

Real Estate.

Business Problem

1.1 Problem Context

Our client is a large Real Estate Investment Trust (REIT).

- They invest in houses, apartments, and (complex of buildings) within a small county in New York state.
- As part of their business, they try to predict the fair transaction price of a property before it's sold.
- They do so to calibrate their internal pricing models and keep a pulse on the market.

1.2 Problem Statement

The REIT has hired us to find a data-driven approach to valuing properties.

- They currently have an untapped dataset of transaction prices for previous properties on the market.
- The data was collected in 2016.
- Our task is to build a real-estate pricing model using that dataset.
- If we can build a model to predict transaction prices with an average error of under US Dollars 70,000, then our client will be very satisfied with the our resultant model.

1.3 Business Objectives and Constraints

- Deliverable: Trained model file
- Win condition: Avg. prediction error < \$70,000
- Model Interpretability will be useful
- No latency requirement

2. Machine Learning Problem

2.1 Data Overview

For this project:

1. The dataset has 1883 observations in the county where the REIT operates.
2. Each observation is for the transaction of one property only.
3. Each transaction was between \$200,000 and \$800,000.

Target Variable

- 'tx_price' - Transaction price in USD

Features of the data:

Public records:

- 'tx_year' - Year the transaction took place
- 'property_tax' - Monthly property tax
- 'insurance' - Cost of monthly homeowner's insurance

Property characteristics:

- 'beds' - Number of bedrooms
- 'baths' - Number of bathrooms
- 'sqft' - Total floor area in squared feet
- 'lot_size' - Total outside area in squared feet
- 'year_built' - Year property was built
- 'active_life' - Number of gyms, yoga studios, and sports venues within 1 mile
- 'basement' - Does the property have a basement?
- 'exterior_walls' - The material used for constructing walls of the house
- 'roof' - The material used for constructing the roof

Location convenience scores:

- 'restaurants' - Number of restaurants within 1 mile
- 'groceries' - Number of grocery stores within 1 mile
- 'nightlife' - Number of nightlife venues within 1 mile
- 'cafes' - Number of cafes within 1 mile
- 'shopping' - Number of stores within 1 mile
- 'arts_entertainment' - Number of arts and entertainment venues within 1 mile
- 'beauty_spas' - Number of beauty and spa locations within 1 mile
- 'active_life' - Number of gyms, yoga studios, and sports venues within 1 mile

Neighborhood demographics:

- 'median_age' - Median age of the neighborhood
- 'married' - Percent of neighborhood who are married
- 'college_grad' - Percent of neighborhood who graduated college

Schools:

- 'num_schools' - Number of public schools within district
- 'median_school' - Median score of the public schools within district, on the range 1 - 10

Mapping business problem to ML problem

2.2.1 Type of Machine Learning Problem

It is a regression problem, where given the above set of features, you need to predict the transaction price of the house.

NB: This project must be submitted upon completion.