Prediction of real estate prices in California

Lukáš Vrábel
ACAI.AI
https://acai.ai/

Abstract

The talk will present our efforts over the past year and a half to launch a start-up for US real estate valuation. The start-up called Babcock & Bonbright was built as part of the Central Europe AI portfolio – a start-up studio focused on using artificial intelligence to innovate in the financial and healthcare sectors.

The presentation will focus mainly on the technological aspect of the project – linking structured data, NLP, and image processing to a comprehensive system for estimating the value of dwellings and apartments using machine learning. Attention will also be paid to the processing of map data, system architecture and data pipeline design.

Our pricing models leverage several types of machine learning models in a hierarchical manner. For the pricing estimation, we use so-called "comparables" – recently sold houses that are close and somewhat similar (comparable) to the subject property. The main pricing model is used to compute the adjusted price for each comparable in the neighborhood using gradient boosted decision trees. The comparables are selected by location, feature similarity and image similarity. The pricing model uses features from both the subject property and comparable property, as well as combined features such as distance, image similarity, ratios of selected original features, common categorical features, common points of interest in the area, etc. The final price estimation for the subject property is determined by averaging price adjustments across all of the comparables.

Bulk of the features are created from structured and semi-structured data about the property – public sale records, offers listings, map and traffic data, school districts, and crime. The text description of the property from the agent/broker listing is also analyzed. Sentiment is extracted by a pretrained sentiment neuron model and rule-based

VADER model. Additionally, a simple bag-of-words like approach is used to extract important keywords that correlate positively/negatively with the price. Both sentiments and keyword occurrence are included in the property feature vector for the main pricing model.

Finally, a Deep Convolutional Neural Network is used for analysis of aerial images. We use transfer learning from standard ImageNet ResNet-50 model to distinguish between cheap and expensive houses. Finetuning forces the network to recognize the most important image patterns for an internal image representation. This representation together with the network classification results then serves as an additional features for the pricing model.

Keywords: Structured data, NLP, Image Processing