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 [5], 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 Data Source

Features used in this project were collected from multiple sources, and were cleaned and organized into a similar format before merging.

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.

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.

Other features were collected multiple sources. For example, the unemployment data was obtained from Data Commons. The housing inventory data was obtained from realtor.com Real Estate Data Library. The GDP, revenue, and expenditure data was obtained from California Open Data Portal. The crime data was obtained from California Department of Justice. And the education and hospital data were downloaded from California Department of Education.

3.2 Data Cleaning

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

Variables	Data
County	Napa County
Date	2020-11
Housing Price	744631
Housing Inventory	311
Unemployment Rate	6.4%
COVID_Cases	1433
COVID_Deaths	3
Cases_Rate	1.0261%
Death_Rate	0.0021%
Fully_Vaccinated	3
Fully_Vaccinated_Rate	0.0021%
Population	136207
Population_Density	42,725
Area	754
GDP	9870652
GDPpp	72.468023
Violent_Crimes	10585
Violent_Crimes_pp	3.3
Property_Crime	59,050
Property_Crime_pp	18.6
Revenue_pp	22016
Expenditures_pp	21083
Hospital	37
Hospital_pp	0.0000113
School	1058
School_pp	0.00032197
Public_School	789
Public_School_pp	0.00024011
Private_School	269
Private_School_pp	0.0000819

Figure 1: Example of Data

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 and to haves a better understanding of the influence of COVID on our society. Some insightful results were listed below. First, we plotted the distribution of price and the housing inventory.

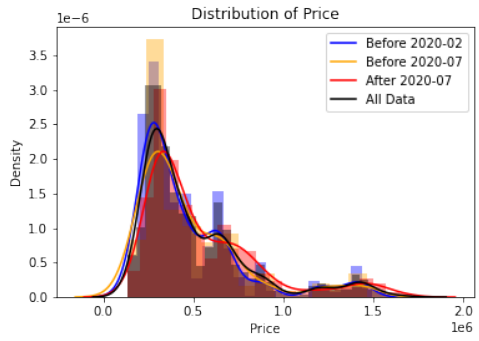

Figure 2: Distribution of Price

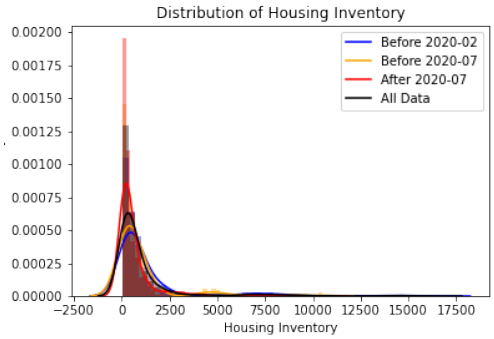

Figure 3: Distribution of Housing Inventory

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.

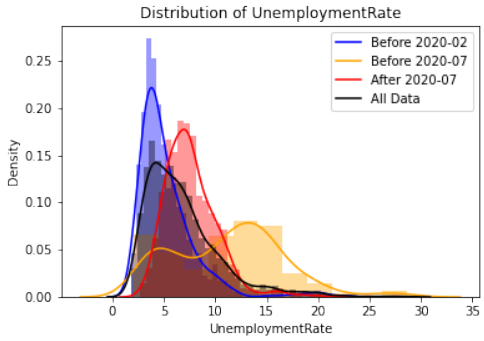

Figure 4: Distribution of Unemployment Rate

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.

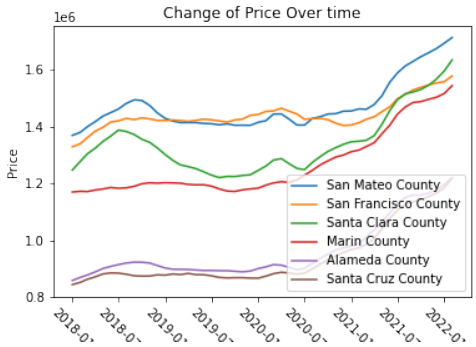

Figure 5: Trend of Price

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.

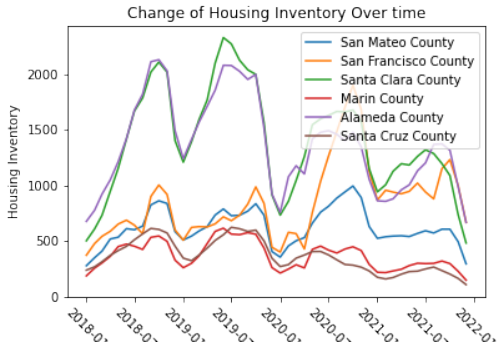

Figure 6: Trend of Housing Inventory

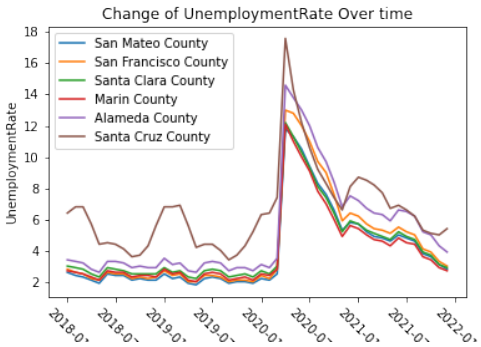

Figure 7: Trend of Unemployment Rate

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.

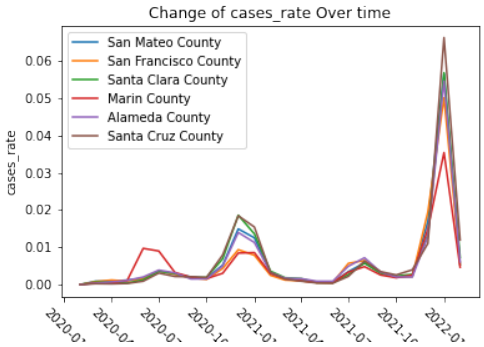

Figure 8: Trend of Case Rate

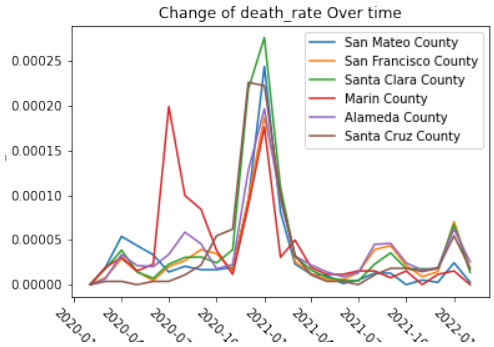

Figure 9: Trend of Death Rate

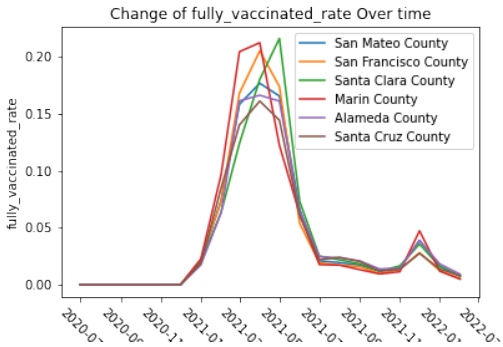

Figure 10: Trend of Fully Vaccination Rate

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.

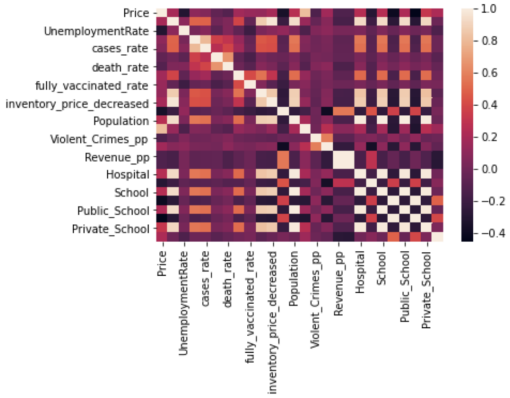

Figure 11: Feature Correlation Heatmap

4.4 Feature Selection

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. [6] (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;

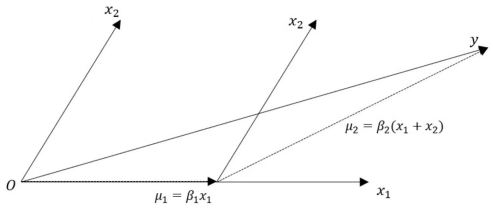

Figure 12: The LARS algorithm in the case of m = 2 covariates;

4.4.3 Result.

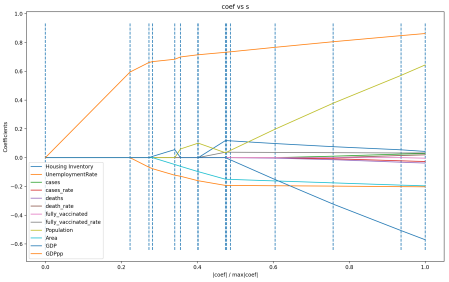

Figure 13: coefficient vs s

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 Dimension Reduction

5.1.1 Principle Component Analysis.

For unsupervised method, well-known algorithms include dimensionality reduction and clustering analysis. We choose Principal component analysis to predict the price [4]. In this part, we use the PCA function in sklearn and try to find which variable gives the most variances. We could adjust the number of components we want to keep, as well as the features we want to test.

We want to do additional data processing before we start PCA, the steps include:

- Change data type for two columns, we have numerical data with type string.
- Split the data by month before running PCA. Given our dataset has spatial and temporal dimensions, we want to split the dataset by temporal dimension and run PCA on subsets of it.
- Normalize the data, given our independent variables have different units and scale, we need to normalize them so their variances are on the same scale.
- Convert dataframe to numpy array for PCA function

In our dataset, we have a suite of COVID related variables. During our investigation, we found that these variables have huge variances that can skew the results. Therefore, we perform PCA on our dataset with/without these variables. Additionally, per-capita data and total data could be more or less correlated and introduces colinearity in our dataset. Therefore, we also explored per-capita only data PCA.

5.2 Time Based Analysis

5.2.1 Lag Time Analysis.

In this part we examined the relationship between the housing price and covid-related features. By comparing two time series, we can obtain useful insights regarding similarities of their changes over time.

To eliminate the influence of unwanted variables, in this section we consider the counties separately and examine the housing price and values of covid-related features within that county. The Tuolumne County was excluded from the analysis since it contains a large proportion of missing housing price and covid-related data, and all other counties were included.

Because the change of covid-related features and the change of the housing price are probably not simultaneously, we introduced the lag time analysis. We set the lag time to different values to examine which value describes the correlation the best.

The correlation coefficients were computed using the *personr* function in the *stats* module of the *scipy* library. The function will return us both the absolute value of the Pearson correlation between the two variables and the two-tailed p-value of the correlation.

5.2.2 Autoregressive Model.

Now, we used the Time-Series model to investigate whether COVID-19 has an effect on housing prices and try to predict the housing price more accurately in a specific area. Thus, we introduced AR models. The AR(Autoregressive) model predicts future behavior based on past behavior. It is used for forecasting when there is some correlation between the value at a certain time and the values in the past. Housing prices, as former studies indicate, can be well-predicted using the AR model. The basic formula of the AR model is shown below.

$$y_t = \delta + \sum_{i=1}^{p} \Phi_i y_{t-i} + \epsilon_t$$

We will use house pricing data from 2015 to 2020 of Los Angeles County(which is the most populated county in California) to train the AR model. Then we will use the model to predict the housing price from 2020 to 2021. If the COVID-19 influenced the housing price, the prediction can be away from the real prices.

5.2.3 ARX model.

ARX (Autoregressive with Exogenous Variables) model includes exogenous variables as input terms. It not only focuses on the correlation of now and past dependent variables but also introduces the exogenous variables to help predict. This model can allow us to use more time-based data such as COVID-19-related data and see its influence on the housing price.The basic formula of

the ARX model is shown below.

$$y_t = \delta + \sum_{i=1}^{p} \Phi_i y_{t-i} + \sum_{i=0}^{s} \Theta_i^* x_{t-i} + \epsilon_t$$

We will add time-based data such as housing inventory, unemployment rate etc. and COVID-19 related data to the former AR model. We expect the performance of the model would be better if we introduced COVID-19 related features.

5.3 Regression Analysis

In this part, we implemented and trained five different regression models: Linear regression model, Ridge regression model, Lasso regression model, Decision Tree regression model, Random Forest regression model, and K-Nearest Neighbors (KNN) 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 did the data preprocessing including train-test split and normalization. Then 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.

5.3.1 *Data Preprocessing.*

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.Then we normalized all the features with functions implemented in the sklearn.

5.3.2 *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. The objective function is $||y - X_w||_2^2$, and it is one of the most basic machine learning models we have learned.

5.3.3 *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.

5.3.4 *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: $(1/(2 \cdot n_samples)) \cdot ||y - X_w||_2^2 + alpha \cdot ||w||_1$.

5.3.5 *Decision Tree Regression.*

Decision Tree is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs, and utility. A regression tree is basically a decision tree that is used for the task of regression which can be used to predict continuous valued outputs instead of discrete outputs. The model observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. We used mean squared error as criterion and set the max depth to be 6.

5.3.6 *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. We used mean squared error as criterion and set the max depth to be 6, estimator number to be 100, and random state to be 3.

5.3.7 K-Nearest Neighbors Regression.

The KNN algorithm is a non-parametric supervised learning method that can be used for both classification and regression. In both cases, the input consists of the k closest training examples in a data set. The output depends is the property value for the object. This value is the average of the values of k nearest neighbors. Basically, The target is predicted by local interpolation of the targets associated of the nearest neighbors in the training set. While implementing the model in sklearn, we set the number of neighbors to be 3.

6 RESULTS AND DISCUSSION

6.1 Dimension Reduction

6.1.1 PCA with all features.

In this part, we run all available features for PCA every 6 months and presented our results in figure 14. We found that death rate and number of vaccination is the predominant component in our dataset. For 2020-02, we believe this is sensible because some of the counties started to accumulate death cases while others are not. Therefore, we see a lot more variances in the variable death rate than in other variables. As time passes, fully vaccinated became the predominant factor, as we believe there's a large variation in terms of vaccination across different states.

	2020.2	2020.8	2021.2	2021.8
PC1	Death Rate	Fully Vaccinated	Fully Vaccinated	Fully Vaccinated
PC2	Violent Crimes	Deaths	Deaths	Population Density
PC3	Population Density	Population Density	Fully Vaccinated Rate	Hospital per person

Figure 14: PCA with all features

6.1.2 PCA without COVID Features.

In previous section, we saw that COVID related factors accounted for most of the variances, but it is not convincing that the number of death case is the primary driver of housing prices. Therefore, we explore independent variables excluding COVID factors and presented our results in figure 15.

Running PCA on the same time point we found that private school, population and hospital per person are the primary driving factors. This is inline with intuition that people tend to find housing closer to good school and more hospital. In the mean time, more people usually means the housing prices are higher. For example, NYC and SF's housing are in general more expensive.

Across different time points, we found the 3 factor accounts for roughly 70, 18 and 3 percent. This number is consistent throughout the past 2 years.

	2020.2	2020.8	2021.2	2021.8
PC1	Private School	Private School	Private School	Private School
PC2	Population Density	Population Density	Population Density	Population Density
PC3	Hospital per person	Hospital per person	Hospital per person	Hospital per person

Figure 15: PCA without COVID Features

6.1.3 Per Capita PCA with all features.

There is a chance there's a large correlation between a variable and the per-capita version of it. In this section, we only keep the per-capita version of the variables and try to see which variable will become the predominant variable.

We have to run more dates because there's not one factor that dominates all the time points. Nonetheless, we found that death rate and and Hospital per person are the major components. However, the principal components are not as convincing as before, given the top explained variance ratio are only around 40 to 60 percent except for the 2020-02. We selectively presented our data in figure 16

	2020.2	2020.8	2021.2	2021.8
PC1	Death Rate	Hospital per person	Fully Vaccinated Rate	Hospital per person
PC2	Hospital per person	Death Rate	Hospital per person	Death Rate
PC3	Cases Rate	Expenditures per person	Revenue per person	Deaths

Figure 16: Per Capita PCA with all features

6.1.4 Per Capita PCA without COVID features.

Similarly, we want to explore which factor influences the housing price most if not for COVID. Therefore, we remove covid variables here.

We ran 8 different months, and we found that hospital per person, expenditures per person and public school per person are the major factors. The explained variance ratio are around 47, 25 and 13 percent. We selectively presented our results in figure 17

We understand that we are missing more granular(monthly) data on variables like number of schools and or monthly expenditure. Therefore, the result across different months looks similar to each other. However, this is still inline with our understanding of the housing market. As a potential followup for future, we can include more granular data on these fields to explore the influences of these factors across time.

	2020.2	2020.8	2021.2	2021.8
PC1	Hospital per person	Hospital per person	Hospital per person	Hospital per person
PC2	Expenditures per person	Expenditures per person	Expenditures per person	Expenditures per person
PC3	Public School per person	Public School per person	Public School per person	Public School per person

Figure 17: Per Capita PCA without COVID features

6.1.5 Conclusion.

We explored a few different scenarios in PCA. We found that for COVID related features, death rate has the largest variance but is somehow biased towards the beginning of our data. We also found hospital has a lot variances no matter whether COVID features are included.

If we exclude all COVID features, private school is a principal factor in our dataset followed by population density and hospital per person. If we only consider per-capita factors, hospital, expenditures and public school are more significant.

In conclusion, in the supervised learning part, we could pay some attention to death rate, hospital per person, school and expenditure when selecting independent variables. We also believe more granular data by time will also be helpful in exploring PCA across time.

6.2 Time Based Analysis

First, for time lag analysis, we first calculated the Pearson correlation between housing prices and the case rates. We sorted the results of all counties using p-values. The first few columns of the result can be seen below.

	county	corr	pvalue
0	Sierra County	0.731611	0.000049
1	Mariposa County	0.719009	0.000075
2	Humboldt County	0.702990	0.000128
3	Plumas County	0.673922	0.000306
4	Inyo County	0.599982	0.001940
5	Mendocino County	0.588648	0.002478
6	Calaveras County	0.563470	0.004141
7	Nevada County	0.540902	0.006348
8	Shasta County	0.538038	0.006688
9	Yuba County	0.535490	0.007004
10	Del Norte County	0.535179	0.007043

Figure 18: Price and Case Rate Correlation

We used another table to further summarize the result. As shown in the table below, in 33 counties the p-value of the correlation is smaller than the 0.05 significant level, and the number of counties grows to 46 when we altered the significant level to 0.1. The result demonstrated that there was some sort of correlation between the housing price and the covid-related data. If we applied lag time analysis and set lag month to 1, the correlation became even more obvious. However, when we increased the lag time, the correlation became weaker.

lag month	describe	p-value<0.05	p-value<0.1
0	correlation of housing price and case rate	31	43
1	correlation of housing price and case rate	33	46
2	correlation of housing price and case rate	12	18
3	correlation of housing price and case rate	12	13

Figure 19: Summary of Price and Case Rate Correlation

We also calculated the housing price change using the housing price feature, and calculated its correlation with case rate. Different from the correlation of price and case rate, only correlation in 9 counties reached the 0.1 significant level. We also calculated the correlation between case rate and the housing price after d months, and found that when d was around six, the correlation between the two variables were the most significant. Namely, the price change is more correlated with the case rate before about half of a year. The numbers of significant correlation coefficients are listed below.

lag month	describe	p-value<0.05	p-value<0.1
4	correlation between price change and case rate	3	15
5	correlation between price change and case rate	20	25
6	correlation between price change and case rate	27	31
7	correlation between price change and case rate	34	41
8	correlation between price change and case rate	25	33

Figure 20: Summary of Price Change and Case Rate Correlation

We also applied the analysis on the death rate, and the results were very similar to that of case rate. The results have shown that there exists some correlation between covid-related features and the housing price. However, the causal relationship between covid and the housing price was still uncertain. Also, in the future, we should try to verify the results using more data.

6.2.1 Autoregressive model.

We specifically used house pricing data from 2015 to 2020 from Los Angeles County(which is the most populated county in California) to train the AR model, and use the model to predict the housing price after the COVID-19 pandemic started(from 2020 to 2021) to see whether it affects the housing price or not. First, we plot the housing price data change and the correlation of price at time t and time t-1.

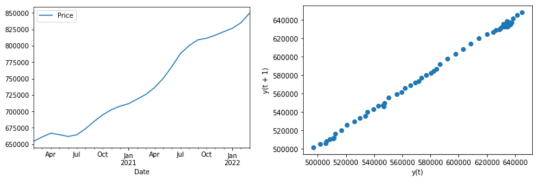

Figure 21: Housing Price Trends and Lag Correlation

As the plot shows, there is a significant correlation between past housing prices with the current ones. To further check the order of the AR model, we plot the autocorrelation and partial autocorrelation plot, as shown below.

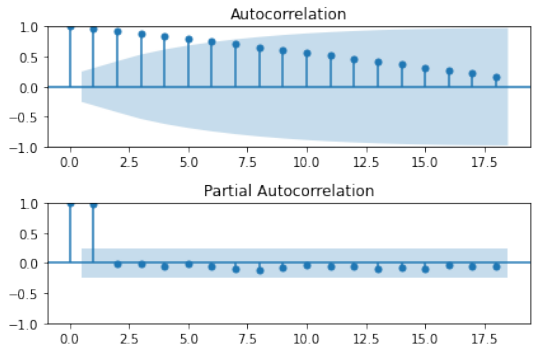

Figure 22: Autocorrelation and Partial Autocorrelation

As the plot shows, a correlation value up to order 2 is high enough in the partial autocorrelation. Thus, we will train the AR model of order 2.

Then we build the model and use the data from 2015/1 to 2019/12 to train it. The AR model results are shown below.

	coef	std err	z	P>|z|	[0.025	0.975]
const	3304.3171	2677.044	1.234	0.217	-1942.592	8551.227
Price.L1	1.5952	0.103	15.415	0.000	1.392	1.798
Price.L2	-0.5992	0.102	-5.852	0.000	-0.800	-0.398

Figure 23: Result of the AR model

The results show that 2 orders of the AR model both are significant. The housing price are highly correlated with the former housing prices. Thus, we make the predictions using housing prices since the pandemic started(from 2020/1 to 2021/12) and compare them with the true value of housing prices.

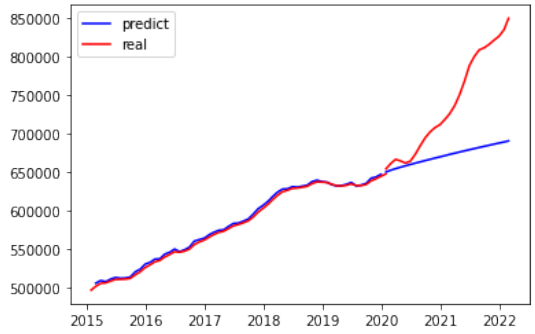

Figure 24: Real and Predicted Housing Prices from 2015 to 2022

The first spline represents housing prices from 2015 to 2020 and their prediction values using the AR model, while the second is 2020 to 2021 ones. The plot first shows that it have a great accuracy on the training set, which is housing price before the pandemic,but also clearly shows that the accuracy of the housing prices after the pandemic are really low comparing to the former ones. The model using data before the COVID-19 pandemic cannot predict the data after that well. The pandemic is highly likely to have a huge impact on housing prices. Thus, we tried to introduce exogenous variables like COVID-19 variables and other time-based variables to build an ARX model to better predict the housing price after the pandemic started in the next section.

6.2.2 *ARX model.*

As shown in the previous AR model, COVID-19 pandemic may have a great influence on the housing price. Thus, we further use the ARX model to take COVID-19 and other variables into account.Since the COVID-19-related data exists only after the pandemic starts, the time-series data we are using are from 2020/2 to 2021/12. We use the last 4 months as test sets and the remaining as training sets. By using step-wise procedures, we did the feature selection and removed insignificant features from the model. The final features we used were housing inventory, the death rate of COVID-19, the fully vaccinated rate, and the number of inventory that prices increased compared to last month. We first used all features except the COVID-19 features to build an ARX. The results are shown below. All coefficients are significant.

	coef	std err	z	P>|z|	[0.025	0.975]
const	6.394e+04	3.39e+04	1.884	0.060	-2583.010	1.3e+05
Price.L1	1.7724	0.171	10.370	0.000	1.437	2.107
Price.L2	-0.8674	0.206	-4.210	0.000	-1.271	-0.464
Housing Inventory	-1.4895	0.745	-2.000	0.046	-2.949	-0.030
inventory_price_increased	33.1882	7.233	4.588	0.000	19.011	47.365

Figure 25: ARX model without COVID-19 features

Then we include the COVID-19 features into the ARX model. The results are shown below. The coefficients of two COVID-19 features are significant and both AIC and BIC dropped, which means this model have a better performance when predicting housing prices.

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.049e+05	2.96e+04	3.542	0.000	4.69e+04	1.63e+05
Price.L1	1.6702	0.139	11.979	0.000	1.397	1.943
Price.L2	-0.7857	0.166	-4.727	0.000	-1.111	-0.460
Housing Inventory	-4.3155	1.046	-4.124	0.000	-6.366	-2.265
death_rate	-8.622e+06	3.8e+06	-2.270	0.023	-1.61e+07	-1.18e+06
fully_vaccinated_rate	-7.674e+04	2.38e+04	-3.218	0.001	-1.23e+05	-3e+04
inventory_price_increased	40.1292	7.148	5.614	0.000	26.120	54.138

Figure 26: ARX model with COVID-19 features

Next, we used the test dataset to predict housing prices on two ARX models. The results are plotted below. As we can discover directly in the plot, the second ARX models, which contains COVID-19 features, have a much better prediction accuracy.

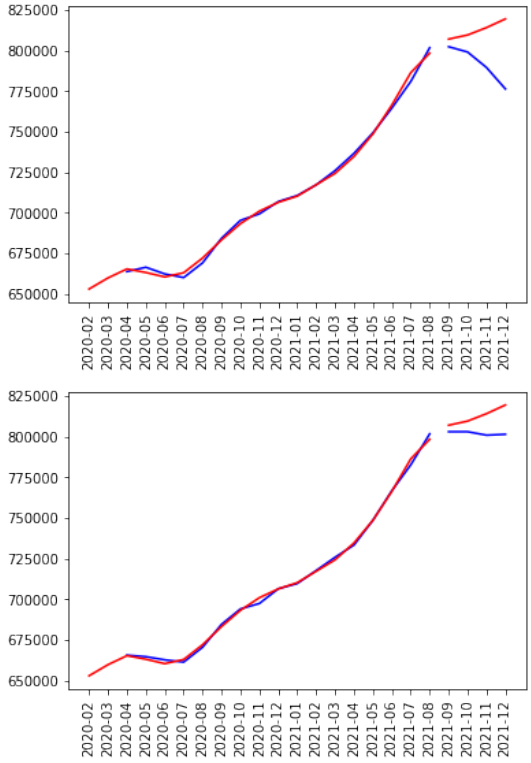

Figure 27: Prediction Results of Model without COVID-19 features(upper) and Model with COVID-19 features(lower)

Considering all AR and ARX models, we found that COVID-19 certainly has a huge impact on the housing price. As the pandemic started, the housing price increased tremendously. Including COVID-19 features into the model can help improve the accuracy.

6.3 Regression Analysis

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. We first compared the linear regression model with the Ridge and Lasso regression models. The evaluation results are shown in the table below:

	Linear	Ridge	Lasso
EVS	0.8484	0.8459	0.8474
R^2	0.8483	0.8460	0.8474
MSE	1.8e10	1.9e10	1.9e10

Figure 28: Comparison of the Linear, Ridge, and Lasso regression models

The results indicate that the three models have highly similar performance on our dataset. We also plotted histograms of residuals for three models on testing set:

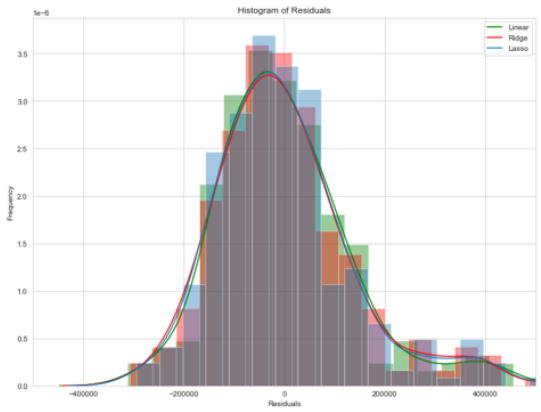

Figure 29: Residuals histograms of the Linear, Ridge, and Lasso regression models

As both results show that the linear, Ridge, and Lasso regression does not have significant difference in performance, we planned to only keep the linear regression model for further comparison with other models. Below table shows the evaluation results of all other models:

	Linear	Decision Tree	Random Forest	KNN
EVS	0.8484	0.9755	0.9829	0.9817
R^2	0.8483	0.9754	0.9829	0.9817
MSE	1.8e10	3.0e9	2.1e9	2.2e9

Figure 30: Comparison of the Linear, Decision Tree, Random Forest, and KNN regression models

The results indicate that the random forest and KNN regression model have the overall best performance, the decision tree regression model has comparable performance, and all of them perform much better than the linear regression model. We plotted the housing price versus predicted housing price and residuals for both linear and random forest regression models to have a more intuitive comparison between them:

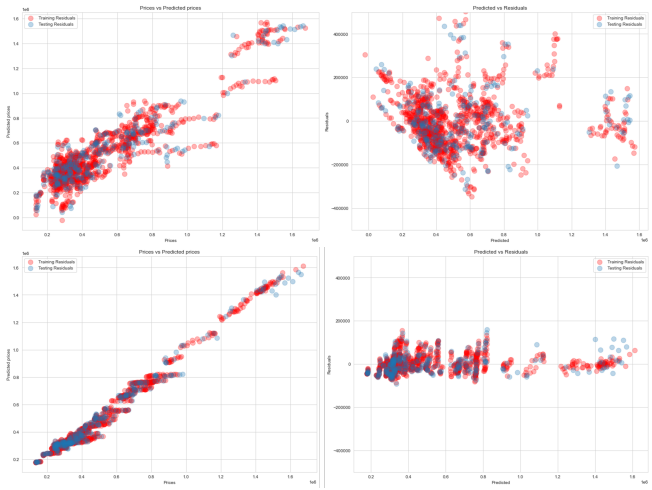

Figure 31: Housing price versus predicted housing price and residuals for the Linear and Random Forest regression models

From the figures we can clearly observe that the random forest model has much better correlation between the price and predicted price and much smaller residuals comparing to the linear regression model. Then we can plot the residuals histogram on both training and testing set to further analysis the performance of the two models. The histograms indicate that the random forest model has much smaller residuals for both training and testing sets, comparing to the linear regression model. However, it has a larger difference between the training and testing sets, which illustrates a potential overfitting to this model.

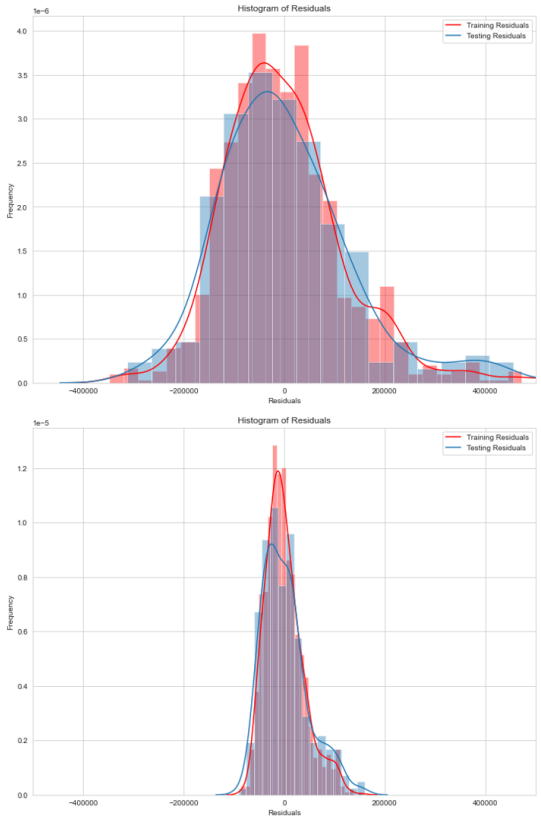

Figure 32: Residuals histograms of the Linear and Random Forest regression models

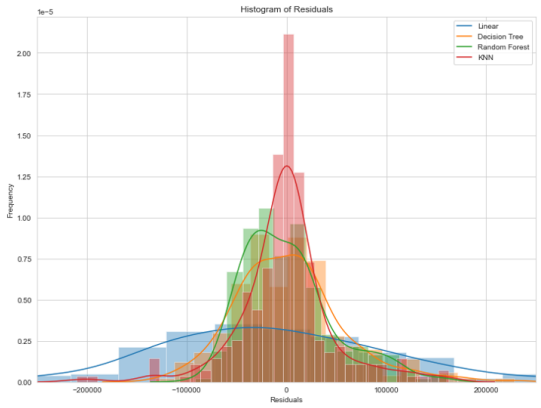

Figure 33: Residuals histograms of all the regression models

Finally, we plotted the residuals histogram of all the models and compare their results. We can observe that although the random forest and KNN models have quite similar results on *EVS*, R^2, and *MSE*, the KNN model has significantly smaller residuals. The random forest model has comparable residuals to the decision tree model, and the linear regression model has much larger residuals than all of the other models.

Last but not least, we would like to discuss why random forest model has a better performance. We first 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.

7 CONCLUSIONS

Overall, we develop different machine learning models to predict housing price in California during the COVID Period. We collected data from various sources and preprocessed them to be ready for machine learning usage. Then we implemented different unsupervised and supervised machine learning models for dimension reduction, time-based analysis and regression. Specifically, through dimension reduction, death rate and fully vaccinated rate are the most significant factors. For per-capita data, death rate, hospital per person and are the most significant factors. While using the time-based analysis model, with COVID-19 features have a much better accuracy, so we could deduce that COVID-19 have a huge influences on housing prices. Finally, for Regression Analysis, With random forest and KNN regression models, the housing prices in different counties at specific time can be predicted with high accuracy.

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] 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.

[5] 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

[6] 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

[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