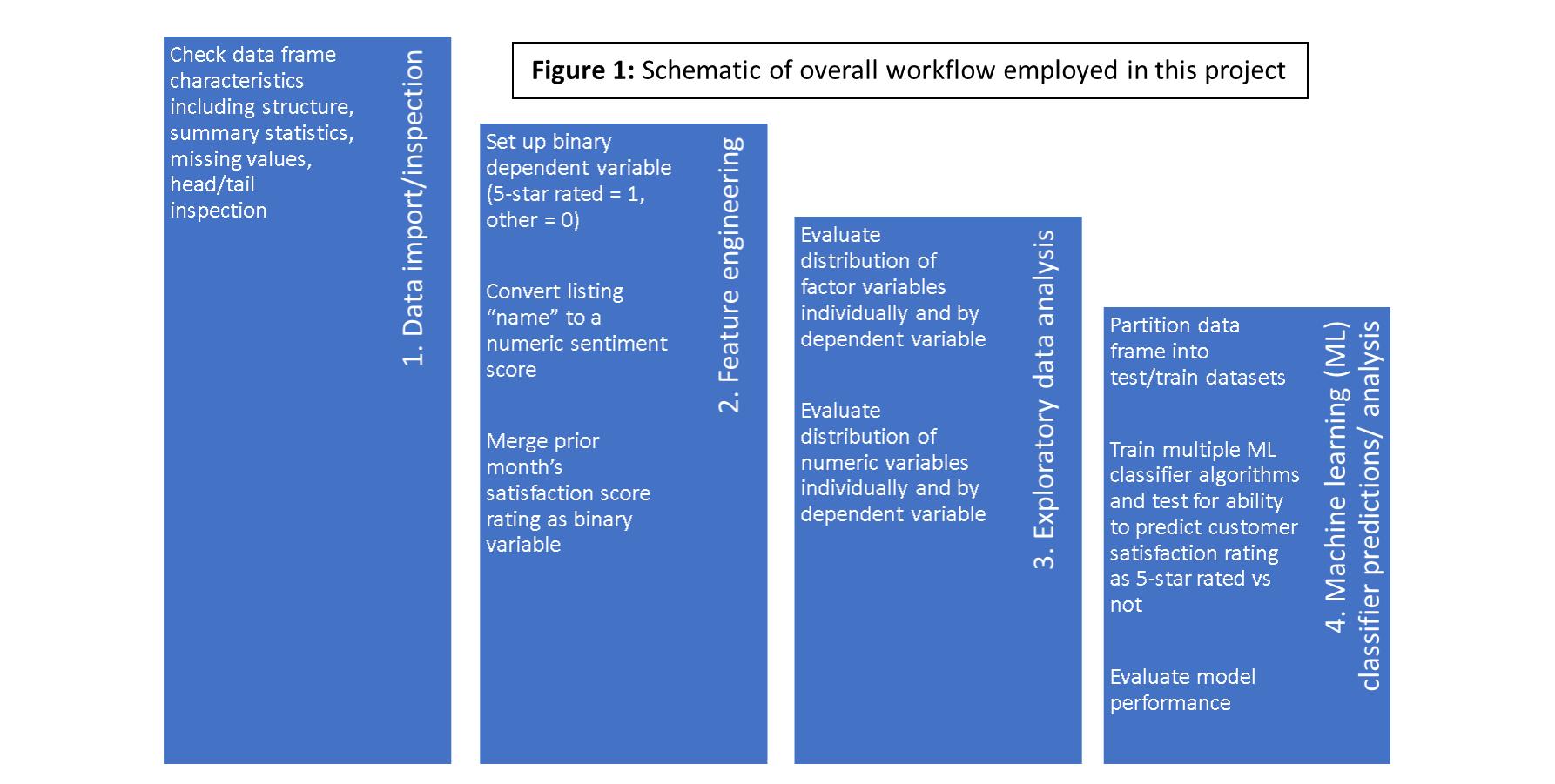
|  |  |
| --- | --- |
| PREDICTIVe MODELS OF overall customer satisfaction FOR airbnb properties in BOSTON | Sricharan Bandhakavi |

**Introduction:** Airbnb Inc ([www.airbnb.com](http://www.airbnb.com)) is a San Francisco based online marketplace/hospitality service with worldwide operations for short-term lodging rentals. Customer ratings of these rentals are an important determinant of the success of these rentals and the reputation of Airbnb. As proof-of-concept, this study evaluates, key metrics of Airbnb listings in the Boston metropolitan area from 2017 July for identifying factors that affect customer satisfaction.

Text mining/sentiment analysis, data analytics, and machine learning approaches were used to identify important drivers of customer satisfaction and predict which listings will receive the highest overall satisfaction score (5-star rated by customers). A similar approach may be extended to additional Airbnb properties for predicting/incentivizing high performing listings.

**Methods:** **Figure 1** outlinesthe workflow used in this project and individual steps within. The first two sections of the workflow are described further in this “methods” section while the remaining two sections are described in the “results and discussion” section.



1. **Data import/inspection:** The “Boston” specific data for this analysis was retrieved as a zip file from the public source:<http://tomslee.net/airbnb-data-collection-get-the-data>**.** The “Boston” zip file contains 23 csv files each of which represent a single “survey” or “scrape” of the Airbnb web site for Boston conducted from 2014 to 2017.

The dataset for this project represents the most recent survey within the “Boston” zip file that was conducted in July 2017 and available as: tomslee\_airbnb\_boston\_1429\_2017-07-10.csv. It consists of 4705 observations and 17 variables (room\_id, survey\_id, host\_id, country, city, neighborhood, reviews, overall\_satisfaction, bedrooms, price in USD, last\_modified, latitude, longitude, and location).

Among these, multiple columns/variables were identified for removal based on majority NA values (“country” column), redundancy of information (“latitude”, “longitude” and “location” columns), and lack of relevance for modeling purposes (“survey\_id” and “host\_id” columns).

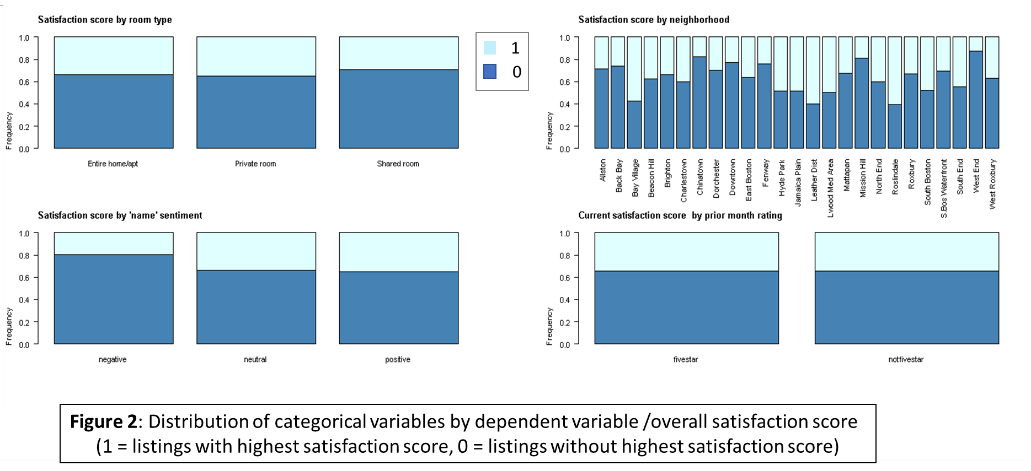
1. **Feature Engineering:** After removal of columns identified for removal in the previous step, feature engineering consisted of three key operations outlined below:

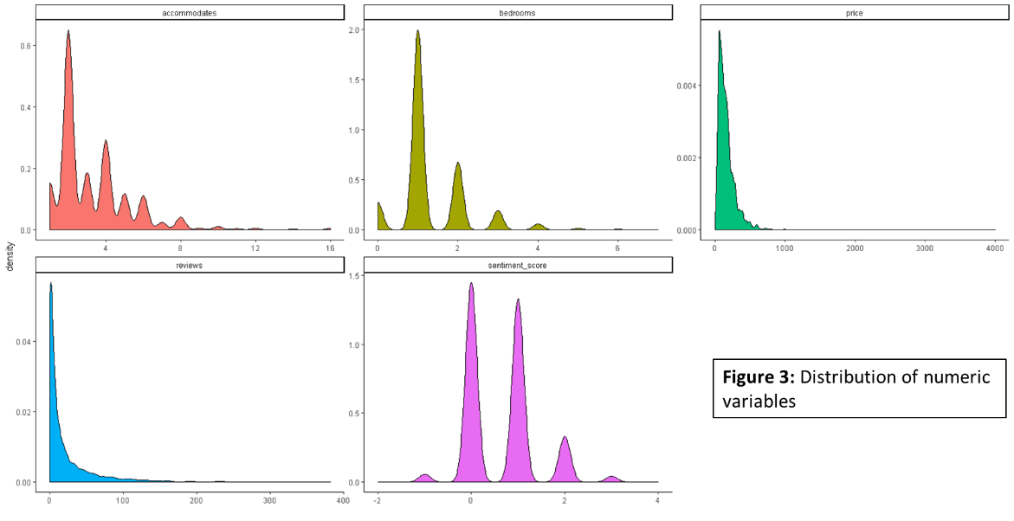
* Dependent variable set up: The dependent variable was set as a binary “factor” variable wherein the highest customer satisfaction score/rating of 5.0 is represented as “1” and all lower ratings represented as “0”. Dummy values of “1” and “0” are used to refer to highest/other satisfaction scores in rest of text.
* Text mining/Sentiment analysis of “name” column: An interesting column within the data frame is the “name” column. The column contains a brief headline description of each listing/property and used as a means to attract attention to the listing on the Airbnb site. Given its potential to attract customers by its inherent messaging, “name” entry for each listing was converted via text mining/sentiment analysis into a numeric value/ “sentiment\_score” using Jeffrey Breen’s algorithm (https://datamatters.blog/tag/sentiment-analysis/).

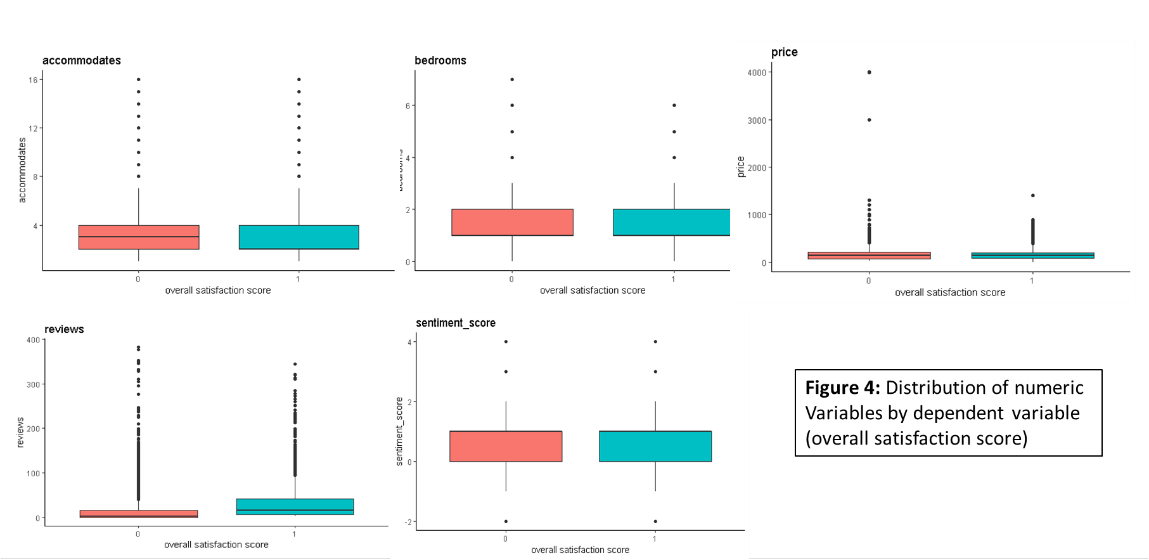
Based on the “sentiment\_score”, a “sentiment” column was further created for labeling the associated sentiment as “positive” (sentiment\_score > 0), “neutral” (sentiment\_score = 0), or “negative” (sentiment\_score < 0). The numeric “sentiment\_score” and categorical label, “sentiment” for each property were used for modeling purposes in the subsequent steps.

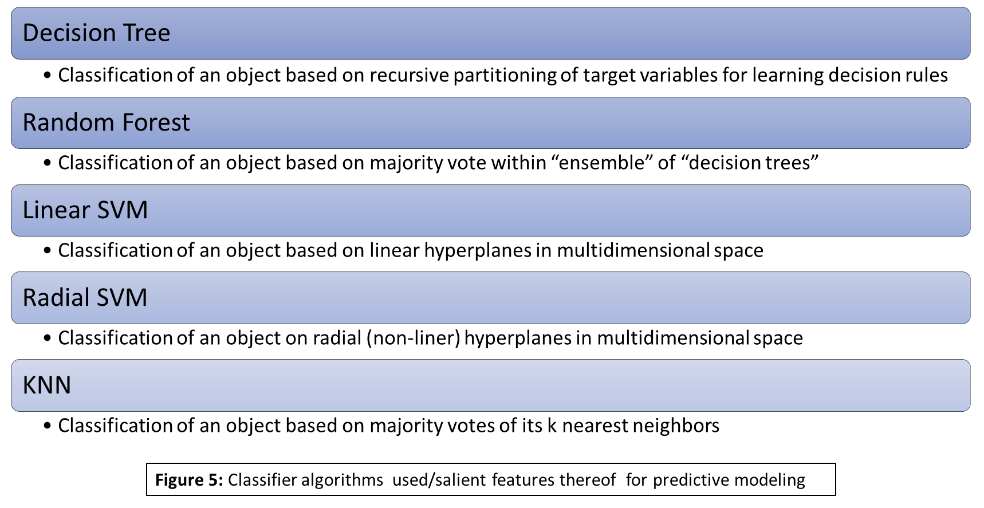
* Prior month’s satisfaction score: Based on the assumption that prior month’s ratings may carry over to the current month, the “overall satisfaction score” for listings also listed in prior month’s dataset (4199/4705) was imported into the current data frame (“merged” in R). This resulted in 506 listings in the July data set that did not have a prior month’s satisfaction score; for these, the mean value of this column (without these listings) was imputed as their assigned value. Finally, the prior month’s satisfaction score was converted to a binary “factor” variable labeling each property/listing as either a five-star rated in prior month or not five-star rated. This variable, “prior\_satisfaction\_score”, was also used for modeling purposes in the subsequent steps.

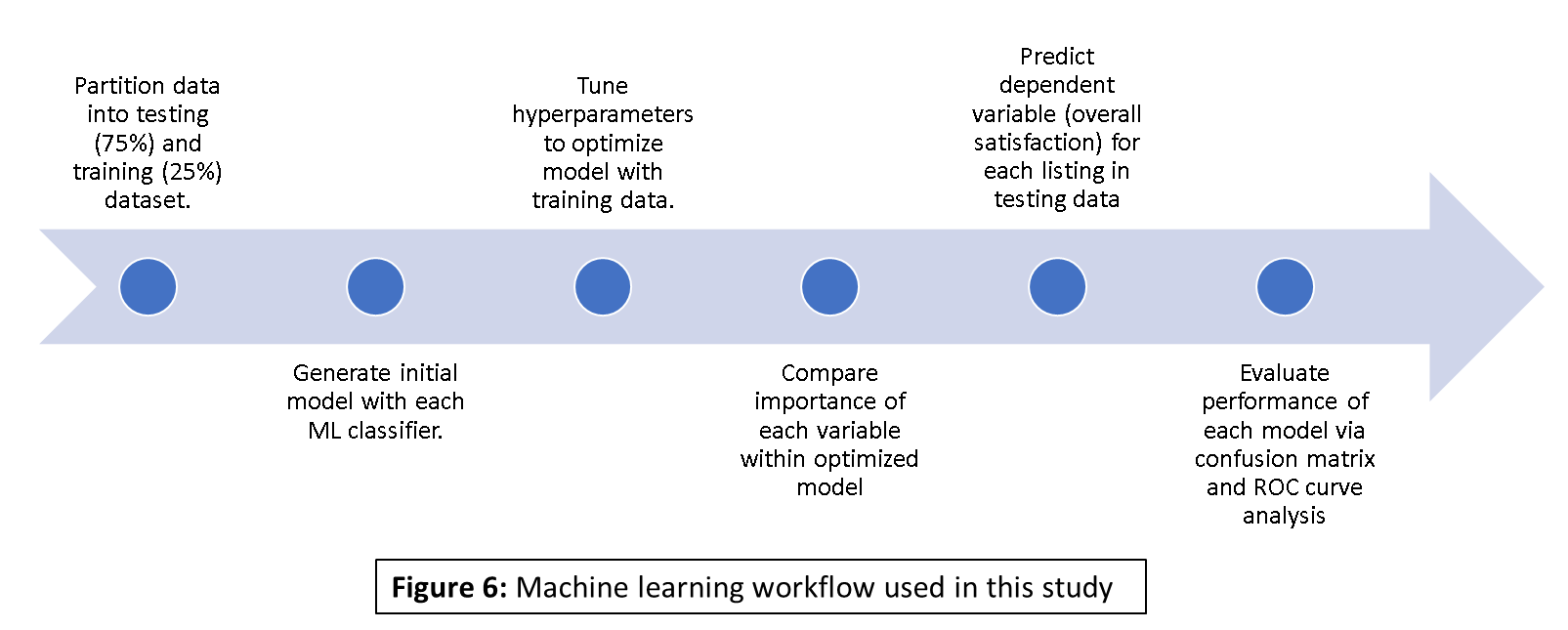
**Results and Discussion:** After generation of the consolidated dataset as outlined above in the “methods” section, as an initial step, **exploratory data analysis (EDA)** was performed to better understand each of the variables within this dataset and their potential relationship with the dependent variable of interest (highest overall customer satisfaction rating). For this step, each of the categorical variables (“room type”, “neighborhood”, “[name] sentiment”, “prior month rating”, “overall satisfaction score”) were analyzed using ‘counts’ and ‘proportions’, while each of the numerical variables (“reviews”, “accommodates”, “bedrooms”, “price”, and “sentiment\_score”) were analyzed using ‘density’ plots and ‘boxplot’ visualizations.

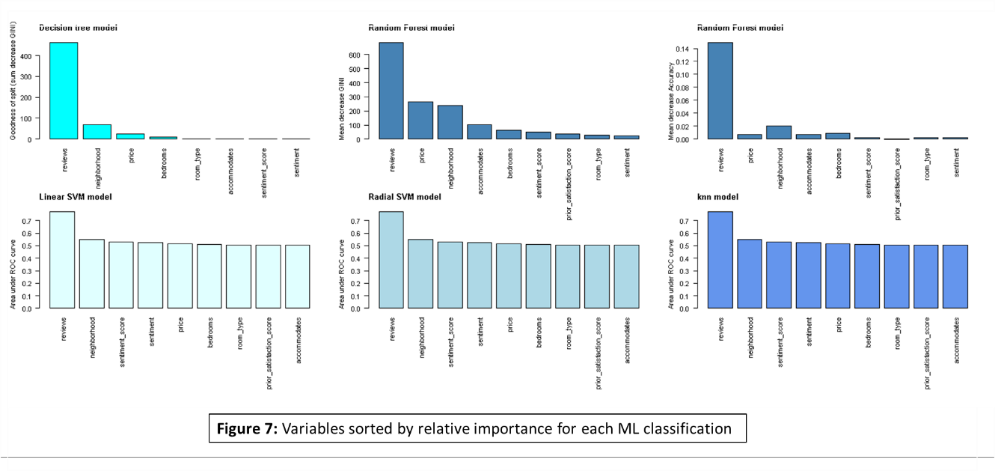
Categorical variables distribution: As shown in **Figure 2** below, overall satisfaction score for listings are not affected by room type (**Figure 2**, top left panel) or prior month ratings (**Figure 2**, bottom right panel). However, some neighborhoods such as Bay Village, Leather District and Roslindale (**Figure 2**, top right panel) have a higher proportion of listings with highest overall satisfaction score (“dummy” value = 1). Also, negative sentiment scores associated with listing “name” have a lower proportion of listings with highest overall satisfaction score (**Figure 2**, bottom left panel). See also **appendix** for distribution of these variables by count.

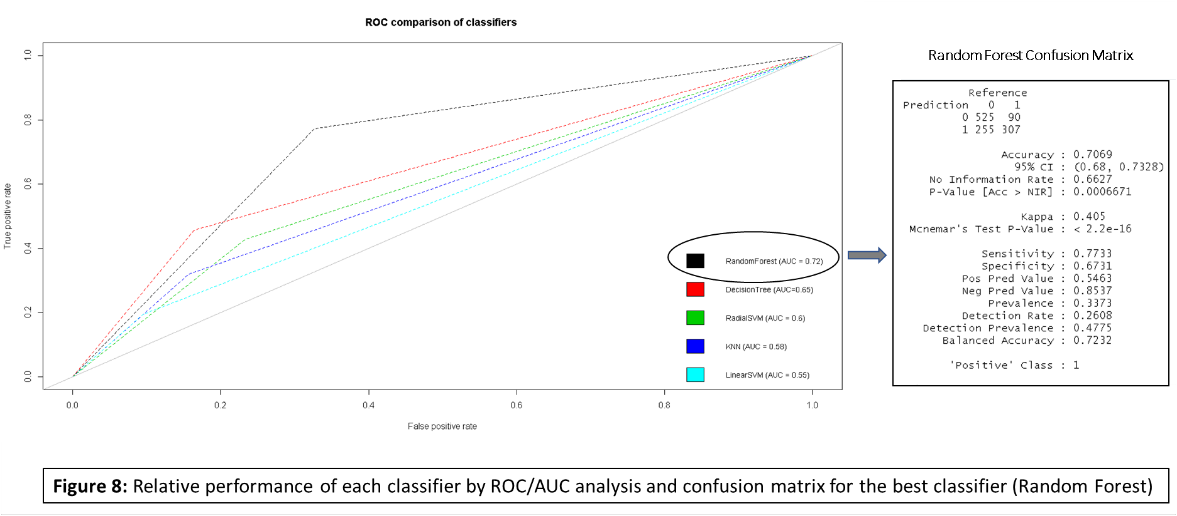
**Numeric variables distributions:** **Figure 3** below shows the distribution profile of each numeric variable within the dataset. As indicated “price” and “reviews” are unimodal in distribution whereas the other three variables are polymodal. By boxplot analysis (see **appendix**), all columns have a significant number of outliers. However, no outlier treatment was performed in order to maintain data integrity.

Next, the distribution of each numeric variable by dependent variable (overall satisfaction) was visualized by boxplot analysis as shown in **Figure 4**. As shown in the figure, listings with highest overall satisfaction score have higher number of reviews. Separately, the median value for “accommodates” is higher for listings with the lower satisfaction score suggesting this feature is anti-correlated with highest ratings. By contrast, none of the remaining variables appear to be able to stratify listings based upon highest vs lower satisfaction score.

1. **Machine learning classifier predictions/analyses:** As the final step in this analysis, a variety of machine learning (ML) classifier algorithms were employed to evaluate their ability to correctly predict those properties with the highest satisfaction score versus not. ML Algorithms used: Five different algorithms representing distinct ML families were employed as it was not clear at the outset which algorithm was going to generate the best predictive performance. See facing **Figure 5** for classifier algorithms employed in this study and salient features associated with each algorithm.

****Machine learning workflow: **Figure 6** outlines the overall workflow used for machine learning with each classifier. Briefly, the Boston Airbnb listings dataset was partitioned into a training and testing dataset at a 75:25 ratio resulting in 3528 listings in the training data and 1177 listings in the testing data. For each ML classifier, the same training dataset was used for generating an initial model; each initial model was “tuned” for critical hyperparameters and cross-validated 10-fold. Next, the importance of each variable in the training dataset was compared. and the optimized model for each classifier was used to predict the satisfaction rating of Airbnb listings in the testing dataset. Finally, the performance of each classifier was evaluated via confusion matrix and ROC curve/AUC analysis.

Variable importance: **Figure 7** in facing page shows the relative importance of each variable in the final model generated using the various ML classifiers. As expected from exploratory data analysis, “reviews” and “neighborhood” are highly important contributors for each model. By contrast, previous month’s ratings (“prior\_satisfaction\_score”) and accommodation capacity (“accommodates”) of each listing have very low importance in the final model regardless of the algorithm used. Interestingly, “sentiment\_score” a variable generated via text mining/sentiment analysis of the “name” of each listing is fairly important in some of the models (linear SVM, radial SVM, and KNN), but not in all models. Prior to further evaluation of variable importance/predictors of customer satisfaction, see next section on relative performance of each model.

Model performance: **Figure 8** below shows relative performance of each classifier in a ROC chart with the AUC metric for each model. The Random forest classifier outperforms all other classifiers with an AUC of 0.72. Random forest picks up ~77% of true positives (sensitivity = 0.7733) at a false positive rate of ~0.33 (1-specificity = 1-0.6731 ~0.33). Given its superior performance relative to other models, variable importance by random forest (and not other models) should be used to identify significant drivers of customer satisfaction from **Figure 7**.

**Conclusion/next steps:** As proof-of-concept, a multi-ML classifier approach was used to identify important predictors of highest customer satisfaction rating of Airbnb listings in Boston. Random Forest was the best classification predictor with accuracy ~0.71, AUC ~0.72, and sensitivity ~0.77.

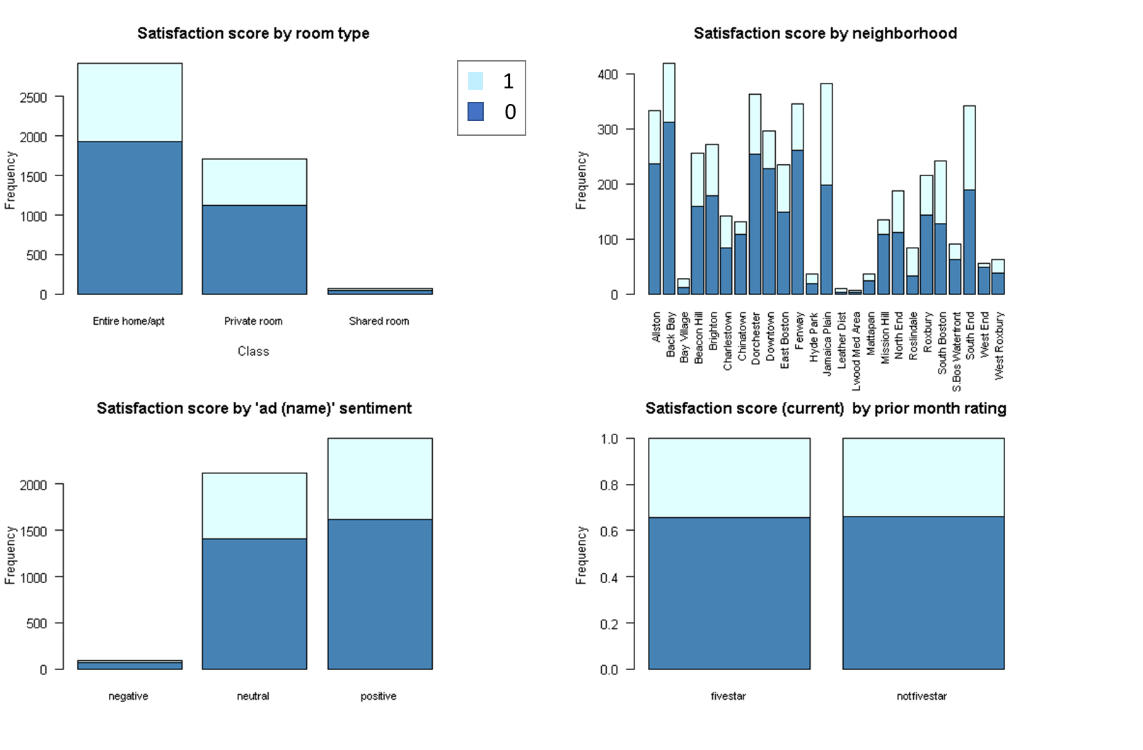
Number of reviews and neighborhood are the most important predictor variables. Thus, as next steps, it would be useful to identify factors promoting # of reviews as satisfaction drivers on a neighborhood-basis. Unfortunately, information within this dataset was insufficient for identifying significant drivers for # of reviews (*data not shown*); thus, a survey focusing on these questions is needed. A similar multi-ML approach may be used on other Airbnb listings across locations for identifying/incentivizing high performance listings.

**APPENDIX IN NEXT PAGE**

**APPENDIX**

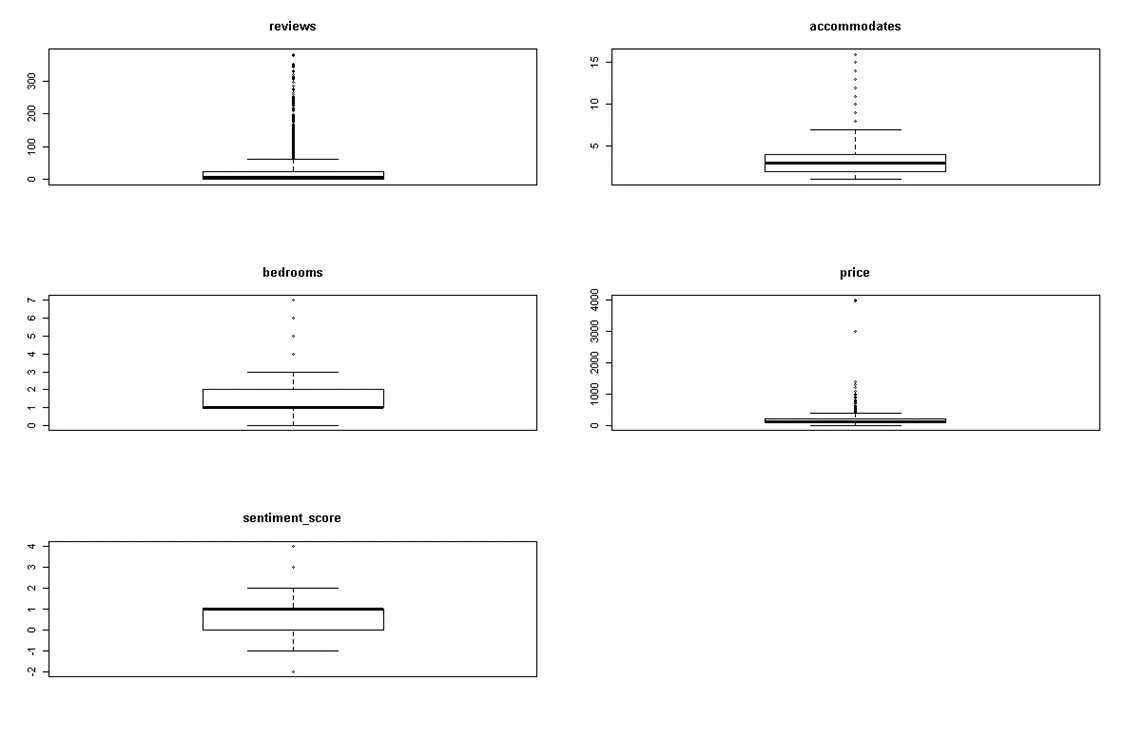
**Counts of categorical variables by satisfaction score**

In the main text, distribution of categorical variables by satisfaction score is shown as a proportion only (**Figure 2**, main text). As a complement to this analysis, the same distributions are shown below as “counts”.

As indicated in **top left figure** in below panel, most properties are entire home/apartments and their satisfaction score is slightly higher than that of private rooms. The **top right panel** shows that Backbay, Jamaica Plain, Allston, Dorchester, East Boston and South End have the highest # of properties whereas L’wood Medical Area, Leather District and Bay Village have the fewest # of properties. Some neighborhoods have a higher proportion of listings with highest satisfaction score. **Bottom left panel** shows that most properties have a positive or neutral sentiment score associated with their name and there is no apparent difference in the proportion of listings with highest satisfaction score among positive vs neutral listings. Finally, **bottom right panel** shows no apparent effect of prior month ratings on current month’s satisfaction score.

**Boxplot analysis of all numeric variables**

In the main text, distribution of all numeric variables was visualized as a density plot (**Figure 3**, main text) and as a complement, boxplot analysis was also performed and shown here. The boxplots show higher # of outliers for “reviews” and “price” variables compared to remaining variables, but all variables have outliers. However, as mentioned in the main text, no outlier treatment was performed to preserve data integrity.

****