dummy variables. The outcome of interest was score (rating) of hotel.

Our dataset (154 variables total, with dummy variables) was partitioned into a training and test data set. Our training dataset had 337 observations, and our training dataset had 167 observations. Five models were fitted using the training data and the mean squared error (MSE) was calculated for each model using the test data. We measured the MSE to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation.

First, we fit a linear model using least squares on all the predictors in the training data. We found that the test MSE was 1.5449. Next, we fit a lasso regression model on all the predictor variables in our training data using a technique that "shrinks" the coefficient estimates towards zero, which reduces variance. Alpha was held at a value of 1 and our final lambda value chosen by cross-validation was 0.06474015. Our test error was 0.9317.

Next we fit a decision tree model on our training data. Cost complexity pruning was used to grow a very large tree, and then prune it back to obtain a subtree. The tuning parameter controls a trade-off between complexity and how well it fits the data. Our final complexity parameter (alpha value) chosen by cross-validation was 0.0130026. The best fit model has two nodes and one split (Figure 4). In this model, having a hotel pool is deemed an important factor in determining the score a user gives the hotel. For those hotels without a pool, the presence of free internet plays a large role in influencing the score that a user gives a hotel. Our test error was 0.9265.

**Figure 4. Best Fit Decision Tree for Predicting Hotel Score**



Fourth, we fit a random forest model on our training dataset. Random forest models can be used to improve prediction accuracy by decorrelating the trees. We used an mtry value of 1:12 and a node size of 1:5. The test error was 0.8937.

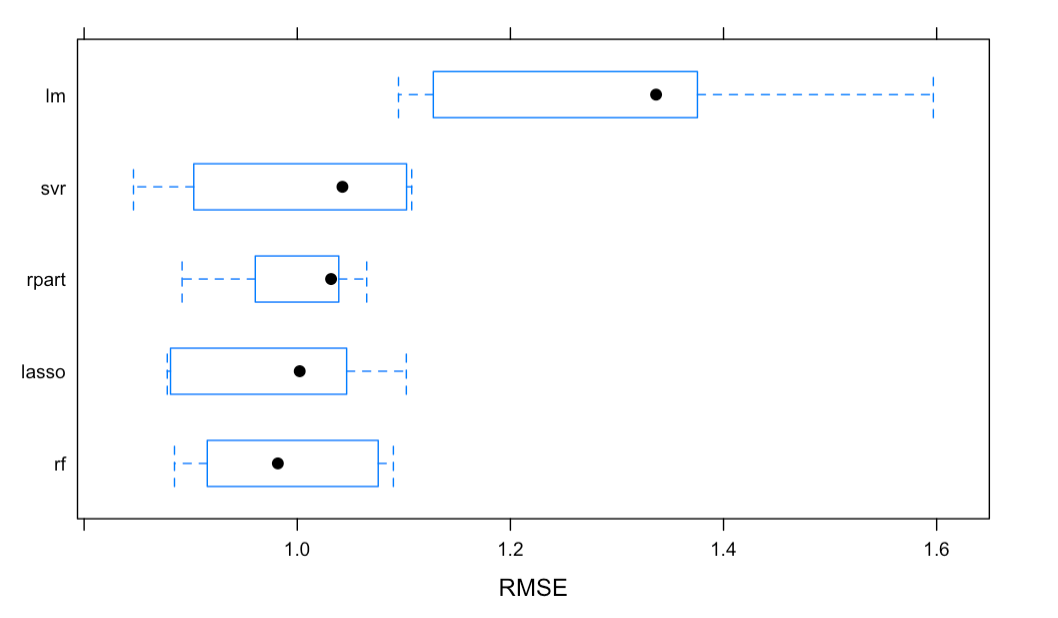
Lastly, we fit a support vector model (linear kernel) to the training data. The tuning parameter “cost” quantifies the penalty associated with having an observation on the wrong side of the classification boundary and can be used to build a linear boundary. The test error was 0.9581.

The best model for predicting hotel score was the random forest model. This model was chosen because it has the smallest cross validation test MSE, a value of 0.8937. Random forest marginally outperformed the decision tree model (MSE: 0.9265), and outperformed the linear model appreciably. Overall, the worst model to predict hotel score was the linear model with the lowest RMSE (1.5449). One disadvantage of the random forest model was that its complexity contributed to making the tree more time-consuming to create and output (i.e., longer time to run in R).

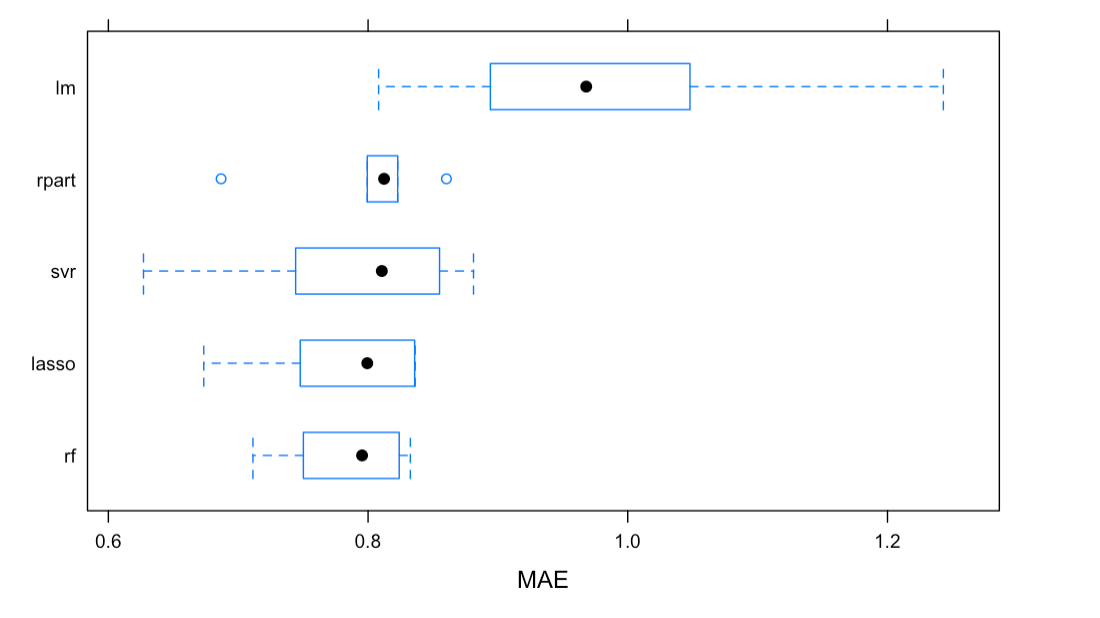
In the scale from 1 to 5 used for the score on TripAdvisor, the random forest (the best model) achieved a test MAE/average absolute deviation of 0.759. This tells us the model, on average, is within one star of predicting the real score.

Box plots of the RMSE and MAE of the training datasets appear in Figures 4 and 5. From the figures, we can see that the median RMSE and MAE of the random forest model are lowest compared to all other models.

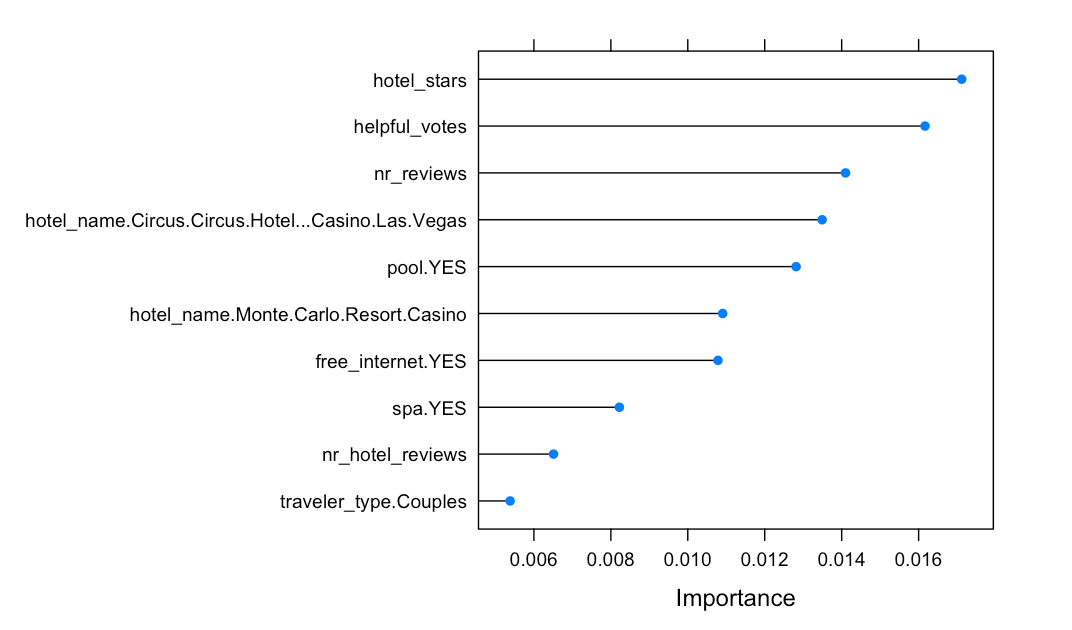
**Figure 5: RMSE of Linear, Lasso, Random Forest, Decision Tree, and Support Vector Models**



**Figure 6: MAE of Linear, Lasso, Random Forest, Decision Tree, and Support Vector Models**



**Figure 7. Top 10 Important Variables in Predicting Hotel Score of TripAdvisor Users**



For the random forest model, the top 10 variables that played important roles in predicting a patient’s total charges are displayed in Figure 7. These variables were chosen using the varImp() function, which automatically chooses a measure of variable importance that is appropriate for given algorithms. The top three variables included number of hotel stars, number of helpful votes of the user, and number of reviews of the user. Characteristics of the TripAdvisor user were particularly important in determining the rating of the hotel, which we found surprising. Other important variables in predicting the rating of the hotel included whether the hotel had a pool, spa, and free internet.

Limitations of this analysis included limited information provided about the hotel itself. It would be interesting to evaluate the cost of the hotel and see if that affected the TripAdvisor user’s rating of it. It would have also been interesting to know the user’s demographic information, particularly their age group.

**Conclusion**

The random forest model predicted TripAdvisor users’ hotel scores the best. The important variables we listed provided us insight into the most important predictors for hotel ratings, particularly (and unexpectedly) that the characteristics of the TripAdvisor user play a large role in predicting the score of the hotel. However, the most important predictor overall was the number of stars of the hotel.

**Dataset**

**Dataset can be found here:** <https://archive.ics.uci.edu/ml/datasets/Las+Vegas+Strip>

**RMD file:** p8106\_Final\_dd2948