AIRBNB PRICE MODELING

HENRY CARPENTER

GENERAL ASSEMBLY – DATA SCIENCE

AUGUST 27, 2019

THE PROBLEM

- You’ve decided to list a bed, room, or apartment on Airbnb to make some cash on the side
    - How much do you charge?
- You manage a family-run hostel and have been hurt by Airbnb’s proliferation
    - What are they charging?
    - How should you price to compete?

ASSUMPTIONS OF THE MODEL

- We assume
  - There is nothing shady in the market - no collusion, predatory pricing, etc.
  - The price an Airbnb settles on is its fair market value
  - That price includes all taxes, fees, and other charges

MODEL PERFORMANCE METRICS

- Since price is what we care about, our key performance metrics all relate to the predicted price's relationship to the actual price
  - The residual is the difference between the true price and the predicted price

- Three main metrics
  - Mean residual
  - Median residual
  - Standard deviation of residual

THE DATA

- 22,600 individual listings
  - 96 columns
- 8.23m individual listing-dates
- Reviews of listings
- Information on neighborhoods

THE DATA

- The table of listings was most important
  - Most of the data was unusable
  - Much of the data was useless
  - Some of the data was so extreme that it was discarded
  - Ended up using 12 predictive columns

THE DATA

- Feature columns:
  - # of people it accommodates
  - # of bathrooms
  - # of bedrooms
  - # of beds
  - # of reviews
  - Host rating
  - Is instantly bookable (y/n)
  - # of reviews per month
  - Cancellation policy
  - Bed type (5 types)
  - Room type (3 types)
  - Neighborhood group (12 total)

THE DATA

- Limitations:
  - Cannot use the price of listing '*x*' on day 1 to predict the price for that same listing on day 2
    - Must remove duplicate listings by the listing ID
    - Cuts the dataset from 1m usable rows to 5,000
    - Cannot use date of listings as a feature column
  - No usable data for the overall quality of the listing
    - A lavish penthouse could appear identical to a 196 square foot boiler room/apartment
    - Can only use the data we have

THE MODEL

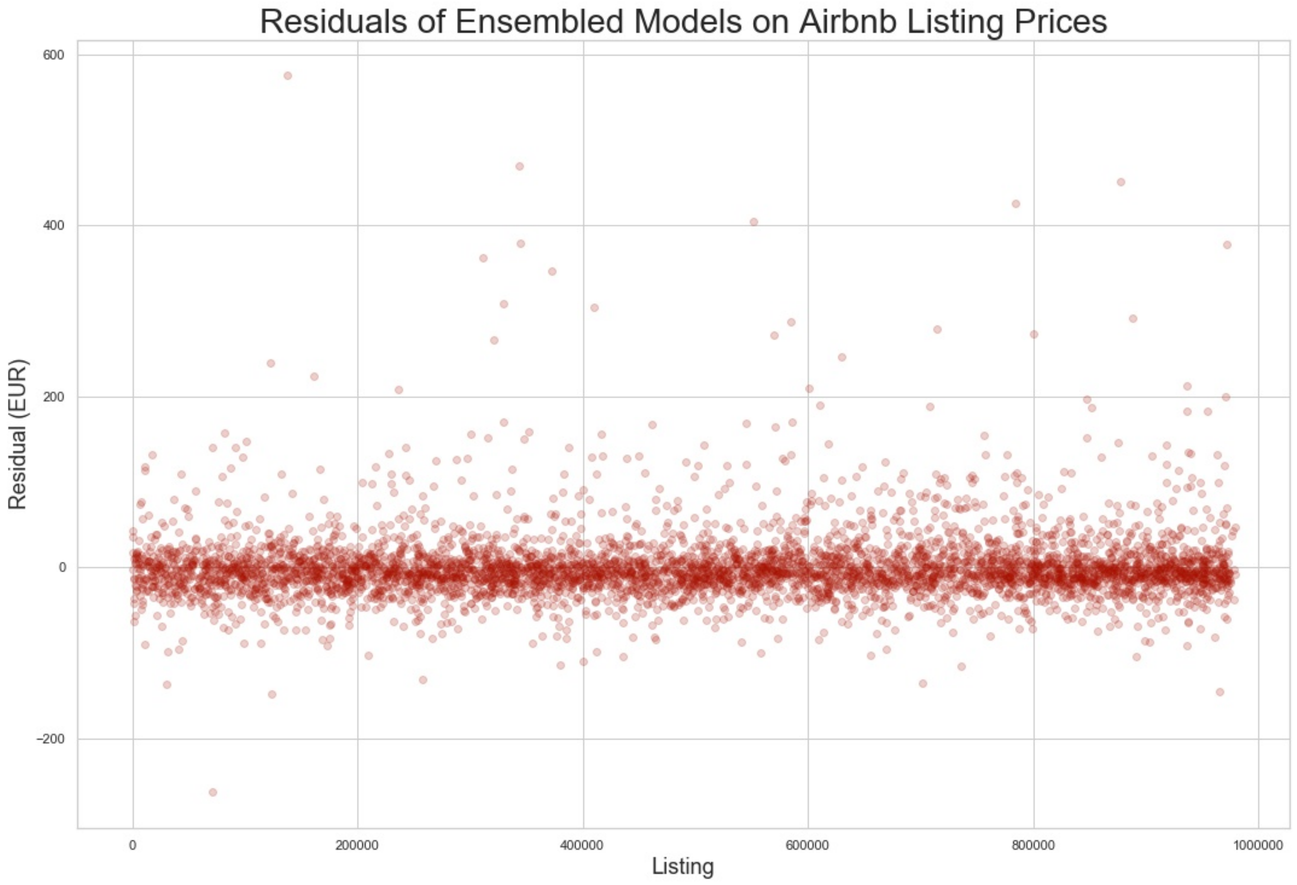
- Two approaches:
  - Linear regressor
  - Random forest regressor

- Average the two predicted prices together per listing to garner a more accurate overall result

PERFORMANCE

Model	Mean of residuals	Median of residuals	Standard deviation of residuals	Percent within $25 of true value
Linear Regressor	$24.48	$15.78	$41.05	67.7%
Random Forest Regressor	$24.71	$15.88	$41.11	69.2%
Ensemble	$23.61	$14.79	$40.19	70.3%

PERFORMANCE



CONSEQUENCES

- Cannot predict perfectly, but a good approximation
- Large inaccuracies are mostly for the very expensive listings
  - The predictions are 15% more accurate on average for listings of less than $200

DESIRED MODEL IMPROVEMENTS

- More data points
- Better representation of listing quality (penthouse v. boiler room)
- Incorporate dates into the model

ANY QUESTIONS?

THANK YOU!