Ryan Filgas

CS545
5-21-22

Machine Learning Grad Paper Review

"Understanding house price appreciation using multi-source big geo-data and machine learning" by Kang et al. uses machine learning and big data to predict house price appreciation. Kang et al discuss methods of data collection, and use of features in classification algorithms in the greater Boston area using a dataset of 20,000 plus houses. Their research includes fine scale house price appreciation estimation, neighborhood-scale house price appreciation estimation, model and determinants analysis, and challenges encountered in conducted work. Additionally, the authors analyze the weight of several different property features as contributors to house appreciation in relation to spatial proximity. The end goal of the research is a proposal for a "comprehensive multi-feature-fusion framework using machine learning to model the house pricing appreciation rate" (Kang et al., 2).

The research will provide insight to buyers, researchers and decision makers in real estate as well as urban planning and development. Researchers focused on filling in perceived gaps in housing analysis, stating the usual focus of studies falls on price modeling rather than appreciation. The study also uses Redfin listing images and street view as features which allow human perspective to be better captured in addition to neighborhood gps data.

The framework proposed encompasses data collection, feature construction, model training and mapping and analysis. Using collected datasets algorithms were developed using machine learning and spatially weighted regression. The four different categories of data used in their calculations are house information, built environment, human mobility patterns, and socio-economic attributes. House prices are gathered across a five-year period from 2014 to 2019 as well as the appreciation rate in that time period.

In gathering feature data the researchers pulled from several sources. For visual features, street view images are used sampling every 100 meters of each neighborhood and kept in the data set if they're within 50m of a house. Image data features are extracted using a deep convolutional neural network. Human mobility patters are captured using the SafeGraph anonymized mobile phone location database at the resolution of census block groups as well as travel time between locations as provided by the Uber Movement Project which captures density of amenities in the area as well. This data is used to calculate travel time for features like grocery stores, universities, entertainment etc. Lastly, socioeconomic data is used from the American community survey which consists of ethnicity and its ratios of population, average income, and average unemployment.

The model researchers trained for image classification is formulated as a 10 category classification task to accelerate training and avoid errors from skewed distribution. Due to the time cost of training the researchers train on image data separately and adopt PCA into their model to reduce feature dimension. Due to the specificity of their component reduction this dataset wouldn't be well generalized to areas outside of Boston, but can serve as a model for future research. This appears to be an intelligent decision given the quantity of data to process.

Kang et al. compare two approaches to test the framework at a fine grain level, the first being multiple linear regression and the other is a machine learning approach called gradient boosting machine with decision trees. At a neighborhood level ordinary least squares regression and geographically weighted regression are used to estimate the variables that influence house appreciation rate. GWR allows the model to consider spatial relationships to nearby amenities. This is important as the variables influencing housing prices show patterns

primarily in their local proximity. Researchers hope that including spatial relationships will increase accuracy and provide new insights. Root mean square error and the coefficient of determination R^2 are used to evaluate model performance for results.

Findings showed that the logarithm house price was the most important feature (a correlation coefficient of .55) in determining house appreciation showing lower cost houses had the highest appreciation. House type and location were also influential factors in predicting appreciation (absolute correlation coefficient of.49). Researchers noted they would like to include more regional factors in their studies including built environment, country, culture, coastal vs inland etc. to generalize the model. They expressed uncertainty in the quality of collected image data due to collection method; they detected that the models formed from street view photos don't fare as well as others.

The research proposed several insights that seem useful, including spatial and image data. It was revealing how image data was used to reduce error, and how PCA was adopted to evaluate data efficiently concerning integration of the image data with the rest of the model. Incorporating image, spatial, and GPS movements into housing prediction may become a standard practice if it's not already, although the mixed results from the image data might require refinement. The researchers found the machine learning model outperformed their statistical model (MLR) as they expected. I would ask the authors if it would be worth training a separate model to recognize houses more precisely in order to improve the accuracy of their prediction models on gathered image data. I would also be curious to see the data for proximity to amenities graphed on a map as a predictor for house appreciation.

Sources Cited

Kang, Yuhao, et al. “Understanding House Price Appreciation Using Multi-Source Big Geo-Data and Machine Learning.” *Land Use Policy*, vol. 111, 24 Nov. 2021, pp. 1–11., https://doi.org/10.1016/j.landusepol.2020.104919.