Airbnb new listing price predictor system

Jeffrey Samuel, Mehul Zawar, Shivendra Kumar, Shruti Gupta, Siddharth Harisankar

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907
jchittar@purdue.edu; mzawar@purdue.edu; kumar394@purdue.edu; gupta592@purdue.edu; sharisan@purdue.edu

Our objective is two-fold. First, we identify the important parameters or features which affect an Airbnb's listing price and estimate the price per night. Second, we develop an R shiny app which provides an interface to input the parameters or features of a new listing and display the estimated price per night.

Listing a home on Airbnb requires considerable investment from a host (our stakeholder). But, how will a host know how to price his/her listing ? How will the inclusion or exclusion of certain amenities influence the listing's price ? We would like to develop an app to enable hosts obtain answers to the above questions and price their listings accordingly

We developed a Random Forest Regression model to identify the key features which influence the pricing of a listing. Then, we used a Gradient Boosting Model (GBM) to estimate the price of a listing based on the inputs provided for the selected features. Once the model was developed, we built an R shiny app to obtain the feature's value from the hosts, pass the values as input to GBM and display the estimated price for the listing.

1. Business Problem:

Estimate the price of a new Airbnb listing based on the features of the listing. Features include the location of the listing, number of bedrooms, amenities (such as TV, Internet), etc. The hosts of Airbnb listings will be the stakeholder.

Solving this problem requires us to know how each feature affects the pricing. For this, we can leverage analytics and identify not only how each feature affects pricing, but also the most important features which affect pricing. This would provide hosts a considerable idea about which feature to include or exclude and to also price their listings correctly

2. Analytics Problem:

Identify the key factors of an Airbnb listing which affect its pricing and estimate the price of a new listing. We will also need to identify the monetary effect each feature has on the estimated price. Key metrics of success is accurately estimating the price for a new listing and enabling adoption of the app among hew Airbnb hosts

3. Data:

We are using files of web-scrapped data of Berlin Airbnb listings in csv formats. The main file is at listing level and has 96 attributes of listings for ~22k listings. For data cleaning, we first removed variables such as URLs, web-scraping IDs and license-related columns from the dataset which had no business significance. We then identified columns which had high number NULLs (we set a threshold of 40%) and dropped such columns. After removing unnecessary columns, we

performed Exploratory Data Analysis (EDA) to identify which columns had outliers and capped/floored outliers based which were two-standard deviations higher/lower than the mean

4. Methodology Selection:

As mentioned above, we performed Exploratory Data Analysis (EDA) to outlier-detection. We also leveraged descriptive analytics and performed visualizations and text analytics (word-clouds) to understand the current business trends to make informed decisions.

We also performed predictive modelling to predict which features affected a listing's price. Predictive analytics was essential for us to understand the effect of how each feature affected a listing's price and estimate the price of a listing.

5. Model Building:

We first ran a multiple linear regression to predict the price of the listing by partitioning the dataset into train (80%) and test. Next, we ran a random forest model to identify the important features which affect the price of the listing. Once we have the most important features, we developed a Gradient Boosting Model (GBM) with 1000 trees. We used cross validation with k=10 folds to identify the RMSE on the test dataset and selected GBM since it had the least test RMSE

The model is the crucial link which bridges the host's inputs to the predicted price and the integration will be done in the server function of the R shiny app

6. Functionality:

The R shiny app takes inputs (the values of the important features identified from the random forest model) from the host and passes them as an argument to the GBM. The GBM predicts the listing price based on the values or selection made by the hosts and prints them through a renderText()

7. GUI Design and Functionality:

We used selectInput, sliderInput and radioButtons functionalities to obtain the values of the features selected from the host. The layout of the app was designed using fluidRow. The host can select values from the given input functions and click on "Update" to obtain the estimated price for the new listing.

8. Conclusions:

We can extend this app to develop dynamic generation of word clouds for text attributes which the host needs to create while enlisting (eg. summary of the listing, transit options, house rules, etc.). We can also predict the occupancy rate for the listing and calculate the expected annual revenue based on the predicted price and occupancy rates. This functionality would enable a host to assess whether he/she will be able to recover the investment costs

9. References:

a. **R shiny layout:**
   i. https://shiny.rstudio.com/articles/layout-guide.html
   ii. https://www.rdocumentation.org/packages/shinybootstrap2/versions/0.2.1/topics/fluidPage
b. **Word Clouds:** https://datascienceplus.com/building-wordclouds-in-r/
c. **ggplot:** https://ggplot2.tidyverse.org/reference/scale_gradient.html