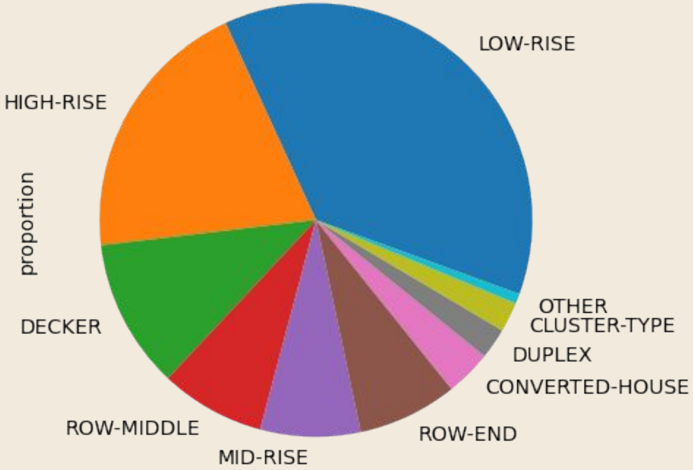
Investigating Brookline Real Estate Market

Is Machine Learning or are Algorithms better at predicting the price of a property?

Grant Corbett

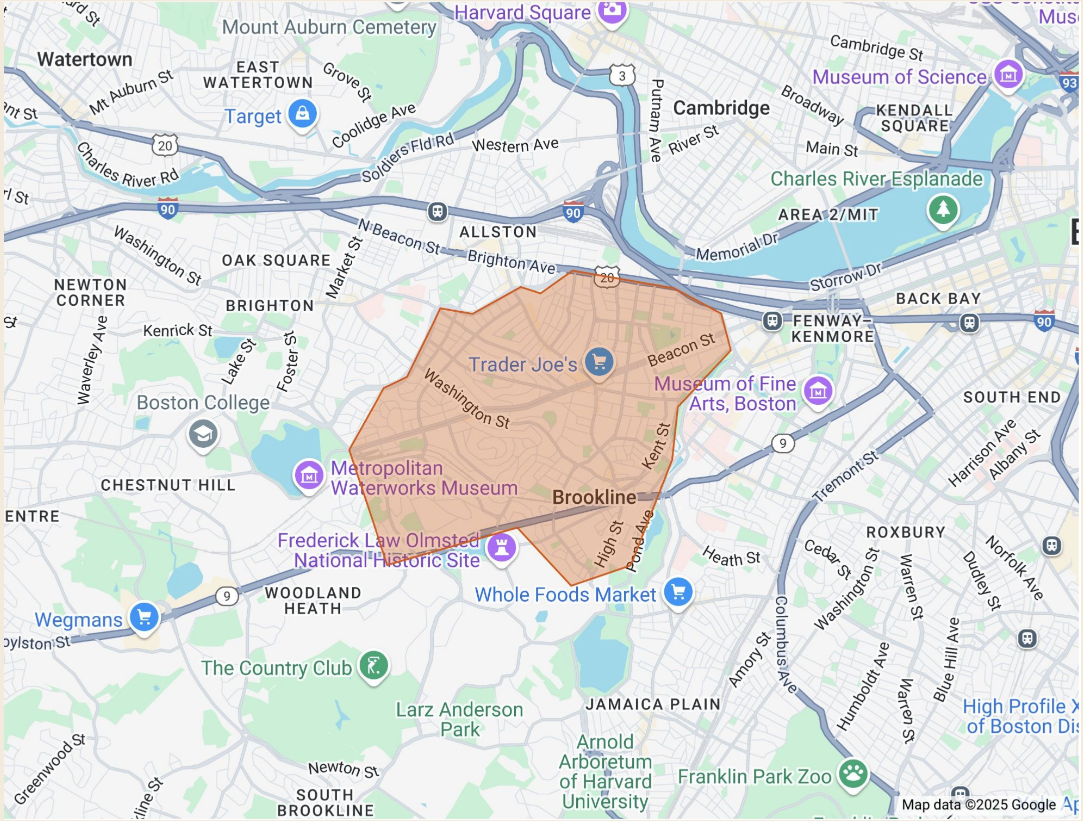
Brookline Data Set Snapshot

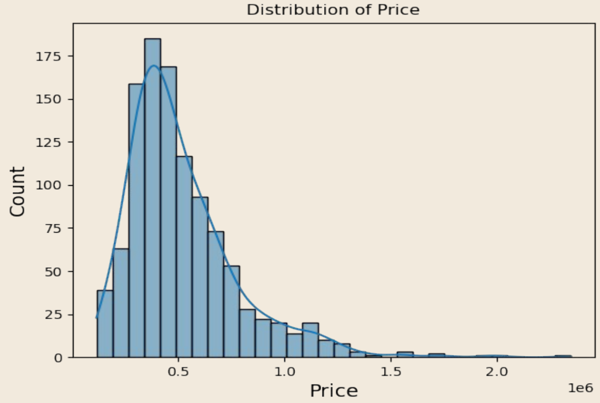
- **Number of Observations:** 1085 homes
- **Unit of Observation:** Individual Properties
- **Number of Variables:** 14 variables
- **Key Variables:**
    - Price (USD)
    - Size (sq ft)
    - Bathrooms
    - Beacon (Boolean)
- **Data Errors / Limitations:**
    - Retrospective Dataset
    - No Neighbor Data
    - No Historic Variables (year built, etc.)

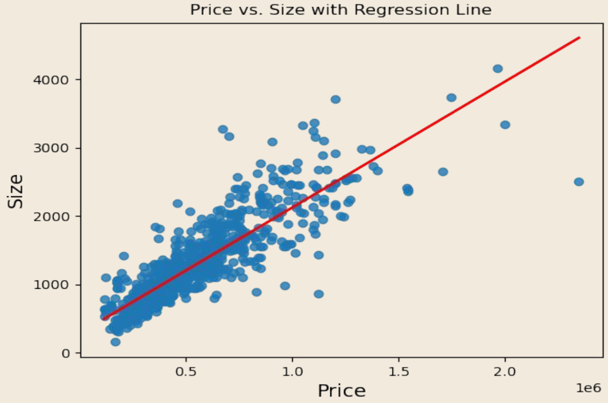


	Mean	Median
Price	\$515,000	\$454,000
Size	1234 ft^2	1105 ft^2

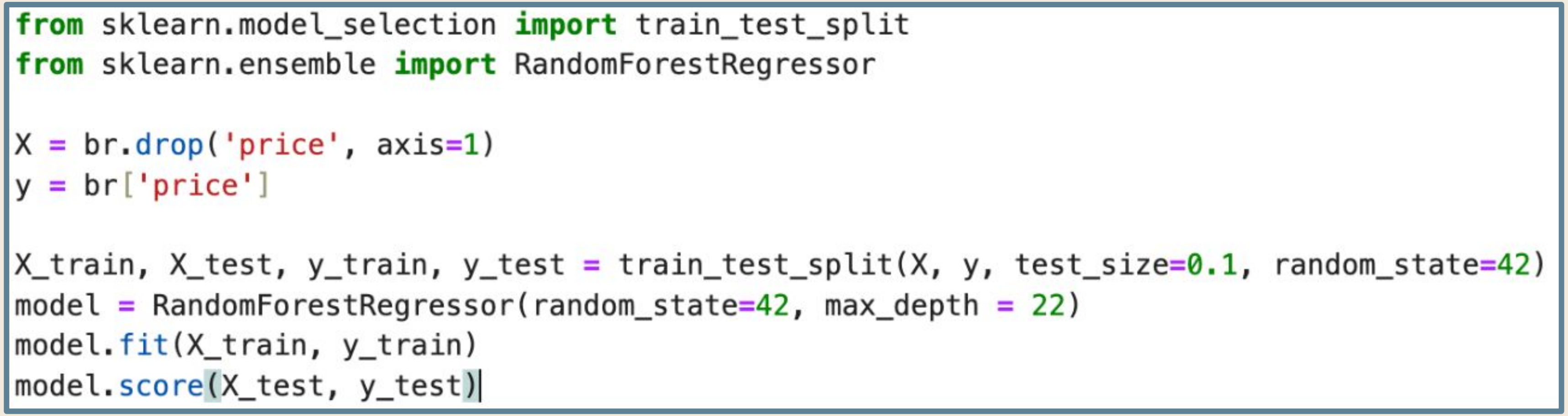
Brookline Data Set Snapshot







Machine Learning Predictions



```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor

X = br.drop('price', axis=1)
y = br['price']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=42)
model = RandomForestRegressor(random_state=42, max_depth = 22)
model.fit(X_train, y_train)
model.score(X_test, y_test)
```

Regressor Forest Score

88.3%

Top 5 Best Predictors:

1. **Size (82%)**
2. Garage (3.6%)
3. Rooms (2.4%)
4. Elevators (2.1%)
5. Mid-rise Building (2.1%)

Predicting Price Using Algorithms

Forward Selection

Start	Empty Regression
Does it improve adjusted R^2 ?	Test Each Variable
Add the variable that improved adjusted R^2 the most	Add Best Variable
Stop when no remaining variables improve adjusted R^2	STOP?
Adjusted R^2	**80%**

Backward Selection

Start	Full Model
What is the variables significance?	Test Each Variable
Remove variable with the largest insignificant p-value	Remove variable?
Stop when every remaining variable in the model is significant	STOP?
Adjusted R^2	**80%**

Why Machine Learning Can More Accurately Price a Property

Traditional Algorithms

- **Limited Variables are Used**
- **Assumes Linear Relationships**
- **Misses Complex Patterns like:**
  - Diminishing returns on size
  - Interactions between variables

Machine Learning

- **Non-linear relationships**
- **Automatically detects interactions**
- **Better at Modeling reality**
  - An additional 500 ft^2 sometimes doesn't add much values

Summary: When you need to classify something as nuanced as the price of a property, machine learning is the better bet as it captures all the nuanced relationships. But neither model is perfect. There's still a 12% error in the random forest regression model. That means that real estate agents get to keep their jobs (for now).