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 [7] [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 DATA COLLECTION

3.1 COVID Data

The COVID-related data was obtained from the Official California State Government Website. The data was collected on a daily basis per county. We first calculated the cases, deaths, and case and death rate per month from Jan. 2020 to Dec. 2021 of every county in California. Thee full vaccination number and rate were calculated similarly. Note-worthily, the vaccination data started from Jul. 2020 since there was no vaccine in the beginning of the pandemic. To solve this problem, we impute those data with 0 since there is no vaccination, and it is reasonable when considering the feature in a specific model.

3.2 Unemployment Data

The unemployment data was obtained from Data Commons. The monthly unemployment rate of every county in CA was downloaded and merged.

3.3 Housing Inventory

The housing inventory data was obtained from realtor.com Real Estate Data Library. The data of counties in CA was selected from the whole data set, which contains data from all counties in the United State.

3.4 Housing Price

Ideally we want to get housing price directly from the market, but actual transaction data are hard to obtain. As a result, we leveraged housing prices from Zillow Home Value Index(ZHVI)[3]. In short, it reflects the typical values for homes for a region, in the unit of dollar price. The dataset comes in different levels of spatial-granularity from Nation to zip code. We used county

level data to balance between enough data points vs the availability of independent variables. We limited the region to California, given it is one of the larger state with more population and more houses. In the mean time, it has a relatively hot housing market which implies there are more transaction data and more accurate price estimation. From a temporal perspective, we used monthly data for data availability purposes. We could only obtain monthly data as ZHVI is only published on a monthly basis. Initially, we used data from the beginning of 2018 to the end of 2021 and could potentially trace back to more ancient data. Finally, we did not distinguish home types like single family or condo in the current phase of the project.

For the housing price itself, it is the average of Zillow's signature Zestimate home value within 35th to 65th percentile range. The Zestimate home price, according to Zillow, is calculated using their neural network-based model taking into account many home characteristics like square footage, location or number of bathrooms. Zillow also provided additional statistics on their website about Zestimate and we believe this is a fairly accurate representation of housing prices.

Variables	Data
County	Napa County
Date	2020-11
Housing Price	744631
Housing Inventory	311
Unemployment Rate	6.4%
COVID-19 Cases	1433
COVID-19 Deaths	3
Cases Rate	1.0261%
Death Rate	0.0021%
Fully Vaccinated	3
Fully Vaccinated	0.0021%
Population	136207
Area	754
GDP	9870652
GDPpp	72.468023

Figure 1: Example of Data

3.5 Data Clean

After data from all sources was collected, we cleaned them separately and standardized the form of the data before merging all features together. Next, we tried to deal with the missing values and outliers. When we took a look at the number of missing variables within each attribute, we found that there are only 1.24%. There are also contextual outliers in the data. When we checked the vaccination data, we found that there are some data points in Jan. 2020, which is unreasonable since it is the beginning of the pandemic and the vaccines are not available until Jul. 2020. Thus, we decided to delete these data since they are wrong.

After preprocessing the data, an example of the current data set looks like the table below. The housing price is the dependent variable, while others are independent variables. COVID-19 related data starts from Jan. 2020, helping us to discover the influence of the COVID-19 on housing prices.

There still might be some work for data preprocessing in the next stage of the project. First, the features before COVID-19 period are insufficient. We may add more reliable features in the next step. Moreover, there can be other forms of the existing data that can perform better in the model. For example, the cumulative sum of vaccination might be more useful. This needs further investigation in the next steps

4 DATA PRE-PROCESSING

In order to get a better understanding of our data, we conducted exploratory data analysis. Based on the critical moments of the development of COVID-19, we divided the data into 3 parts. The first part contains data before the pandemic, the second part contains data for the early stage of the pandemic, and the third part contains data for the later stage of the pandemic.

4.1 Data Distribution

After dividing the data into three parts, we plotted the distribution of data in these parts as well as in the complete data set to examine the quality of our data. First, we plotted the distribution of price and the housing inventory.

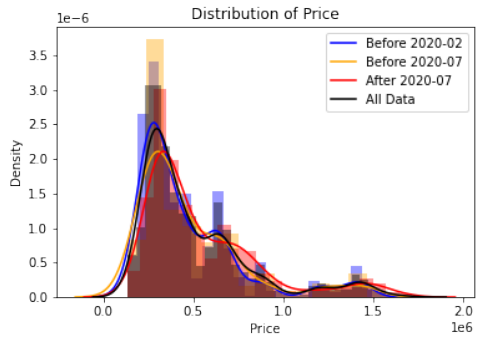

Figure 2: Distribution of Price

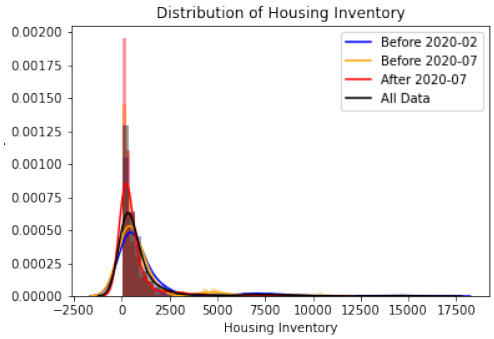

Figure 3: Distribution of Housing Inventory

We can observe from the graph of price distribution that there are more low housing prices before the outbreak of COVID-19, and more high housing prices after the outbreak of COVID-19, which means that in general, the housing prices have experienced an increase due to the pandemic. Similarly, we can observe from the graph that the housing inventory decreased after COVID-19.

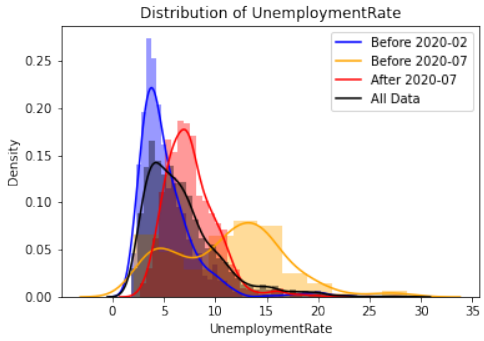

Figure 4: Distribution of Unemployment Rate

The distribution of the unemployment rate gives a clearer view of the influence of the pandemic. The unemployment increased significantly after the outbreak of the pandemic, then recovered partly after vaccination came into use. The distribution plots confirmed the fact that the housing price as well as features influencing the price were significantly impacted by the pandemic, so building a model that can predict the housing price after COVID-19 is necessary.

4.2 Data Trend

Next, we plotted the change of variable price over time, with each line in the graph representing the housing price in one county. To better observe the change, we only showed the trend of top 6 counties with the highest housing prices.

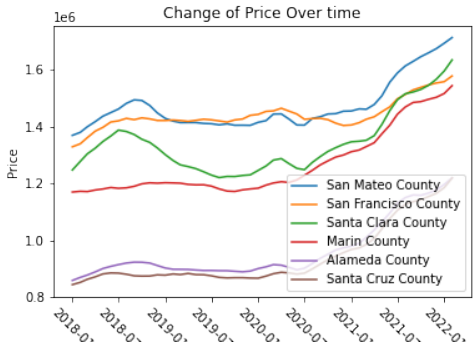

Figure 5: Trend of Price

There is a significant increase in housing prices, which can be attributed to the pandemic. The impact can also be observed from the line graph of the housing inventory and the unemployment rate.

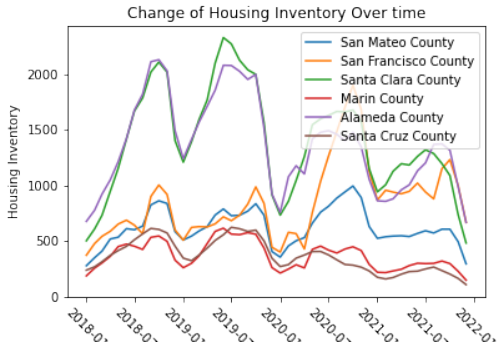

Figure 6: Trend of Housing Inventory

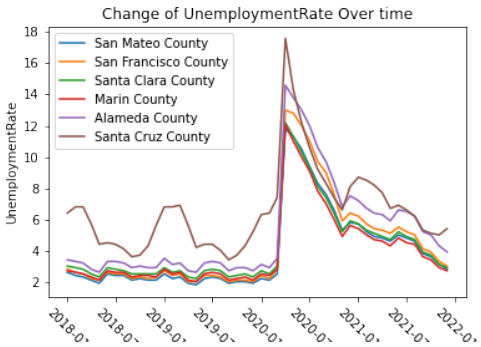

Figure 7: Trend of Unemployment Rate

We also examined the trend of COVID-19 cases and deaths numbers. To eliminate the impact of the difference in population between counties, we plotted the rate of them.

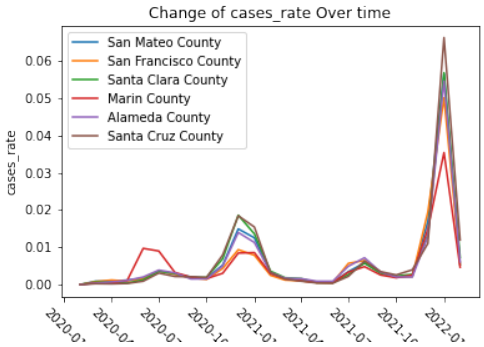

Figure 8: Trend of Case Rate

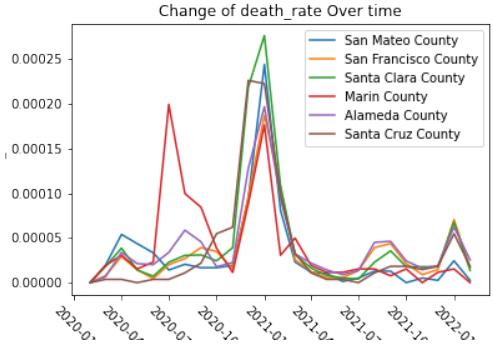

Figure 9: Trend of Death Rate

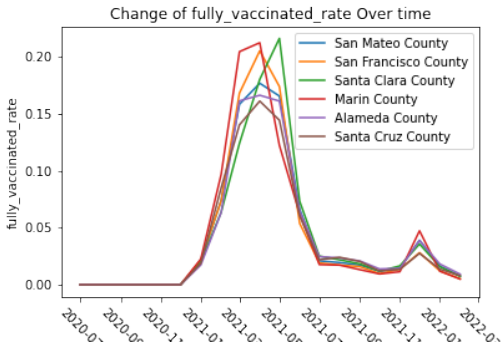

Figure 10: Trend of Fully Vaccination Rate

There is a difference between the cases and deaths plots. The spread of omicron affected the number of cases significantly but had a limited impact on the number of deaths. The change in vaccination and vaccination rate also has two peaks, and is more similar to the deaths and deaths rate plots, as shown below.

4.3 Feature Correlation

Then we drew the heatmap to examine the correlation between variables. As shown below, there are some correlations between independent variables. Therefore conducting the principal component analysis or variable selection algorithm might improve the performance of later models.

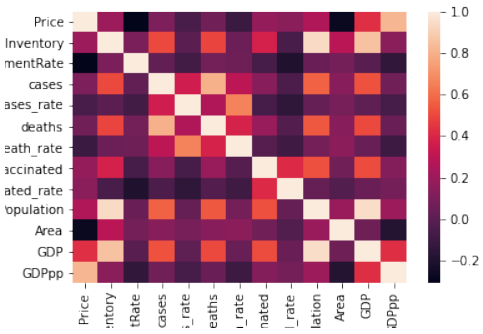

Figure 11: Feature Correlation Heatmap

4.4 Dimensionality Reduction

The features of a dataset are often referred to as dimensions.High dimensions do not mean better train result. We need to do dimensionality reduction to move the features not so important.

There are some ways like Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA), LASSO and RIDGE Regression. LASSO was chosen, because only LASSO does true dimensionality reduction since it forces many of the beta coefficients to be 0 while RIDGE and Elastic Net force small coefficients to be near to 0.

4.4.1 Lasso. [5] (Least absolute shrinkage and selection operator):
Suppose that we have data $\left(\mathrm{x}^i, y_i\right), i = 1, 2, \ldots, N$, where $\mathrm{x}' = \left(x_{i1}, \ldots, x_{ip}\right)^{\mathrm{T}}$ are the predictor variables and y_i are the responses. As in the usual regression set-up, we assume either that the observations are independent or that the y_i s are conditionally independent given the x_i s. We assume that the x_{ij} are standardized so that $\Sigma_i x_{ij}/N = 0, \Sigma_i x_{ij}^2/N = 1$. Letting $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)^{\mathrm{T}}$, the lasso estimate $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ is defined by

$$(\hat{\alpha}, \hat{\beta}) = \arg\min \left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\}$$

$$\text{subject to } \sum_j \left|\beta_j\right| \leqslant t.$$

Here $t \geqslant 0$ is a tuning parameter. Now, for all t, the solution for α is $\hat{\alpha} = \bar{y}$. We can assume without loss of generality that $\bar{y} = 0$ and hence omit α.
Write it in 1 Norm form:

$$Q(\beta) = \|y - \boldsymbol{X}\beta\|^2 + \lambda\|\beta\|_1$$

$$\Longleftrightarrow \arg\min \|y - \boldsymbol{X}\beta\|^2 \quad s.t. \quad \sum \left|\beta_j\right| \leq s$$

4.4.2 LARS Algorithm.
Here use LARS[2] algorithm to compute. As with classic Forward Selection, we start with all coefficients equal to zero, and find the predictor most correlated with the response, say x_{j1} . We take the largest step possible in the direction of this predictor until some other predictor, say x_{j2} , has as much correlation with the current residual. At this point LARS parts company with Forward Selection. Instead of continuing along xj1 , LARS proceeds in a direction equiangular between the two predictors until a third variable x_{j3} earns its way into the "most correlated" set. LARS then proceeds equiangularly between x_{j1} , x_{j2} and xj3 , that is, along the "least angle direction," until a fourth variable enters, and so on.

$\overline{\mathrm{y}}_2$ is the projection of $\mathbf{y}$ into $\mathcal{L}\left(\mathbf{x}_1, \mathbf{x}_2\right)$. Beginning at $\widehat{\mu}_0 = 0$, the residual vector $\tilde{\mathrm{y}}_2 - \widehat{\mu}_0$ has greater correlation with x_1 than x_2; the next LARS estimate is $\widehat{\mu}_1 = \widehat{\mu}_0 + \widehat{\gamma}_1 \mathbf{x}_1$, where $\widehat{\gamma}_1$ is chosen such that $\overline{\mathrm{y}}_2 - \widehat{\mu}_1$ bisects the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$; then $\widehat{\mu}_2 = \widehat{\mu}_1 + \widehat{\gamma}_2 \mathbf{u}_2$, where $\mathbf{u}_2$ is the unit bisector; $\widehat{\mu}_2 = \overline{\mathrm{y}}_2$ in the case $m = 2$, but not for the case $m > 2$; see Figure 4. The staircase indicates a typical Stagewise path. Here LARS gives the Stagewise track as $\varepsilon \to 0$, but a modification is necessary to guarantee agreement in higher dimensions;

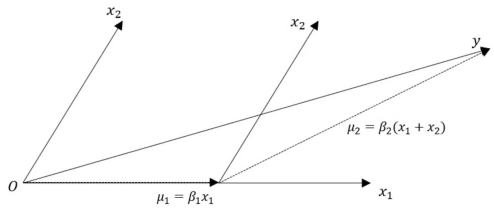

Figure 12: The LARS algorithm in the case of m = 2 covariates;

4.4.3 Result.

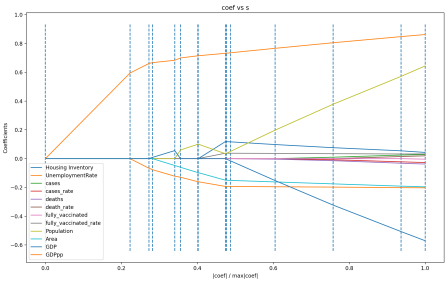

Figure 13: coefficient vs s

To explain this result. Here, the smaller the s is selected, the more the parameters are compressed to 0. If the s is very large, the coefficients will not be compressed, basically.The abscissa s is the s in $\sum_j \left|\beta_j\right| \leqslant s$ From the picture we can see, per capita GDP is a very important feature, the coefficient even can not decrease to zero for any situation. If we set s= 0.6 we can see

the features we should keep are GDPpp, Population, Housing Inventory, GDP, fully vaccinated rate, Area, Unemployment Rate. The coefficients of the remaining variables are compressed to 0.

5 METHOD

5.1 Train-test Split

The train-test split process is essential for estimating the performance of our machine learning models that are used to make predictions on data not used to train the model. We split the data into the train set and the test set: the train dataset is used to fit the machine learning model, while the test dataset is used to evaluate the fit machine learning model. We set the size of the test set to be 0.2, which means 20% of the dataset is split as the test set.

5.2 Supervised Learning

In this part, we implemented and trained five different regression models: linear regression model, Ridge regression model, Lasso regression model, Bayesian regression model, and random forest regression model. We used the pre-built algorithms provided by the scikit-learn (sklearn) package in Python for implementation of all the models. We followed the same process to implement, train, and test all the models: first, we begun with instantiating objects of the class available in the sklearn module; next, we fitted the variables in the train set to the models; and once the model was fit, we evaluated the models' performance by predicting the housing price with the test set and apply different machine learning performance metrics including the explained variance score, R-squared score, and mean squared error to the results.

5.2.1 Linear Regression. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. As it is specified by sklearn, linear regression fits a linear model with coefficients $w = (w_1, \ldots, w_p)$ to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. It is one of the most basic machine learning models we have learned.

5.2.2 Ridge Regression. Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated. The model minimizes the objective function: $||y - X_w||_2^2 + alpha \cdot ||w||_2^2$. It model solves a regression model where the loss function is the linear least squares function and regularization is given by the l2-norm.

5.2.3 Lasso Regression. Lasso (least absolute shrinkage and selection operator) regression is an analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model. It is a linear model that trained with L1 prior as regularizer. The optimization objective for Lasso is: $(1/(2 \cdot n_samples)) \cdot ||y - X_w||_2^2 + alpha \cdot ||w||_1$.

5.2.4 Bayesian Ridge Regression. Bayesian linear regression is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference. This technique can be used to include regularization parameters in the estimation procedure: the regularization parameter is not set in a hard sense but tuned to the data at hand.

5.2.5 Random Forest Regression. Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

5.3 Unsupervised Learning

In the future of this project, we will also try to implement the unsupervised and combination of supervised and unsupervised method. [6] 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.

6 RESULTS AND DISCUSSION

We evaluated the models' performance by predicting the housing price with the test set and apply different machine learning performance metrics including the explained variance score (EVS), R-squared score (R^2), and mean squared error (MSE) to the results. The model evaluation results are shown in the table below:

	Linear	Ridge	Lasso	Bayesian	RF
EVS	0.7675	0.7676	0.7675	0.4729	0.9716
R^2	0.7672	0.7673	0.7672	0.4684	0.9706
MSE	2.8e11	2.8e11	2.8e11	6.5e11	3.6e10

The results indicate that the random forest regression model has the overall best performance, while Bayesian Ridge regression model has the worst performance. The linear, Ridge and Lasso has nearly the same performance. To explain why random forest performs well, we need to understand the difference between it and linear regression. Random forest is an ensemble of random decision trees, while the decision trees can be thought of as a bunch of if-else conditions, and the values pass from the very top with one node to the end of the leaf node. Decision Trees are great for obtaining non-linear relationships between input features and the target variable, therefore, a random forest's nonlinear nature can improve its performance for smaller dataset, comparing to the linear regression. However, linear regression has much fewer parameters than random forests, which means that random forests will overfit more easily than a linear regression.

There are several future directions of this project to be discussed: 1. When we implemented and tested the model, we found the amount of our features were not enough for some models. Therefore, we will include more features to the dataset. 2. Up to now, we did not consider the spatio or temporal features in our dataset. Therefore, for the next step, we will investigate and implement the Spatio-Temporal Kernel Density Estimation (STKDE), which is a temporal extension of the traditional kernel density estimation KDE used for identifying spatio temporal patterns.

Overall, 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 over time. We hope to take the advantage of the detailed and comprehensive spatial-temporal data and optimize the model to predict the housing pricewith 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] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. Least angle regression. *The Annals of Statistics* 32, 2 (2004), 407 – 499. https://doi.org/10.1214/009053604000000067

[3] Hryniw. 2019. Zillow Home Value Index Methodology, 2019 Revision: Getting Under the Hood. (2019). https://www.zillow.com/research/zhvi-methodology-2019-deep-26226/

[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] Robert Tibshirani. 1996. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1996.tb02080.x

[6] 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

[7] 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