Spatial-Temporal Prediction of Housing Price in California during the COVID Period with Machine Learning Methods

Zixi Chen
zchen807@gatech.edu

Xinyue Huang
xinyue@gatech.edu

Yiwei Sun
ysun682@gatech.edu

Jingjing Ye
jingjing@gatech.edu

Zhan Zhang
zzhang601@gatech.edu

1 INTRODUCTION

At the end of 2019, COVID-19 broke out and quickly grew to a global pandemic. It had tremendous impact on the world, hitting every aspect of people's lives, from global supply chain to people working from home whenever possible. In the meantime, we witnessed housing price soared in the past couple of years in California and the United States as a whole. Housing price prediction has been a classic application in machine learning [4], and recent studies have incorporated spatial correlation into traditional time series models [6] [1]. However, there is no study that involves COVID-related factors into housing price prediction. In this study, we would like to investigate if COVID fundamentally changed how houses are priced and do spatial-temporal prediction of housing price during the COVID period with different machine learning methods.

2 PROBLEM DEFINITION

In this project, we plan to explore the relationship between housing prices in California and various factors during the COVID-19 period. We have collected datasets from Zillow that gives the level of housing prices in different counties in California over the years. We also believe the housing price is related to various economic factors such as household income or unemployment rates, demographic factors such as ethnicity or age, and more importantly, COVID related factors such as infection and vaccination cases during the past 2 years. We will primarily focus on the housing price changes during COVID period and make comparison to other times if necessary.

3 METHOD

3.1 Data and Feature Engineering

We will start the data cleaning by removing any duplicate and NaN of the data, then convert any decoded data or numeric dates to decoded data and string for better performance. Then we may do some exploratory data analysis (EDA) to explore the quality of the data. Next, we will do the feature engineering including outliers handling, binning, and scaling to prepare the proper input dataset and improve the performance of the machine learning models.

3.2 Unsupervised Learning

For unsupervised method, well-known algorithms include dimensionality reduction and clustering analysis. We choose Principal component analysis to predict the price [3].

3.3 Supervised Learning

3.3.1 Support Vector Regression. SVR uses the same idea of SVM but here it tries to predict the real values. This algorithm uses hyperplanes to segregate the data. In case this separation is not possible then it uses kernel trick where the dimension is increased and then the data points become separable by a hyperplane.

3.3.2 Lasso Regression. LASSO stands for Least Absolute Selection Shrinkage Operator. Shrinkage is basically defined as a constraint on attributes or parameters.

3.3.3 Random Forest. Random Forests are an ensemble(combination) of decision trees. It is a Supervised Learning algorithm used for classification and regression. The input data is passed through multiple decision trees.

In this project, we can also try to combine supervised and unsupervised method. [5] In baso's study, various supervised algorithms (random forest, gradient boosting, support vector machine, etc.) were combined with unsupervised algorithms such as k-means to improve the performance of credit scoring models.

3.4 Spatio-Temporal Kernel Density Estimation

Spatio-Temporal Kernel Density Estimation [2] is a temporal extension of the traditional kernel density estimation KDE used for identifying spatio temporal patterns.The density estimates are visualized within the space-time cube framework using two-dimension spatial (x,y) and a temporal dimension (t). After running STKDE,it will provide a 3D raster volume as output where each voxel is assigned a density estimate based on the surrounding point data. The space-time density is estimated by the following equation

4 POTENTIAL RESULTS AND DISCUSSION

We expect to build up a spatial-temporal predictor of housing price in California during the COVID period and output the following potential results: 1. Comparison of the performance of different machine learning methods and determine the overall best model for our task. 2. Accuracy of the predicted housing price with the testing data overtime. 3. Correlation between the COVID situation (infection and vaccination cases) and local housing price overtime. We hope to take the advantage of the detailed and comprehensive spatial-temporal data and optimize the model to predict the housing price with a high accuracy. In addition, as this project mainly focuses on the influence of COVID on housing price, we will evaluate and analysis all the COVID-related factors and draw some conclusions on them.

Every member has **20 % contribution.**

REFERENCES

[1] Henry Crosby, Paul Davis, Theo Damoulas, and Stephen A. Jarvis. 2016. A Spatio-Temporal, Gaussian Process Regression, Real-Estate Price Predictor. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPACIAL '16)*. Association for Computing Machinery, New York, NY, USA, Article 68, 4 pages. https://doi.org/10.1145/2996913.2996960

[2] Yujie Hu, Fahui Wang, Cecile Guin, and Haojie Zhu. 2018. A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation. *Applied Geography* 99 (2018), 89–97. https://doi.org/10.1016/j.apgeog.2018.08.001

[3] Changro Lee. 2021. PREDICTING LAND PRICES AND MEASURING UNCERTAINTY BY COMBINING SUPERVISED AND UNSUPERVISED LEARNING. *International journal of strategic property management* 25, 2 (2021), 169–178.

[4] Lianfa Li. 2019. Geographically Weighted Machine Learning and Downscaling for High-Resolution Spatiotemporal Estimations of Wind Speed. *Remote Sensing* 11, 11 (2019). https://doi.org/10.3390/rs11111378

[5] Bao Wang, Yue Kong, Yongtao Zhang, Dapeng Liu, and Lianju Ning. 2019. Integration of Unsupervised and Supervised Machine Learning Algorithms for Credit Risk Assessment. *Expert Systems with Applications* 128 (03 2019). https://doi.org/10.1016/j.eswa.2019.02.033

[6] Linlin Zhu and Hui Zhang. 2021. Analysis of the diffusion effect of urban housing prices in China based on the spatial-temporal model. *Cities* 109 (2021), 103015. https://doi.org/10.1016/j.cities.2020.103015