Predicting New York City Airbnb Prices

Obtaining an accurate predictive model with business implications

Data Science – ELTECON
Rebeka Éva Cook (ZN5K7X)
Jenő Tóth (JSK7AT)
15/12/2021

Agenda

1. Introduction
2. Dataset and Methodology
3. Results
4. Business Implications and Conclusion

Agenda

1. **Introduction**
2. Data and Methodology
3. Results
4. Business Implications and Conclusion

Airbnb has gained popularity over the past decade, changing the dynamics of the housing market

airbnb

Airbnb as a platform

- Many people have shifted from long-term rentals to Airbnbs
- Airbnb is a platform for **short-term accomodations** such as holiday rentals and tourism
- 5.6 million active users over 100,000 cities
- The price of listings vary by **numerous characteristics**

Our goal

- Choose the most important **predictors of price**
- Gain insight into the variables related to price and how they change
- Obtain the most accurate **price predicting model** with our given variables by minimizing the root mean squared error
- Direct our model to **Airbnb beginners**

Our research question

Predicting New York City Airbnb prices based on observable variables with business implications

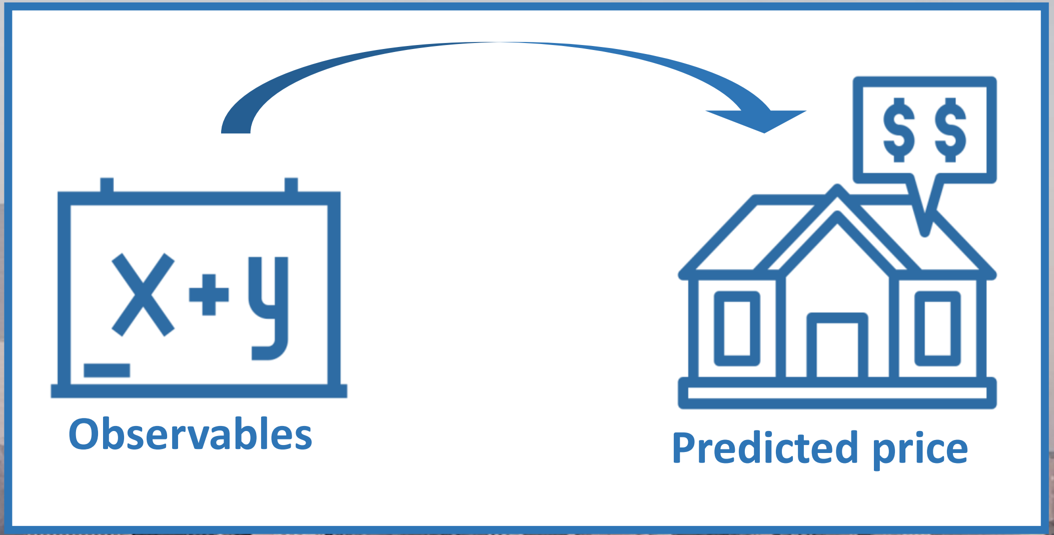
Agenda

1. Introduction
2. **Dataset and Methodology**
3. Results
4. Business Implications and Conclusion

We used data from ‘Inside Airbnb’ to conduct our methodology with price as the dependent variable

Our dataset

- **New York City** Airbnb data from ‘Inside Airbnb’
- **74 variables** over more than **3price5,000 observations**
- Our dependent variable was



1. Step: Cleaning the data

Although are dataset was very thorough, it left significant room for **cleaning**.

2. Step : Exploratory data analysis

Since we had an abundance of variables, we needed to gain **insight** to choose the best predictors

3. Step: Predictive models

We compared many different models based on **RMSE** to obtain the **best predicions**

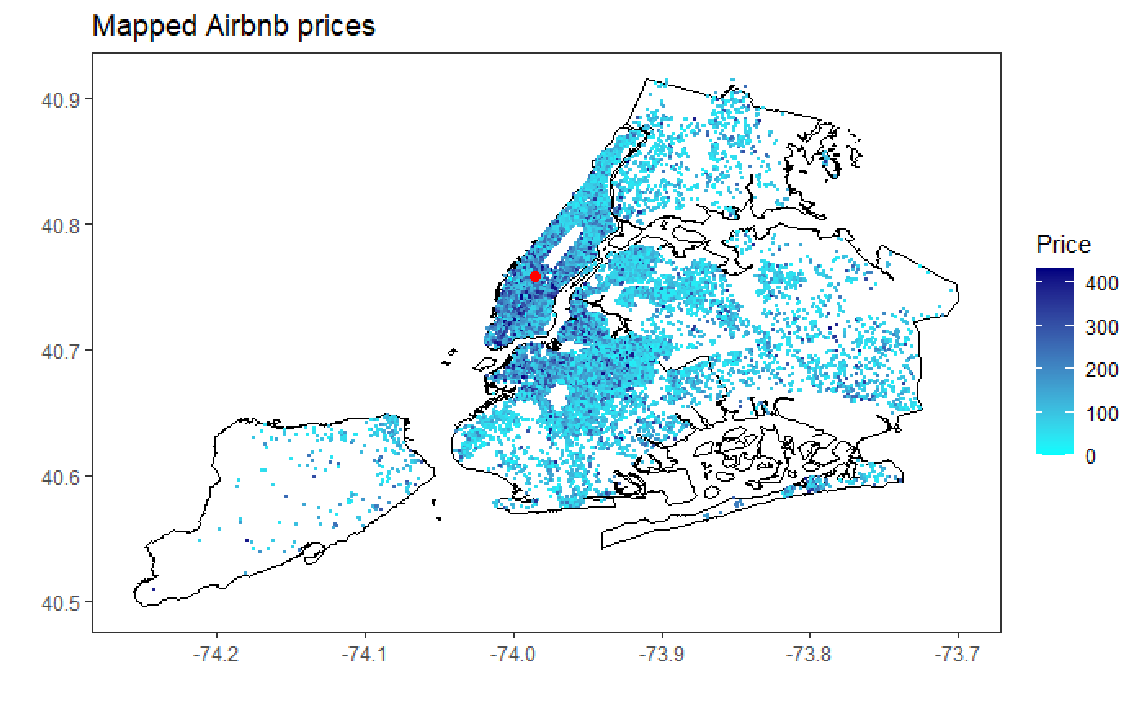
The explanatory variables we used

We created two important new variables

Sentiment score:

"Beautiful, spacious skylit studio in the heart of Midtown, Manhattan…" ▶ 13

Distance from Times Square:



Other variables we used

Host-related variables
- Host account age
- Host response/acceptance rate
- Is the host a superhost?
- Host’s number of Airbnb listings
- Does the host have profile picture?
- Is the host verified?

Airbnb-related variables
- Airbnb type (home, private room, etc.)
- Number of accommodates
- Number of bathrooms, bedrooms, etc.

Review-related variables
- Number of reviews (last 12 months)
- Review scores

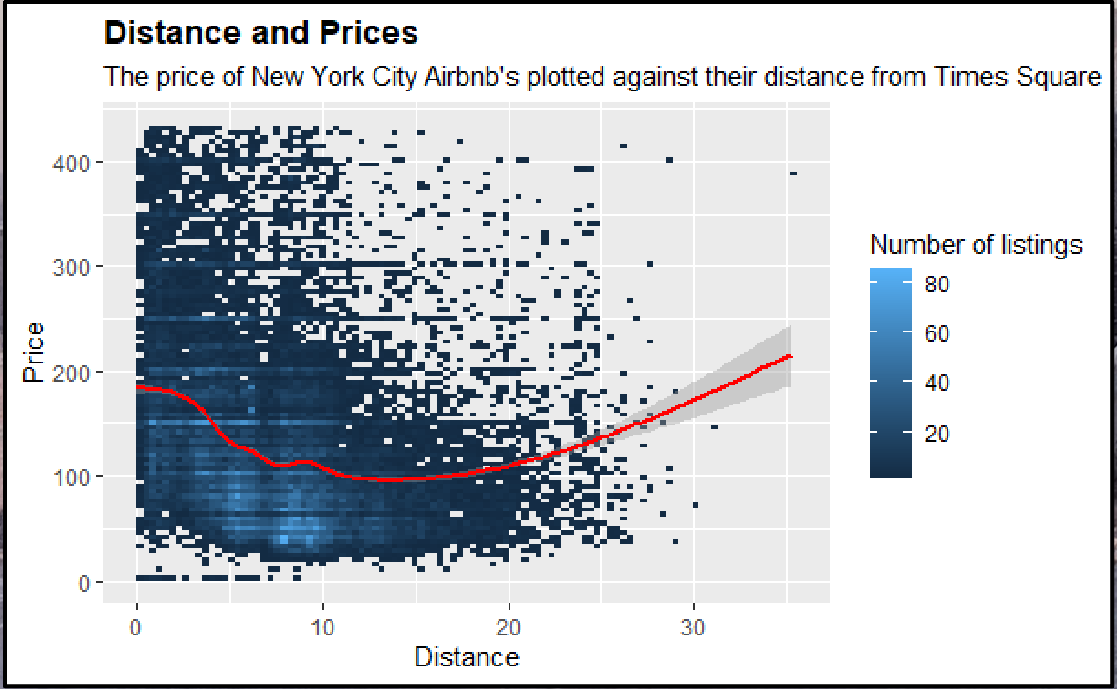
Agenda

1. Introduction
2. Dataset and Methodology
3. **Results**
4. Business Implications and Conclusion

In our baseline model our only predictor of price is the distance from the city centre

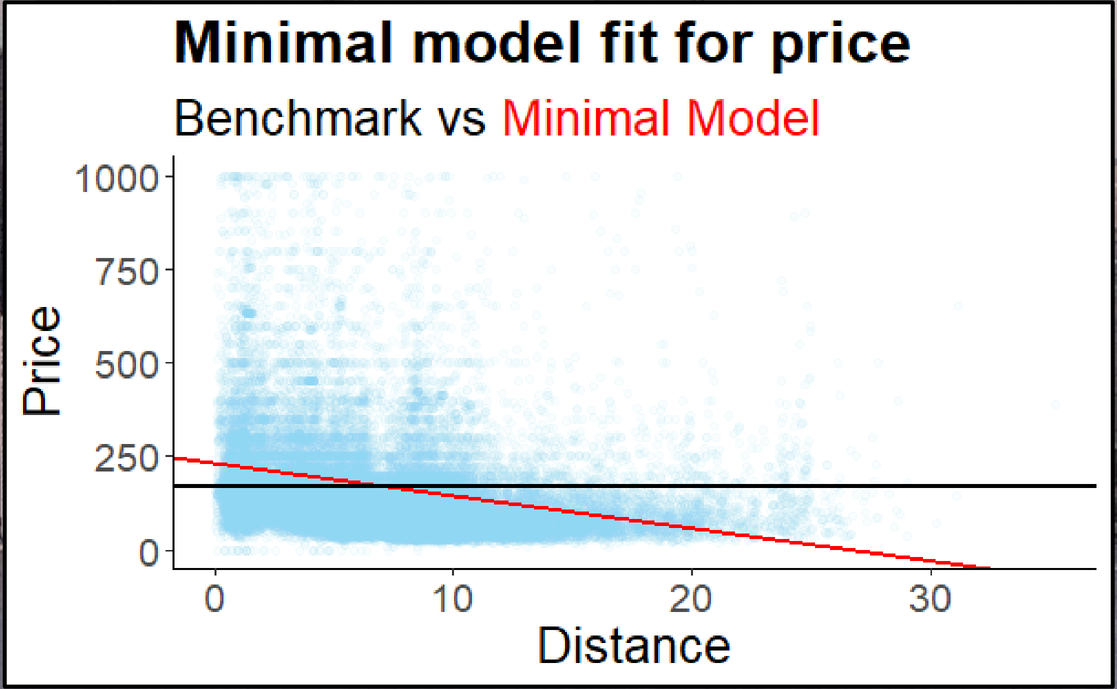
The relationship between distance and price

- Most listings are **within 5 to 10 kilometres** from the centre
- Generally **prices decrease** as we go **farther away**

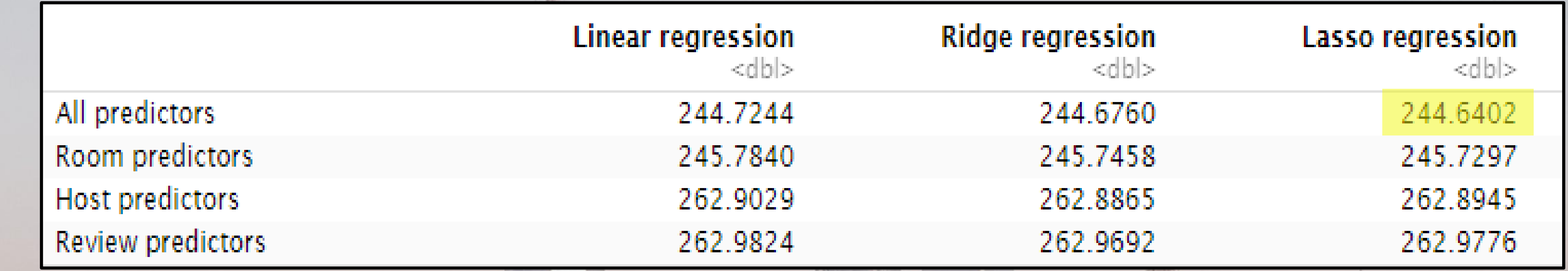


The baseline predictive model

- The baseline model captures an overall trend
- However, it is not adequate enough for accurate predictions with an **RMSE = 296.29**



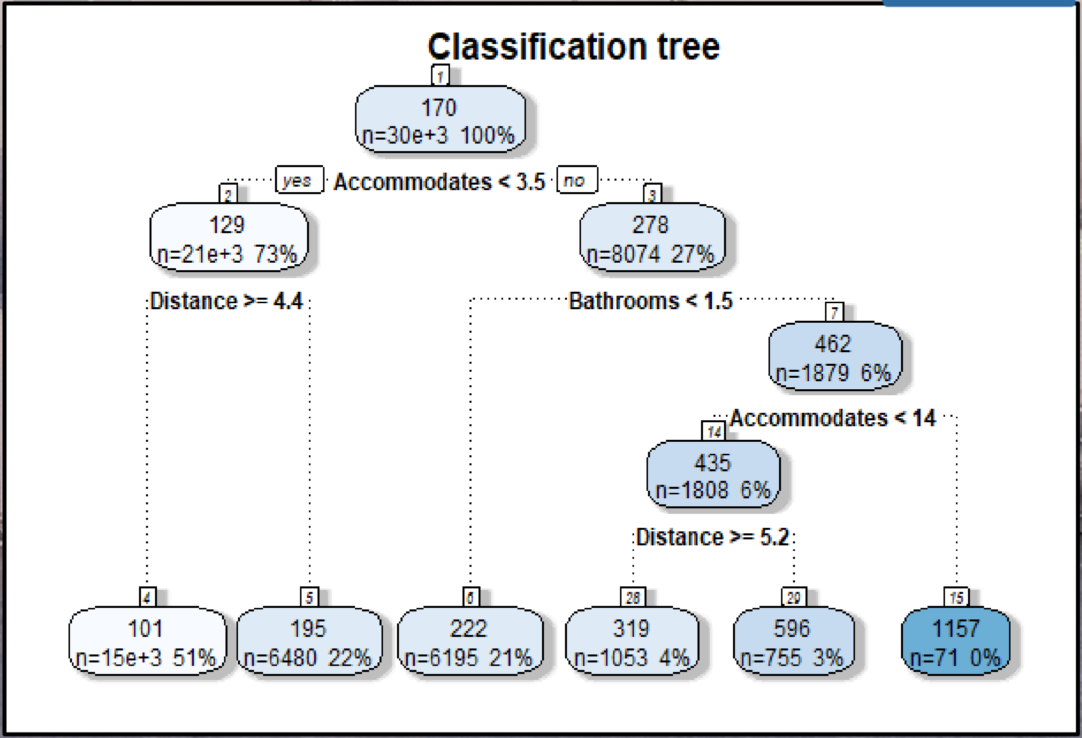
To find the best possible price prediction, we tried 12 different models and a random forest



	Linear regression <dbl>	Ridge regression <dbl>	Lasso regression <dbl>
All predictors	244.7244	244.6760	244.6402
Room predictors	245.7840	245.7458	245.7297
Host predictors	262.9029	262.8865	262.8945
Review predictors	262.9824	262.9692	262.9776

Steps of prediction

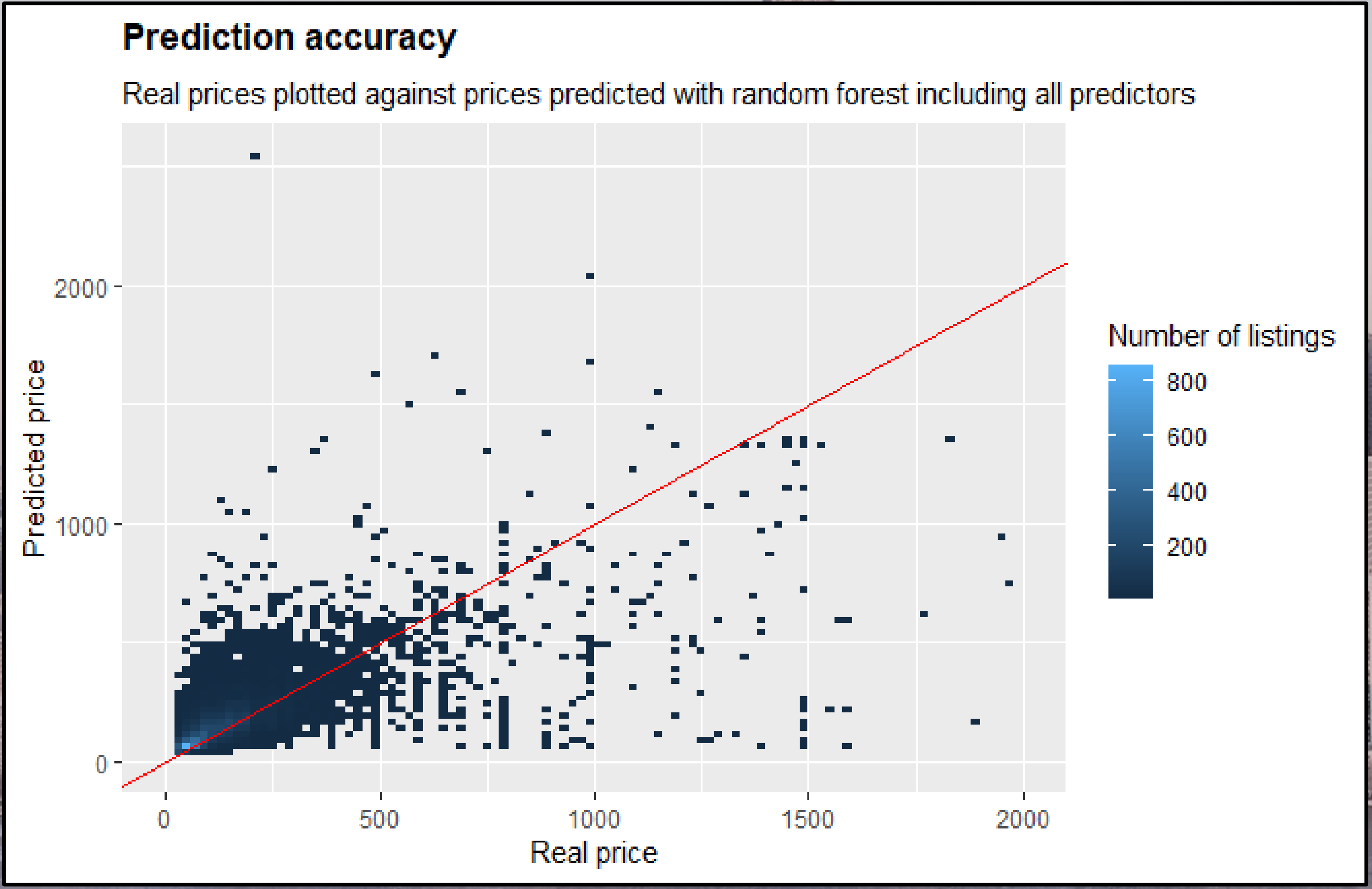
- We tried **3 different model types**
  1. Linear regression
  2. Ridge regression
  3. Lasso regression
- Since the **lasso regression** was the best with **RMSE=244.64**, we also ran a **random forest**
- The random forest prediction was the best, with a **RMSE=233.25**
- A random forest is a collection of **decision trees** such as the one on the right



Agenda

1. Introduction
2. Dataset and Methodology
3. Results
4. **Business Implications and Conclusion**

The best model to predict Airbnb prices is a random forest model using all relevant observable variables

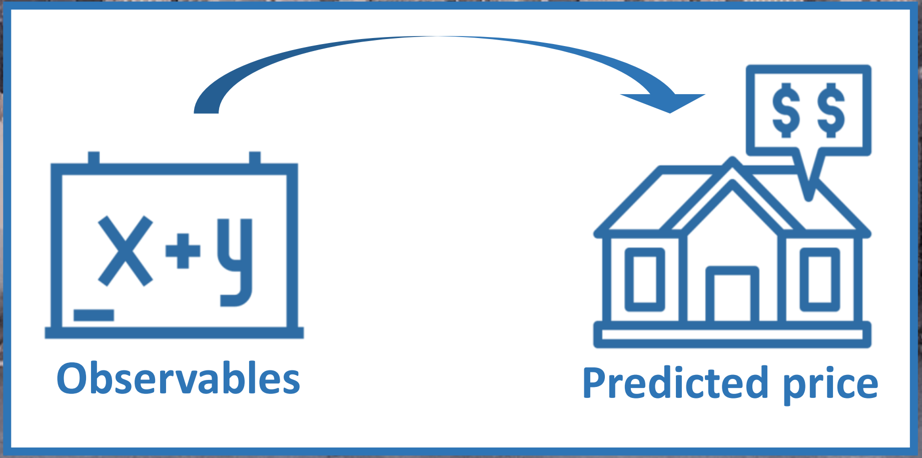


Prediction accuracy

Real prices plotted against prices predicted with random forest including all predictors

General observations

- The model works best for **lower prices** (light blue area)
- The prediction is useful for **an average Airbnb user**, typically beginners or amateurs
- The model gives a relatively accurate **guideline to new hosts**



Observables

Predicted price

Thank you for your attention!

Do you have any questions?