Boston Airbnb Pricing

An investigation using machine learning

Group 7: Abhay Kasturia, Nakul Camasamudram, Philip Parker

Northeastern University

Content

Introduction

Background Information

Background:

- Airbnb is a popular online company where property owners (known as “hosts”) can short-term rent their spaces
- A host must decide what daily price to charge for his or her space
- With data, this is clearly a supervised ML problem
- Issue: many real estate datasets include categorical location features with large numbers of levels

With this in mind, in this project we investigate two questions:

Questions to Investigate

1. To what degree can supervised machine learning techniques be used to assist an Airbnb host in determining an appropriate listing price for their property?
2. For Airbnb data, can the categorical feature of “neighborhood” be replaced with a continuous feature of driving distance to a geographic point of interest (e.g., an airport) and have comparable results?

Methodology

Data Collection and Preprocessing

Data Source: Inside Airbnb Project - 4870 rows, 96 columns

Features Selected:

- host_is_superhost
- host_identity_verified
- neighborhood
- property_type
- room_type
- accommodates
- bathrooms
- bedrooms
- beds
- bed_type
- guests_included
- minimum_nights
- number_of_reviews
- instant_bookable
- is_business_travel_ready
- cancellation_policy
- price

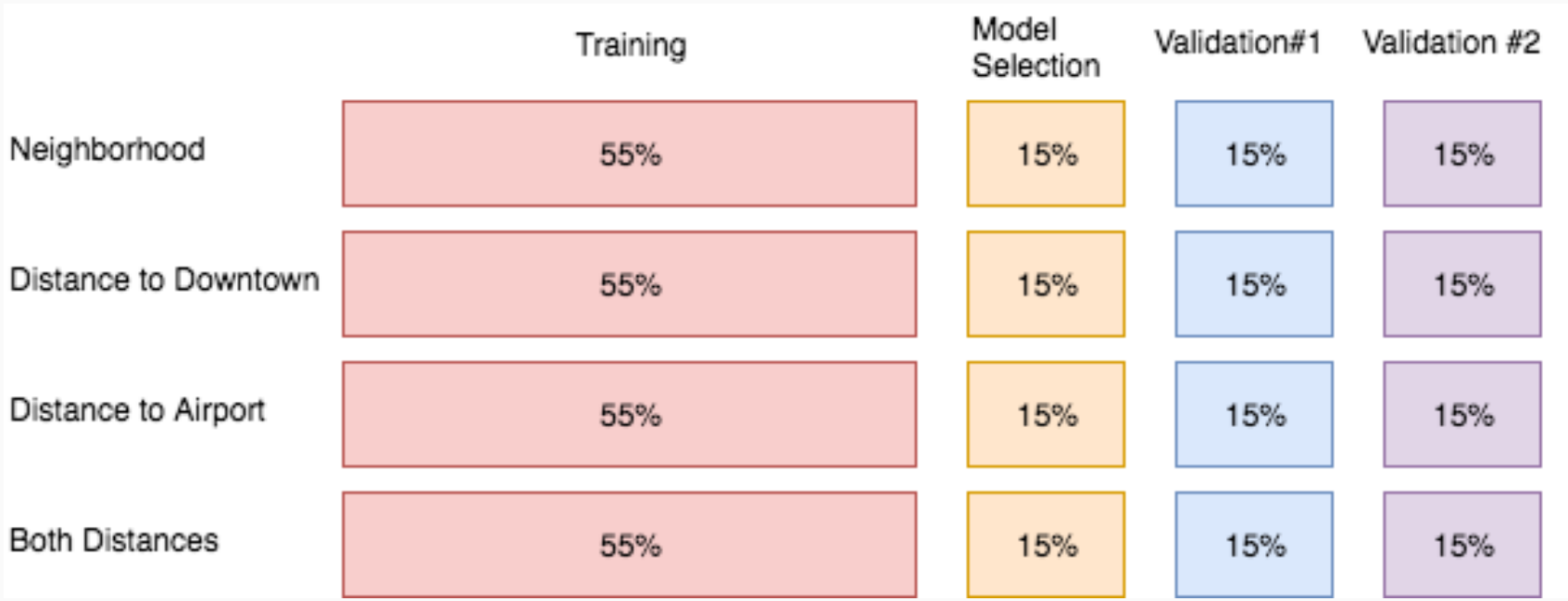
Outliers: 240 rows removed above the 95th percentile for price

Missing Values: 14 rows with missing values removed

Data Transformations and Partitions

Transformations: Neighborhood feature replaced with distance to BOS Airport, to Downtown Crossing, and both

Partitions: Data partitioned into training, model selection, and two validation sets



Applying Methods

Methods Chosen: Linear Regression, GAMs, and Regression Trees

- For each method, we apply it to the different transformations in order to compare the effect of replacing the “neighborhood” feature
- Afterward, we take the best transformation/method combination, and predict using the second validation set for a quality assessment of performance

Linear Regression

Models: All predictors, Subset Selection, Lasso, Ridge

Best: Lasso with $\lambda = 0.1135332$ and 58 predictors

Generalized Additive Models

Setup: Cubic splines for continuous features

Term selection:

- Regression Subset Selection
- Forward Selection
- **Smoothness penalties with additional shrinkage term**

Regression Trees

Considerations:

- Boosting used, as the importance of interpretability is small for our questions
- Cross-validation used to select number of trees = 10,000

Results

Most Important Predictors

Linear Regression: neighborhood.South.End, neighborhood.Downtown, neighborhood.Beacon.Hill, neighborhood.Back.Bay, room_type.Shared.room, room_type.Private.room, property_type.Other, bedrooms, accommodates

GAMs: neighborhood, property_type, room_type, instant_bookable, accommodates, bedrooms, guests_included

Regression Trees:

	NVar	NRel	DDVar	DDRel	DAVar	DARel	DBVar	DBRel
First	room_type	46.9	room_type	50.6	room_type	51.7	room_type	50.1
Second	neighborhood	20.3	bedrooms	15.7	dairport	15.7	bedrooms	15.8
Third	bedrooms	15.7	ddowntown	15.1	bedrooms	15.1	ddowntown	11.5

Final Results

Method/Transformation Comparison:

	Linear Regression	GAM	Regression Trees
Neighborhood	56.82	55.15	55.75
Distance to Downtown	58.95	58.05	57.00
Distance to Airport	60.76	56.24	56.57
Both Distances	58.16	55.77	56.54

Best Method/Transformation Performance:

GAM on the original dataset with MSE = 52.30

Discussion

Question 1: Performance Useful?

Question: To what extent can supervised ML techniques be used to assist a host in determining listing price?

Answer: Useful, but information outside of these features is important.

Question 2: Transformations Reasonable?

Question: For Airbnb data, can the "neighborhood" feature be replaced with distance to a geographic point of interest?

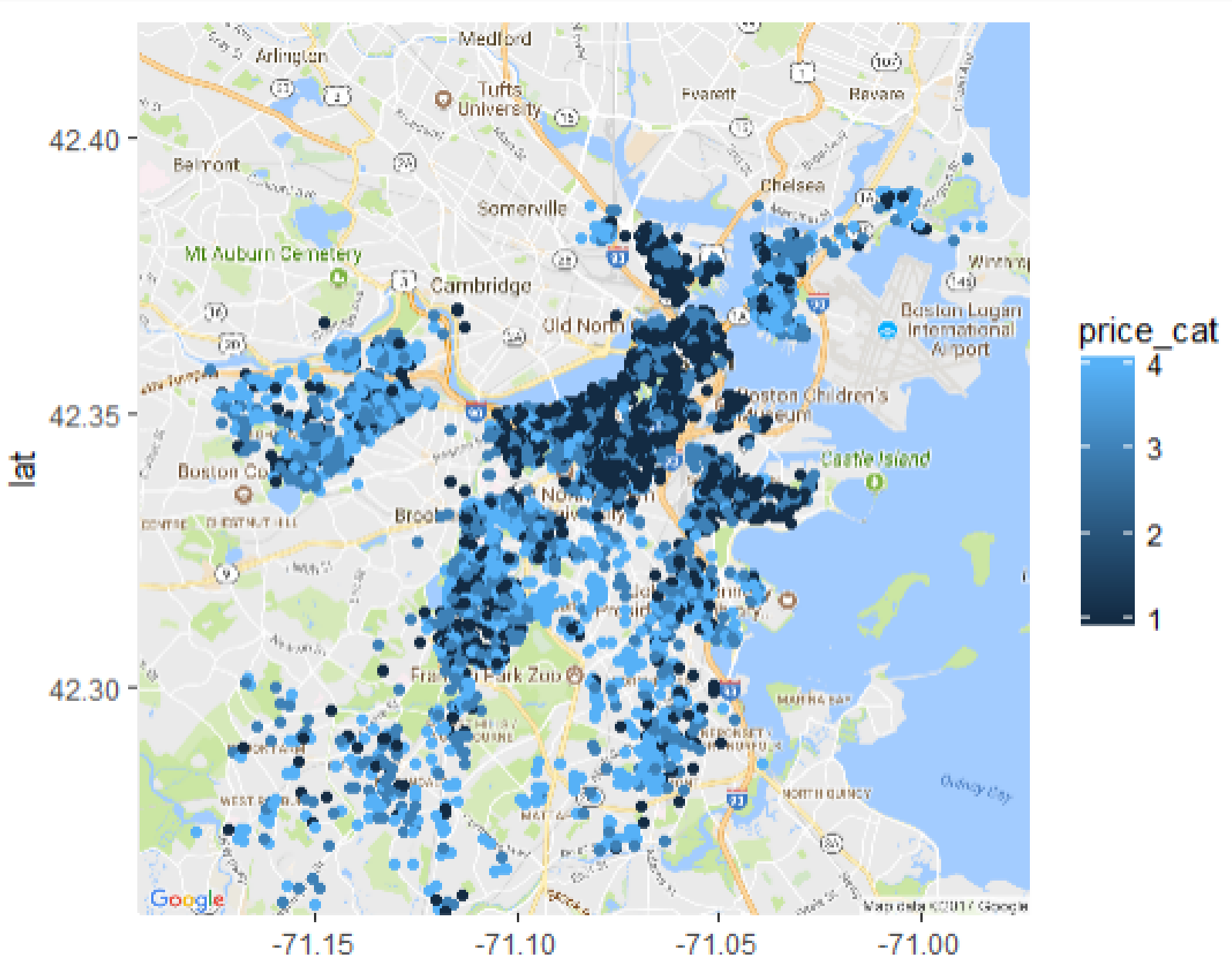
Answer: Yes.

Potential Improvements

- More sophisticated selection of POI
- Try to access information contained in textual data (e.g., sentiment analysis in reviews)

Visualizations

Listing price distribution across Boston (Darker = Expensive)



Audience Questions

Questions?