Capstone 2: Project Proposal

Category: Real Estate

Problem Statement

How can we automate the identification of profitable investment properties for both flipping and renting based on images?

Wait – so are we trying to predict what houses will sell for? Or are we just trying to identify profitable properties? *BOTH!* In order to identify profitable properties, we have to *FIRST* be able to predict what they’ll sell for. *BUT* – we need to take it a step further and have a flag for a ‘fixer-upper’. Some are labeled this way, but many aren’t. And/or if focusing on one very specific market/area, you can know if something has potential based on the asking price alone in many cases

In real life/practice, there are plenty of automation mechanisms to predict current anticipated value based on text metrics alone – recent sales of “comps” – nearby houses with similar features on paper, zip code, #bedrooms, #bathrooms, square footage, age, etc. Then a human will go in or look at the pictures to visually asses the property and tweak that number accordingly, based on condition, updatedness, yard size, special features such as a pool or furnished basement, neighbors/neighborhood, etc – all things that can’t be factored in and captured by automated text-only-based predictive models.

The goal of this is to us AI/neural networks to visually asses the property for us, based on pictures. We can train the model, telling it examples of what should boost the value and what should hurt it, perhaps through a simple scoring system / multiplier

What We’ll Need

* A dataset of real estate listing images and their associated sale prices
  + <https://www.kaggle.com/amir22010/house-price-estimation-from-image-and-text-feature/data>
* A script to pull images of sold listings

Need a way to scan pictures of listings to focus on “old” properties or ones in bad shape. This can be done through training the model with examples of these types of pictures. Then there needs to be filters for certain parameters such as min/max number of bedrooms and bathrooms. We should also work in a geofence to only target certain areas. This would be for Chicagoland

Other stuff:

Kaggle competition on predicting housing prices

Tricky to train data set with mixing pictures and text

See if there’s a readymade dataset with the pictures

Use selenium web driver to mimic a person using a website

Websites have hidden data they driven

Convolution Neural Networks – brief on that next week. For dataset with images

Deep Learning