Predicting Airbnb Prices with Random Forest Methods

MAX KARSOK

DECEMBER 2018

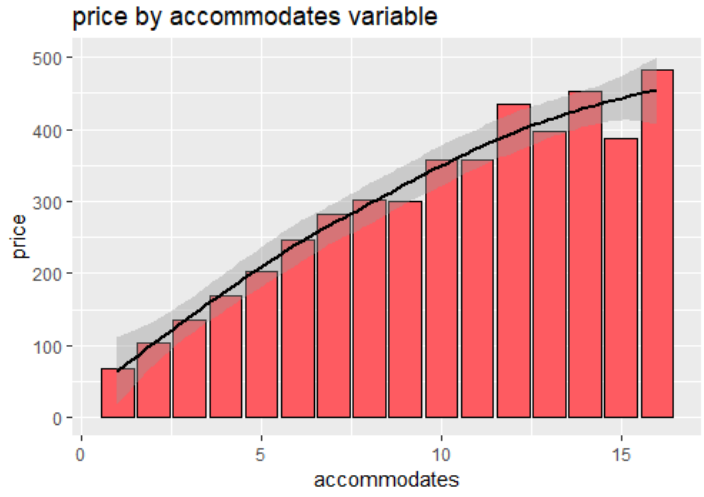
I. Introduction

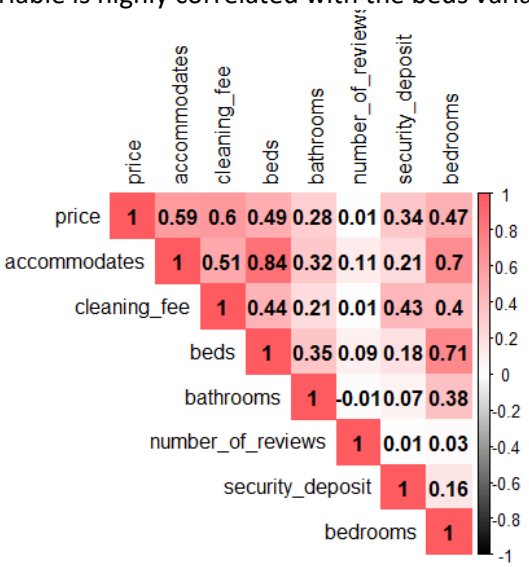
The final iteration of the model designed to predict Airbnb prices in New York City produced a root mean squared error (RMSE) of 54.5494, outperforming a standard linear regression model by 14%.This work will summarize the data normalization and augmentation of the data set to prepare for modeling, the feature selection process, the comparison of evaluated models, the end results of the model's final iteration, and future considerations for continued model development or the model hypothetically integrating to an Airbnb production environment.

II. Exploring the Data

With price of the listing as the dependent variable in this model, the primary purposes of exploratory analysis were to understand (a) the completeness of the dataset and (2) potentially valuable features for the model as it relates to predicting price (using intuition alone). While significant data exploration was conducted ahead of modeling, two examples will be provided for brevity purposes:

In the first example, average prices are calculated against the number of people the listing accommodates, confirming intuition about a relationship between the accommodates variable and the price of the listing. A second example examines correlations of some variables against the dependent variable and against each other. This exploration helped identify potentially significant attributes as well as multi-collinearity. While the final model was a random forest that accounts for multi-collinearity, this process is helpful during the exploratory stage for understanding the business situation (ex: the accommodates variable is highly correlated with the beds variable):





III. Preparing the Data for Analysis

Data preparation occurred in three phases during this process: (1) the modification of existing data, (2) integration of non-provided and publicly available data points, and (3) feature selection against the normalized dataset.

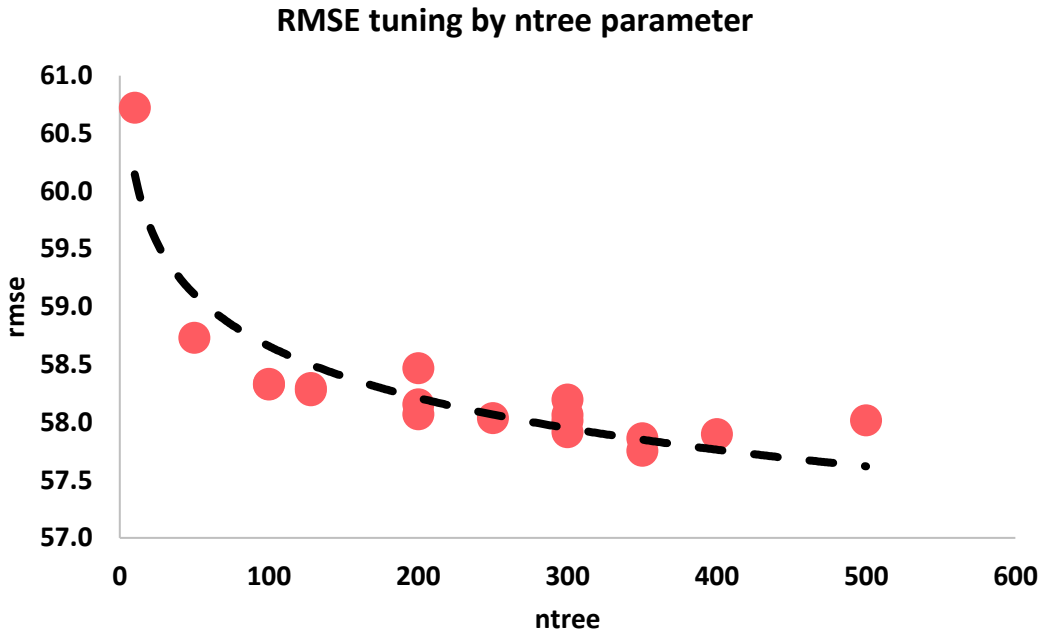
1. Modification of existing data (data cleansing): The first set of feature modifications were made on existing data points. Some of these data cleansing tactics managed NA values in different capacities; for example, null values in the security deposit attribute were replaced with $0 on the assumption that there was no security deposit for these listings. In other cases, replaces a numeric field with the mean of that column was the best normalization approach, especially if there were limited amounts of empty values. Other text cleaning processes included a separate test around high priced neighborhoods and separating these neighborhoods as a variable, and some basic text processing concepts based off of the description fields provided by the lister.
2. Integration of Census Data: The second major feature addition to the model was the incorporation of non-provided census bureau data. The total population and median household income for each zip code in New York City was added to each listing through a publicly available data source here: https://factfinder.census.gov/. Testing all available census attributes was not in the scope of the project, so intuition and for model interpretability, only zip code population and median household income were included. Income proved to be a very significant attribute in the model, while population only moderately significant.

3. Feature Selection: To identify which of the 53 prepared features were best suited for analysis, a forward stepwise selection was executed. This process brings variables into the model one at a time based on their significance to the dependent variable. This process selected 39 variables to be included in any modeling done on the dataset. Forward stepwise selection was used and the feature selection process (as opposed to lasso, backward stepwise, etc.) for two reasons: (1) interpretability of the results for a non-technical user, and (2) the ability to eliminate variables with high collinearity. For reference, the top variables selected by the forward stepwise are shown below:

Rank (generated by AIC in forward stepwise)
1 – *accommodates* (provided, not modified)
2 – *median household income* (user added via census)
3 – *cleaning fee* (provided, NA = $0)
4 – *Manhattan flag* (user created, logical variable if listing located in Manhattan
5 – *Room Type = Entire* (user created, logical variable if listing was the entire apartment)

IV. Modeling Techniques

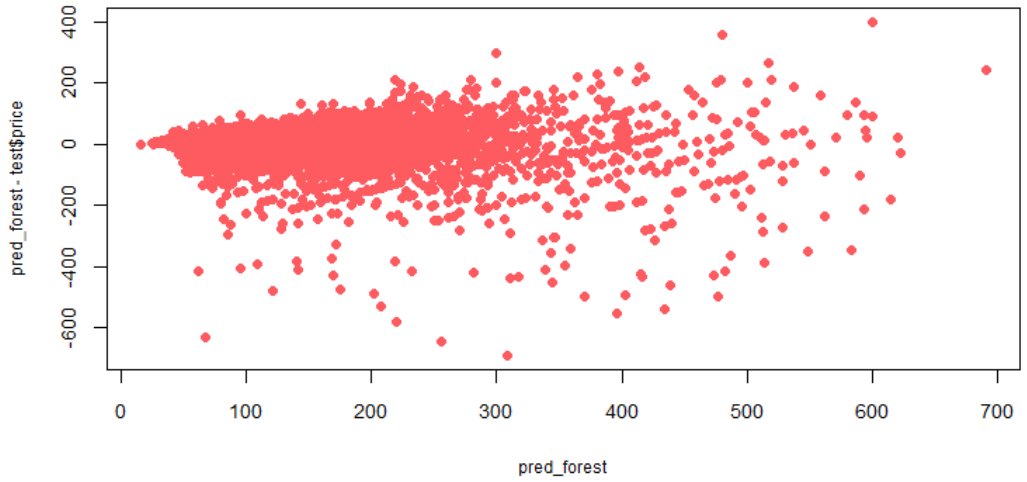
The random forest model chosen to predict the provided Airbnb listings creates 400 decision trees, evaluating up to 6 variables at each split in each tree. In the randomForest package, the number of trees in the random forest and the number of variables to be examined at a given split are the two user-definable parameters. After several manual iterations of testing, the combination of ntree = 400 and mtry = 6 provided the lowest RMSE on the test dataset. See below for the different combinations of these parameters and the resulting impact on the test RMSE:



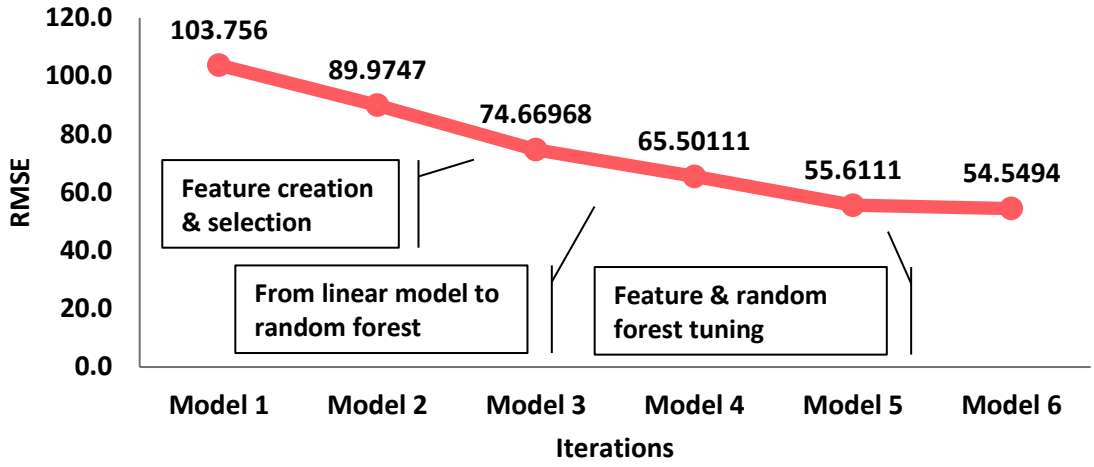
The decision to build a random forest model with the selected features was made in comparison to a linear regression model, linear support vector machine model, and other tree models. These models were not as accurate using the same features as the random forest (ex: for comparison, a linear regression using the same features generates an RMSE of 63.448).

V. Results

The final model generated an RMSE of 54.5494 and the below residuals. Analysis of the residual plot indicates that the model's most inaccurate predictions were more commonly less than the actual price value as opposed to greater than the actual price value. The majority of the residuals just slightly overestimate the actual price of the listing, and the outliers tend to underestimate the actual value.



The model's performance improved drastically over time. Below illustrates the various iterations and relevant points along the development of the final model to arrive at the RMSE below 55:



VI. Discussion

I was satisfied with the results of the feature selection process and random forest model implementation against the Airbnb data set. There are 4 areas that I believe could have been valuable for increasing the model's predictive power:

1. Elements of Seasonality: Each listing scraped in both the base and scoring dataset was posted on Airbnb between 03/04/2018 and 03/06/2018. While this was adequate for this exercise because all postings were taken during the same time period, I feel that an influential variable in this analysis could be the date that listing was posted. For example, an Airbnb owner in Midtown is likely charging a different price for an apartment near Times Square on New Year's Eve compared to a Tuesday night in April.
2. Further text analysis: Some of the most valuable attributes in the random forest model were very simple aggregations of text fields in dataset. Further implementations of NLP techniques (sentiment analysis, topic modeling, etc.) would help augment the analysis.
3. Historical prices of the same listing ID: If this price prediction model were ever to move to production at Airbnb, it could leverage the historical prices of the same listing to predict future prices of that listing. For example, if the listing has been rented 100 times in the past three years, the model can consume the average value of the listing over time, account for potential inflation, and leverage that history to predict what the current and future value of that listing should be.
4. Unsupervised methods: I believe that unsupervised statistical methods such as K-Means clustering could be valuable for analyzing and predicting these Airbnb listing prices. Because of the geographical components of the housing market, as well as the ability to group similar listings by common features makes unsupervised methods a logical next step to take this model.