Predicting the Nightly Price of Toronto’s Airbnb Listings

SPRINGBOARD—CAPSTONE PROJECT 1

BY GEORGE TANG

SEPTEMBER, 2019

What is AirBnb?

- Online marketplace for offering lodging (primarily homestay) and tourism services
- Hosts use the platform to list their properties for accommodations, in which Airbnb receives commissions
- Started with the founders’ idea of “putting an air mattress in their living room and turning it into a bed and breakfast for a few bucks” in 2008
- Now a multinational company, with annual revenue of $2.6 billion in 2017

Hosts' pricing challenge

- Pricing the listings too high or too low may drive customers away, or leave money on the table
- Difficult to identify properties with similar features in the vicinity for reference
- Hard to identify features that affect prices

Potential Solution

- Data analytics to thoroughly examine Toronto’s Airbnb market
- Identify features that affect price
- Build a machine learning model with historical listings data that predicts prices

Project Tasks

1. Data cleaning, wrangling, quality checking and feature engineering (covered in report)
2. Explore Airbnb’s listing dataset and identify features that may affect price
3. Develop and evaluate machine learning models for price prediction
4. Provide Recommendations to hosts and investors, and suggest future works

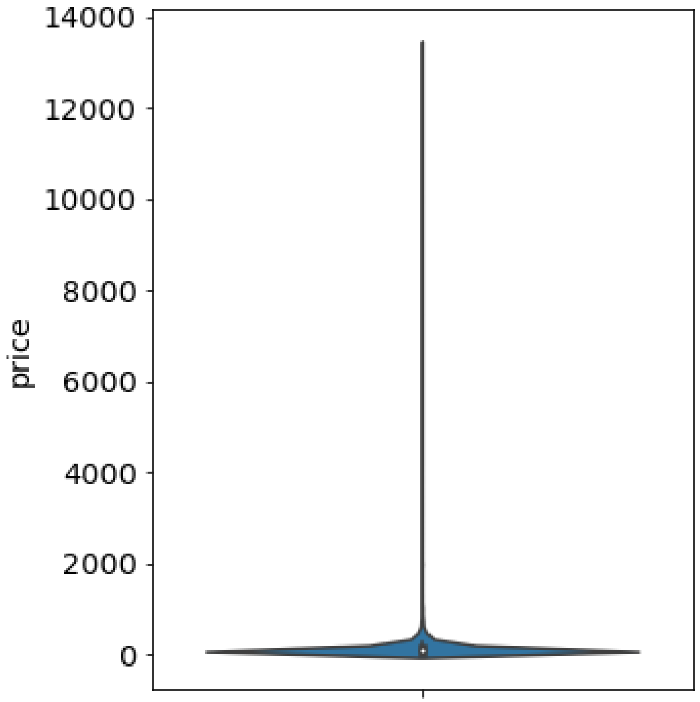
Dataset

- Location: Toronto
  - Largest city in Canada; 4th largest in North America
- 20769 listings, 106 features
- Feature categories:
  - Host information (name, host location, superhost status, response rate, etc)
  - Geographical information (longitude, latitude, zipcode, neighbourhood, etc)
  - Property information (number of beds, bedrooms, bathrooms and accommodates, amenities, etc)
  - Booking information (price, cleaning fee, cancellation policy, min & max nights of stay, etc)
  - Availability (current, number of available days in the next 30, 60, 90, 365 days, etc)
  - Customer reviews (Overall, accuracy, cleanliness, host response, etc)
  - Airbnb listing information (listing urls, host picture urls, etc)
  - Web scrapping information (date scraped, scrape id)

Price

Highly right skewed

Lowest price: $13, Highest price: $13,422

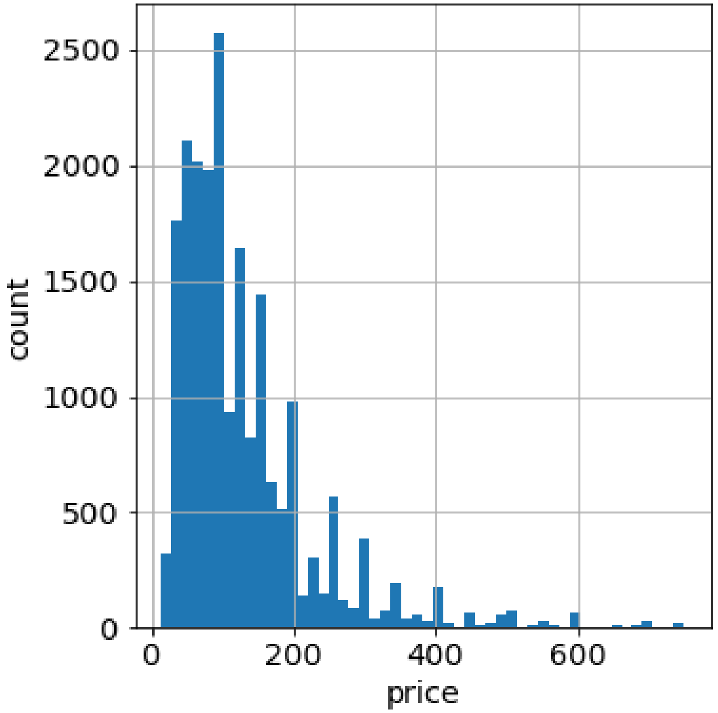


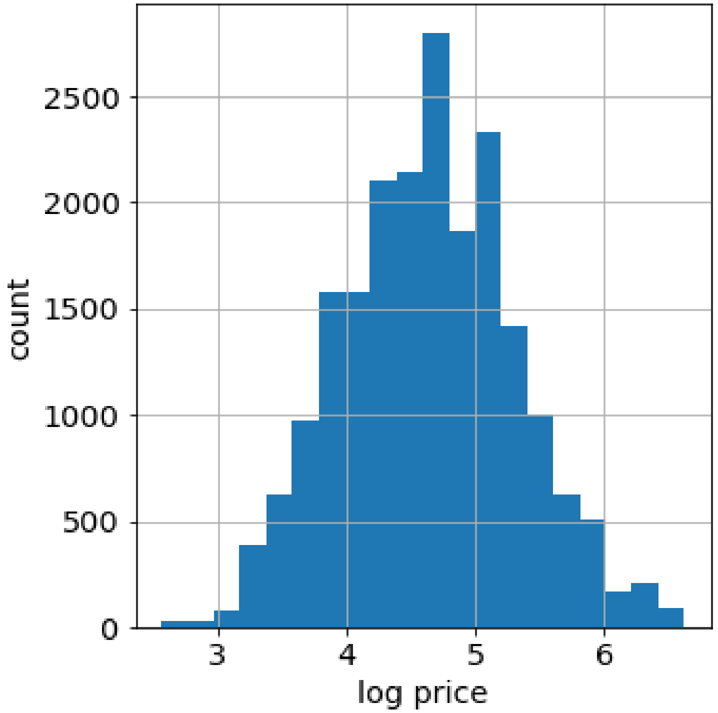
Art Collector’s Penthouse ($13,422/night)

Source: Airbnb

Price

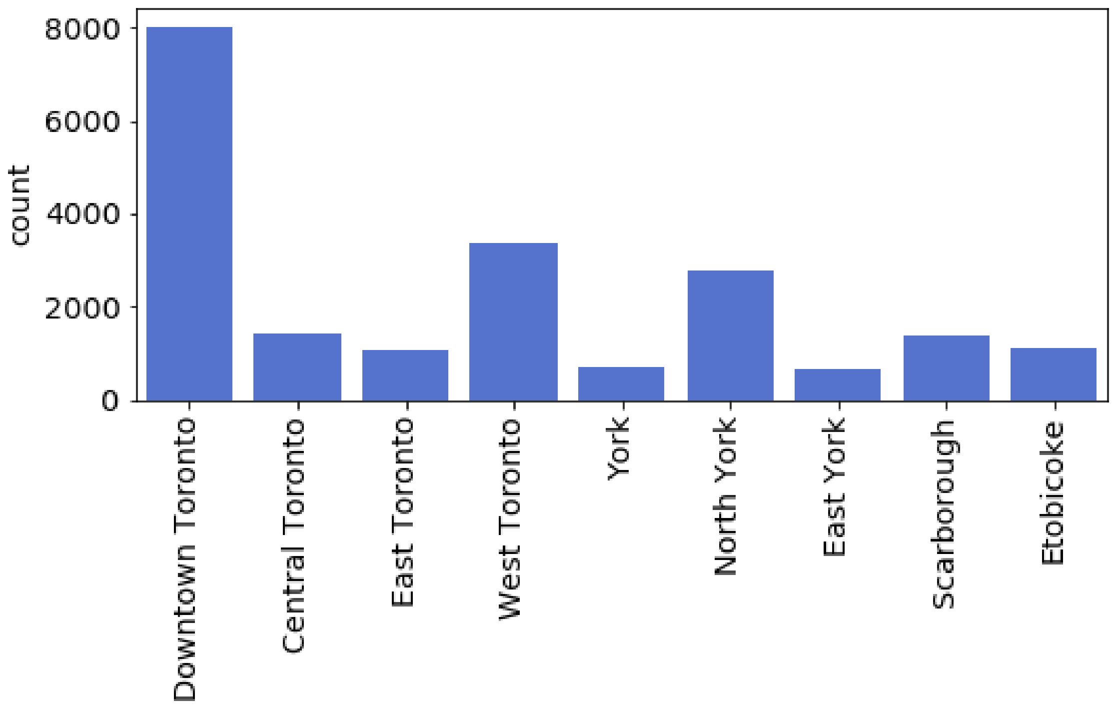
- Price capped at 99th percentile value ($750) to reduce influence of outliers
- Log transformation to normalize data

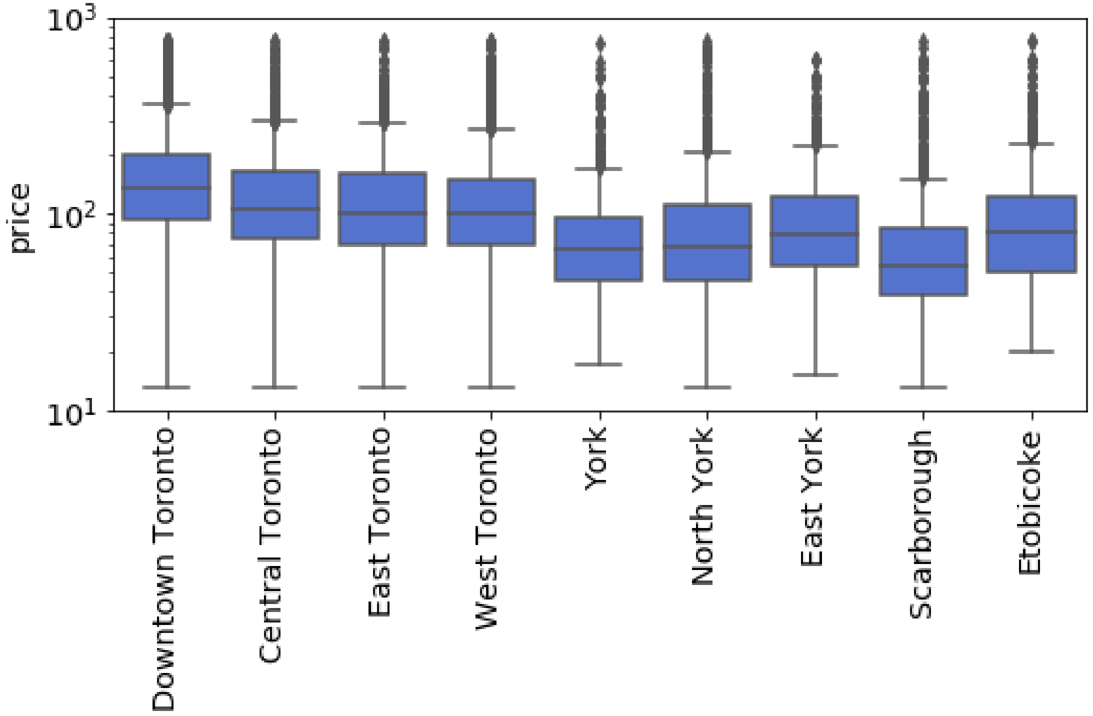




Geographical Location

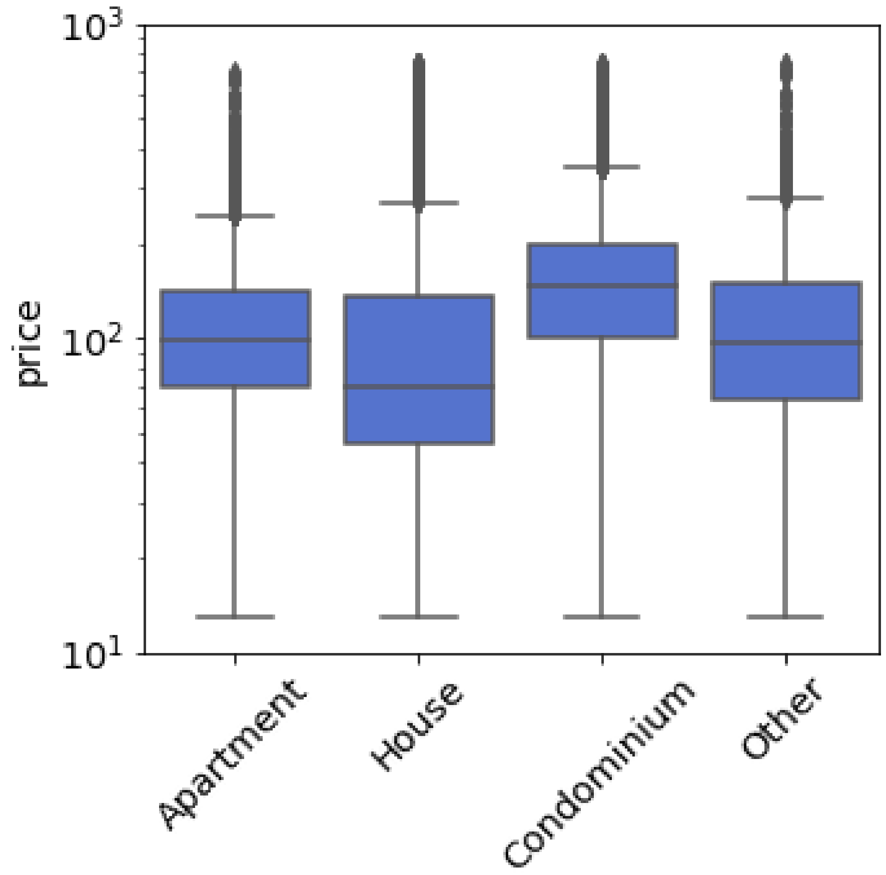
- Downtown Toronto has the most listings, also the most expensive (median price)





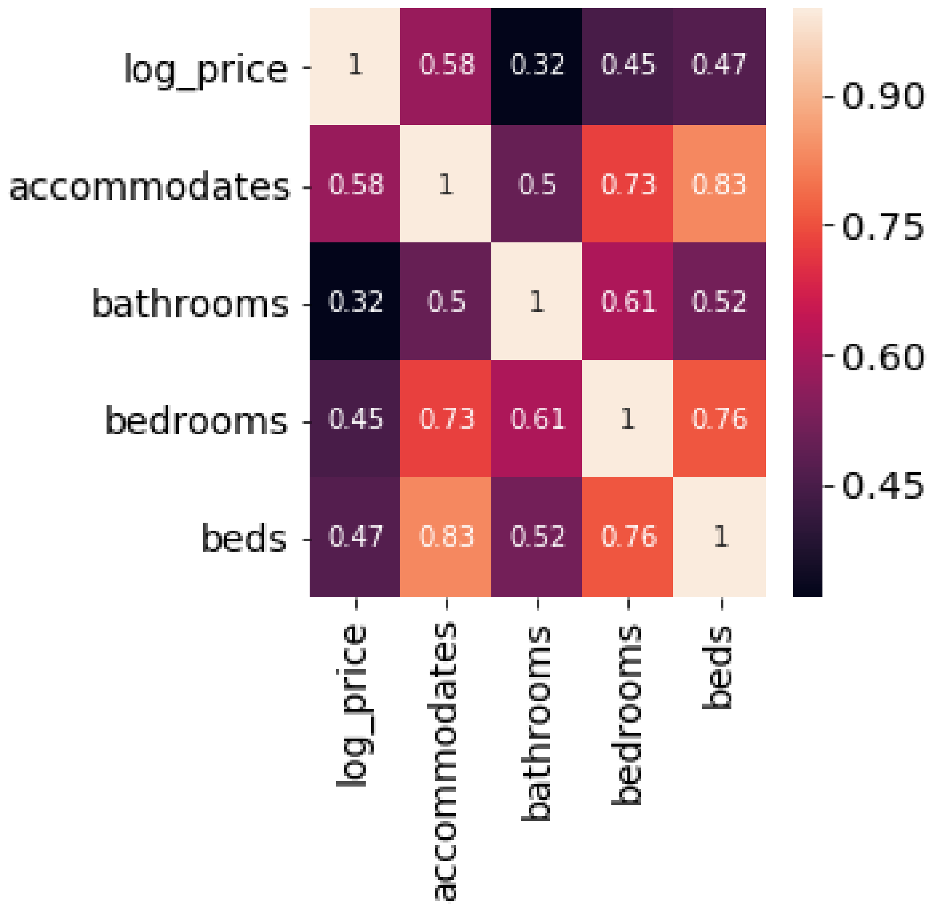
Property Type

- 30 unique property types, with 16 of them fewer than 10 listings
  - Rare types: parking space, tree House, earth house, tent, cave, aparthotel, castle …
- Bungalow assigned to “house”
- Rare categories assigned to “Other”
- Condos the most expensive, houses the least expensive



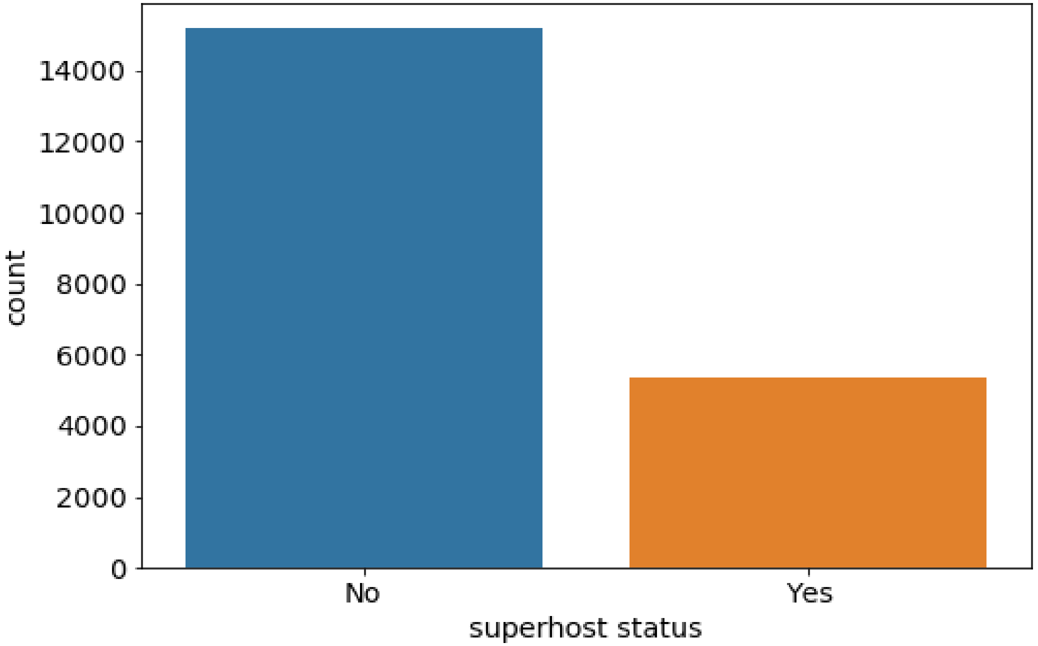
Property Features

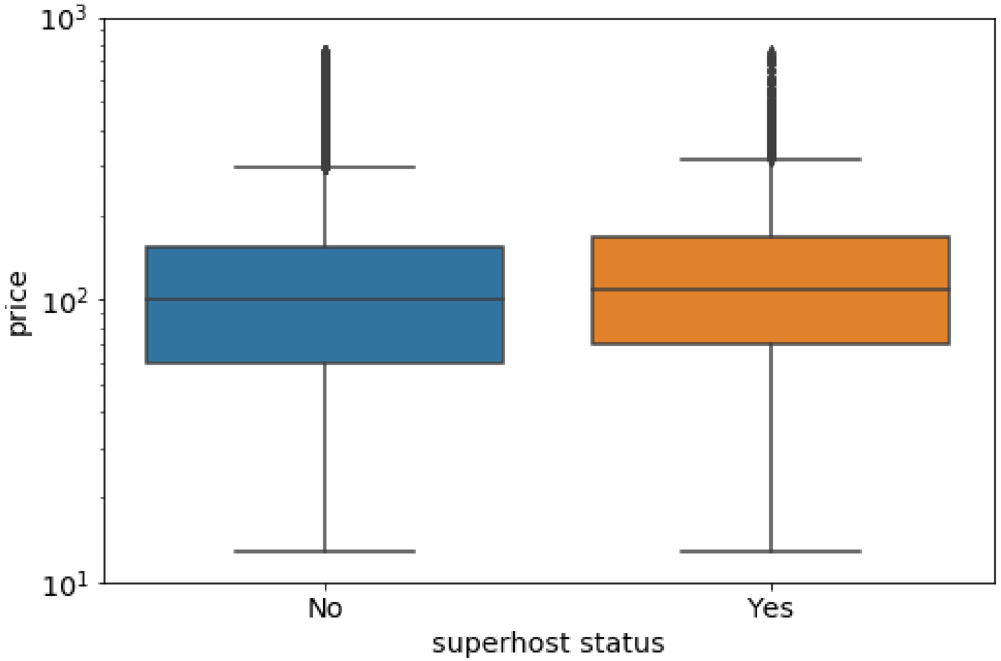
- Log price is somewhat positively correlated with no. of accommodates, bathrooms, bedrooms and no. of beds
- High correlation among those features



Host: Superhost status

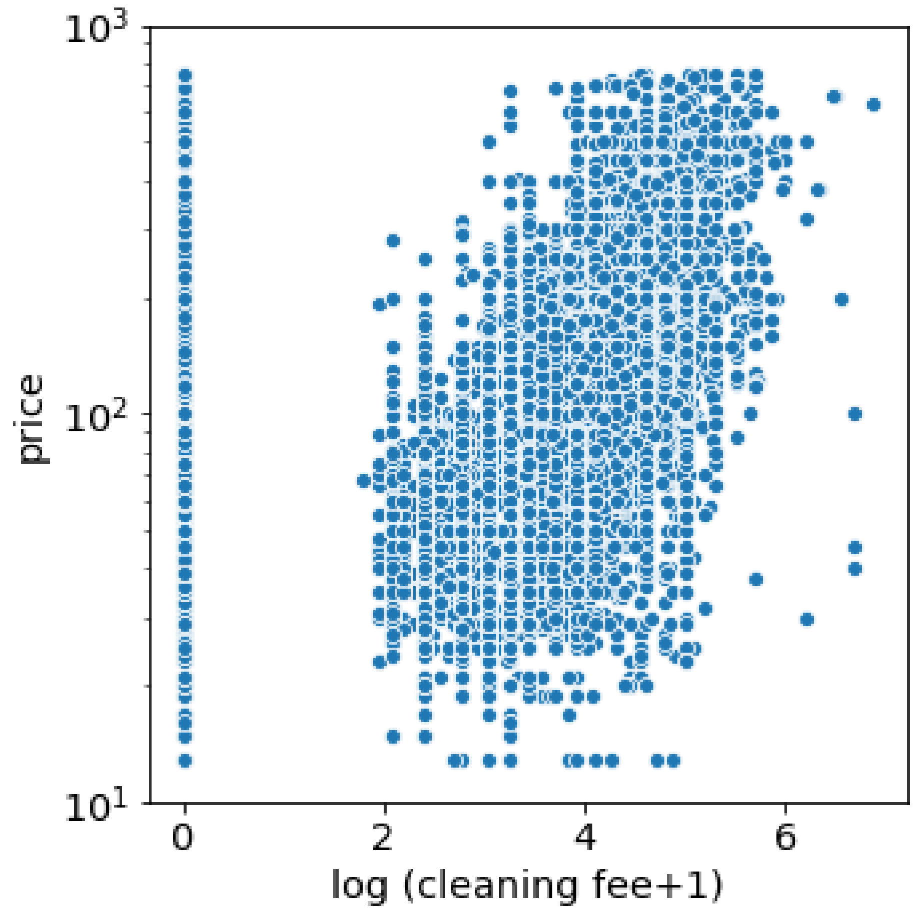
- Status granted by Airbnb to “experienced hosts who provide a shining example for other hosts, and extraordinary experiences for their guests”
- About 26% of listings provided by superhosts
- Median prices for superhosts slightly higher





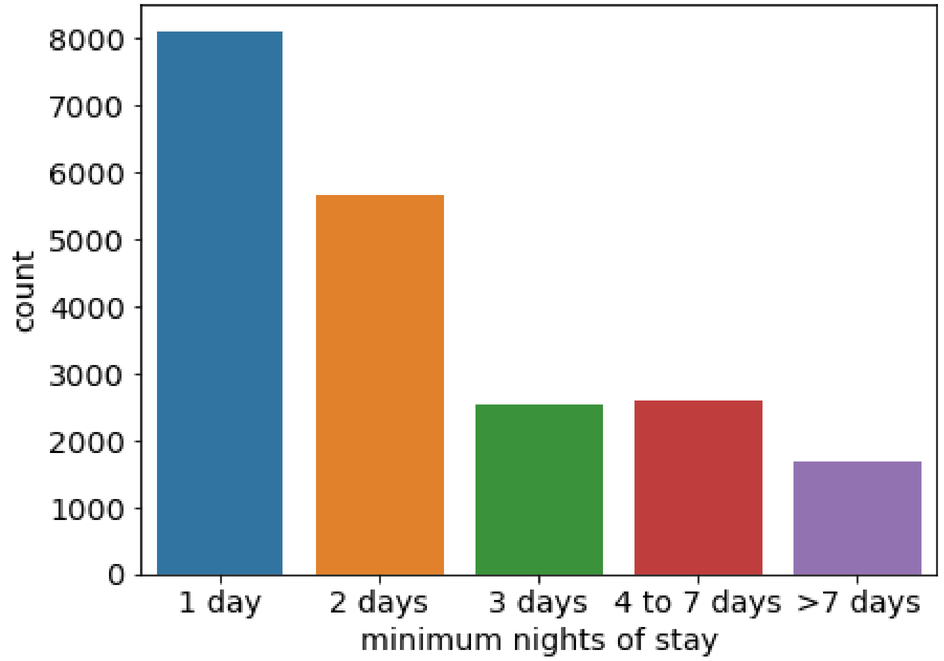
Cleaning fee

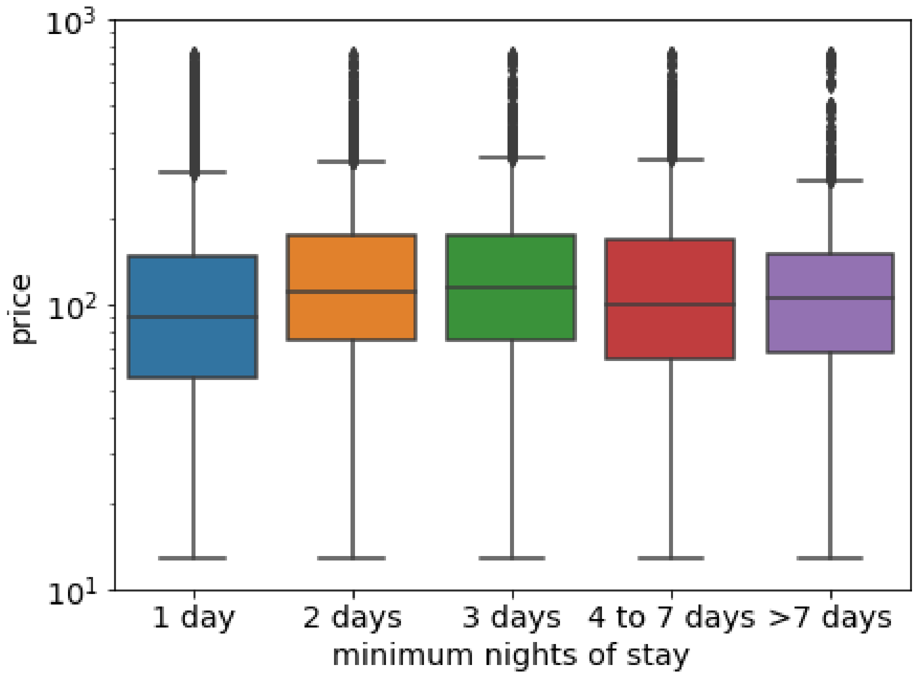
- One time, fixed fee charged to customers regardless of the duration of stay
- Some hosts opt to not charge cleaning fee at all
- For those do, the fee is somewhat correlated with price



Minimum nights of stay

- Minimum of 1 night; maximum of 1125 night (over 3 years!)
- 79% of listings require minimum stays of 3 days or less





Machine Learning

- Models that can be used by hosts to predict prices based on their property features and booking policies
- Log price as the target variable

1. Normalize and split dataset into 75% training data / 25% test data
2. Hyperparameter tuning
   - Linear regression Model (baseline model, no tuning)
   - Random forest (out-of-bag sample validation)
   - Gradient boosting (grid search + cross validation)
   - XGBoost (grid search + cross validation)
3. Build models with best hyperparameters
4. Compare models with test data

Model Performance on Test Data (Log Price as Target)

- The three advanced models are superior to the linear model
- XGBoost has the best performance metrics

	Linear Regression	Random Forest	Gradient Boosting	XGBoost
R^2	0.654	0.708	0.726	0.732
RMSE	0.396	0.364	0.352	0.348
MAPE	6.538	5.868	5.736	5.628

***MAPE: mean absolute percent error**

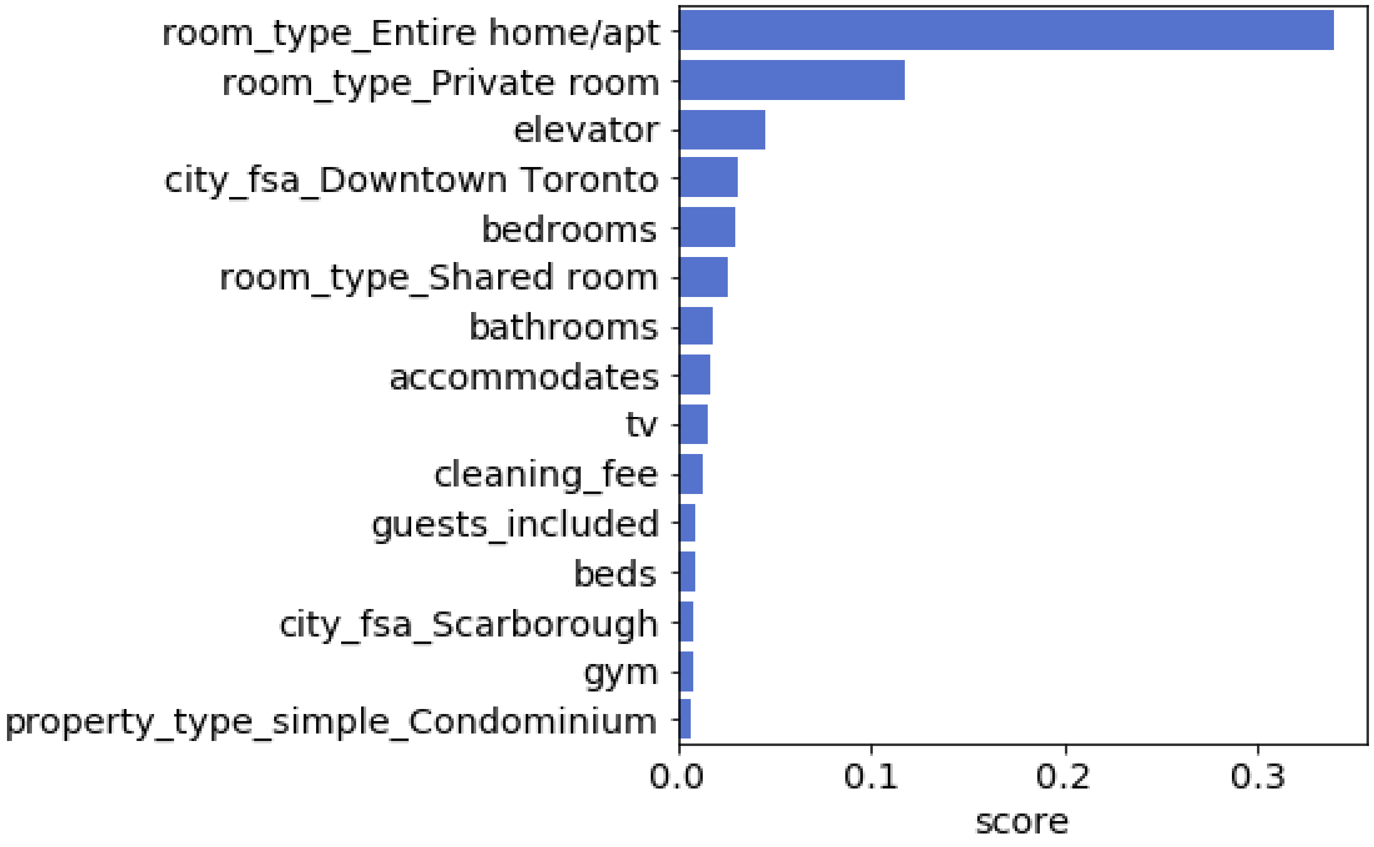
Model Performance on Test Data (Price as Target)

- XGBoost also has the best performance metrics
- 95% chance the predicted price is between -$82 (overpredict) and $150 (underpredict) from actual price
- 95% absolute percentage error below 72%

	Linear Regression	Random Forest	Gradient Boosting	XGBoost
R^2	0.468	0.607	0.638	0.647
RMSE	73.74	63.39	60.81	60.10
MAPE	32.20	28.55	27.79	27.31

Important Features

- Room type the most important feature
- Location (Downtown Toronto) and number of accommodates also on the list
- Some amenities (elevator, tv, gym) also on list



Recommendations

- Examine causes if listing price is significant different from predicted price
  - Compare listings to those with prices similar to predicted price
  - Adjust price to increase revenue or increase competitiveness
- Consider offering entire house / apartment instead of private room or shared room
- Consider home improvements to increase number of accommodates

Future Works

- A product that enable hosts to compare their listings to others with similar price and/or features
  - Allow hosts to know how their properties compare with competitors and adjust prices accordingly
- Split dataset into “economy” and “luxury” categories based on price
  - Models can be more generalizable to the different categories
- Analyse text features and customer reviews
  - Understand what matters to customers most
- Understand pricing dynamics
  - Strategize pricing based on day of week, holidays, special events to maximize revenue

Thank you!

Contact Information

- George Tang, aspiring data scientist
- Email: georgecctang@gmail.com
- Linkedin: https://www.linkedin.com/in/george-tang-b2b8005/