A Real-world Real Estate Problem- Find the Top 10 Undervalued Projects

ECON 445 Course Project

Keren He (605646313), Suyun Cao (905630669),
Zongyang Li (105636910), Kunyi Jiang (005642073)

***Abstract*—We mainly focus on building an optimal model to predict the actual value of a property. To be specific, we select 10 features which are most helpful in prediction. And we tried Multiple Linear Regression model, Ridge Regression, LASSO Regression, ElasticNetCV, and Support Vector Machine to predict the actual value. At last, we find that the LASSO Regression model has the highest average cross validation score. Thus, we can conclude that this model is the optimal model for this project to predict the actual value of a property. And then we report top 10 undervalued projects based on this model.**

I. INTRODUCTION

The dataset that we are working on consists of the market value of house properties in California in decades. We plan to initiate an investigation on what variables might be essential in affecting the market value of the properties through appropriate models.

After a thorough discussion of potential variables that are necessary for our project study. We decided to involve both qualitative and quantitative variables, as we assume the price of the house properties is not only related to quantitative factors such as the size of floor or number of bedrooms but also is highly correlated to the physical location of the properties, as well as their use types.

Therefore, we have selected the net value of house property as our dependent variable, and other 9 elements as independent variables.All the 10 variables are showed in table 1.

Variables	***Variables Description***
netTaxableValue	net value of house
EffectiveYearBuilt	effective year built
SQFTmain	total square footage
Bedrooms	number of bedrooms
Bathrooms	number of bathrooms
Units	number of living units
PropertyType	property use type
GeneralUseType	general use type
IsTaxableParcel	Y=taxable fee parcel; N=non-taxable non-fee
AdministrativeRegion	assessor's administrative office

II. DATA PROCESSING

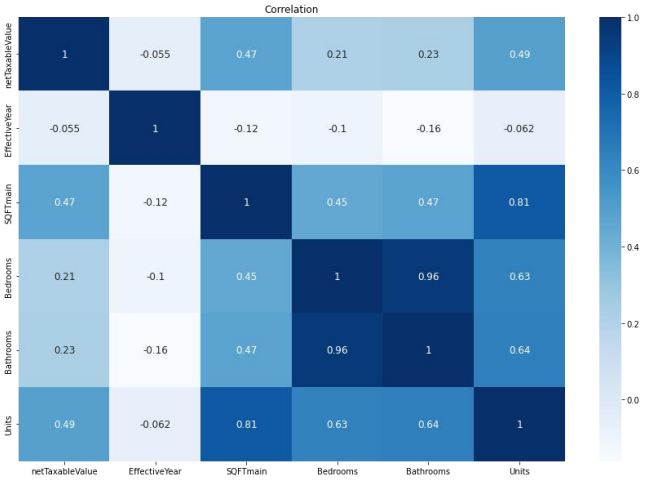
The first step of our analysis will be the pre-processing of our dataset to make it more manageable for conducting next-step modeling. With previewing the raw dataset, we found there are a few observations that contain missing values or NAs, we managed to eliminate the observations that contain unavailable information.

Considering the raw dataset has over two million observations, which makes the modeling on the complete dataset unrealistic with our current electronic devices. To maintain the objectivity of the project, we decided to initiate a random sampling on the raw dataset to withdraw enough observations randomly as our sample dataset so that it will reduce the possibility of bias to the largest extent.

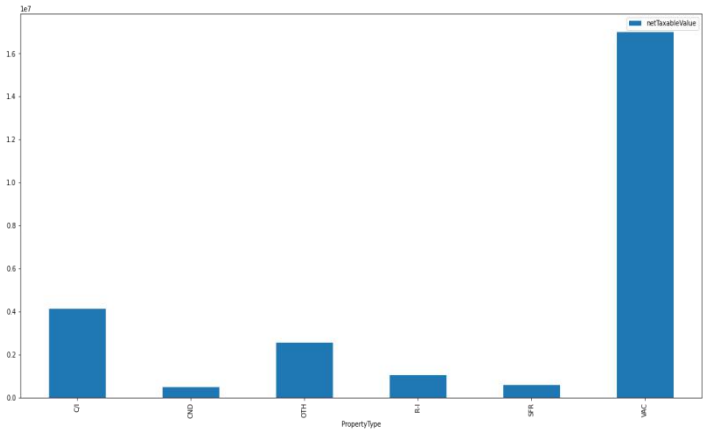
With the above processing, we have then acquired a sample dataset consisting of 3000 observations picked by random sampling method with no missing values existing. This sample dataset will be appropriate for further modeling and analysis.

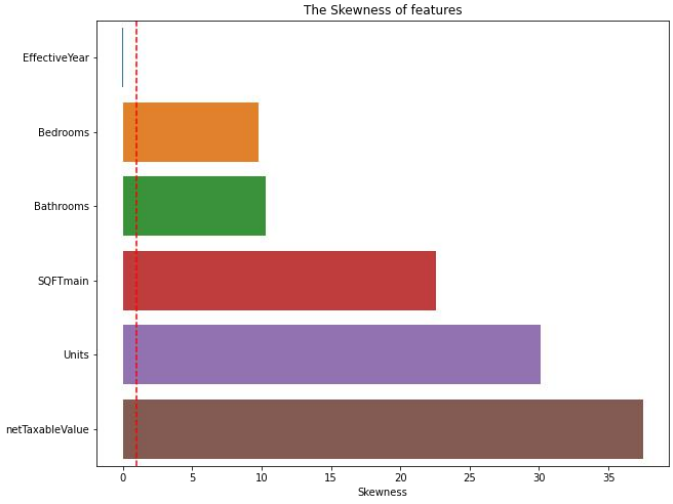
Since we have included both qualitative and quantitative variables, it would be irrelevant to directly fit models with this dataset without transformations on the variables. As such, we have classified all qualitative variables through visualization tools and transformed them into dummy variables. To be more specific, we intend to analyze whether a certain qualitative variable would be effective to our dependent variable. For example, for a qualitative variable such as "Property Types", there are originally 6 different types, by using a bar plot, we observed that the majority number of types goes to type "VAC", and in this case, it would be most appropriate to measure whether "VAC" can be an effective component that affects the net value of the house properties. We then set type "VAC" as 1 and other types as 0. Similarly, the rest of the qualitative variables were also transformed by the above way of thinking.

Meanwhile, we also conducted a correlation check on our qualitative variables to see if there is any correlation between them. By using heatmap as the visualization tool, we observed that there is a strong correlation of 0.96 between variables "Bedroom" and "Bathroom". This conclusion is in fact consistent with our assumptions as if the number of bedrooms increased for a house, the number of bathrooms should also increase to satisfy the basic needs for house living. To avoid problems of multicollinearity, we decided to drop variable "Bathroom" so that our results will present in better objectivity and less bias.

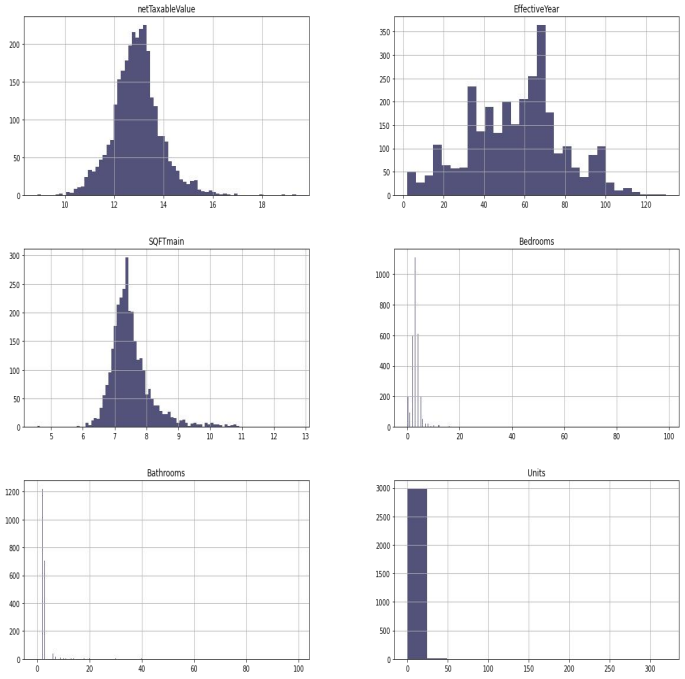


After checking the correlations between the qualitative variables, we then generated another bar plot to measure the skewness of the variables. As result, we observed that only "Effective Year" is not highly skewed, while keeping "Bedrooms" and "Bathrooms" and "Units" the same, we decided to take log transformation on "SQFTMain" and "NetTaxableValue" for better modeling purposes.





With the log transformation on “SQFTMain” and “NetTaxableValue”, we generated distribution plots to ensure the distribution of the variables followed a normal distribution. The reason for us not forcing the variables “Bedrooms”, “Bathrooms” or “Units” to follow a normal distribution because as such will be violating the natural property of these variables as the number of these variables have never followed a random normal pattern, but only obeyed on the actual needs based on the design of the house property itself.



III.MACHINE LEARNING MODELS

In this part, we will use five algorithms, Multiple Linear Regression model, Ridge Regression, LASSO Regression, ElasticNetCV, and Support Vector Machine to predict the actual value. The target label is netTaxableValue, and we include 8 independent variables in the model, which are EffectiveYear, SQFTmain, Bedrooms, Units, PropertyType Map, GeneralUseType Map, isTaxableParcel Map and AdministrativeRegion Map.

- Step 1: Split dataset

In this project, we randomly select 80% of the raw data as the training set, and the remaining 20% of the data is the testing set. That is, the training set contains 2400 observations.

- Step 2: Run the training on 5 algorithms

According to the average cross validation score and performance measurements, such as MAE, MSE and R^2 of each model, we can figure out which model outperforms the others.

- Step 3: Report top 10 undervalued projects based on the best model.

A. Multiple Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. And the performance of this model is shown in the following table:

Training set score	0.39
Test set score	0.30
average cross validation score	0.3892

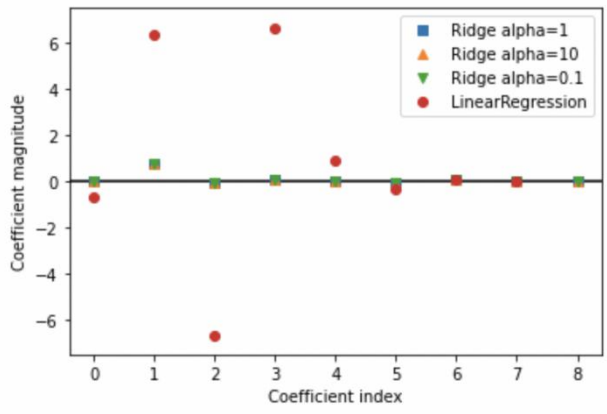
B. Ridge Regression

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated. Here, three different Alphas are applied. And the performance of three Ridge Regression models is shown in the following table:

Ridge alpha=1	Training set score	0.30
	Test set score	0.29
Ridge alpha=10	Training set score	0.39
	Test set score	0.30
Ridge alpha=0.1	Training set score	0.40
	Test set score	0.31

According to the above table, Ridge alpha=0.1 is chosen to be the parameter for Ridge Regression. And its average cross validation score is 0.3893.

For Multiple Linear Regression and Ridge Regression, coefficient magnitudes of each independent variable are plotted in the following figure.



From the figure above, the coefficient magnitudes of each independent variable in Linear Regression are more significant than in Ridge Regression. This means that it is more difficult to explain effects of dependent variables on house price in Ridge Regression than MLR.

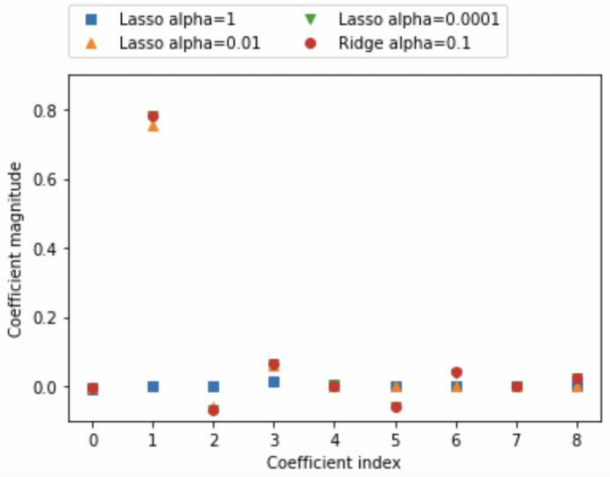
C. LASSO Regression

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). Here, three different Alphas are applied. And the performance of three Ridge Regression models is shown in the following table:

Lasso alpha=1	Training set score	0.12
	Test set score	0.10
Lasso alpha=0.01 (5 features used)	Training set score	0.40
	Test set score	0.31
Lasso alpha=0.0001 (8 features used)	Training set score	0.40
	Test set score	0.31

According to the above table, Lasso alpha=0.01 is chosen to be the parameter for Lasso Regression. And its average cross validation score is 0.3895.

For Ridge Regression and Lasso Regression, coefficient magnitudes of each independent variable are plotted in the following figure.



From the figure above, the coefficient magnitudes of each independent variable in Ridge Regression and Lasso Regression are very similar. This means that there is not too much difference between Ridge Regression and Lasso Regression for this dataset.

D. ElasticNetCV

ElasticNetCV is a cross-validation class that can search multiple alpha values and applies the best one. According to this model, only 5 features have

been used in the model. And the performance of this model is shown in the following table:

Training set score	0.40
Test set score	0.31
average cross validation score	0.3895

E. Support Vector Machine

Support-vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. Here, a Linear SVC (Support Vector Classifier) is used to fit the dataset. And the performance of this model is shown in the following table:

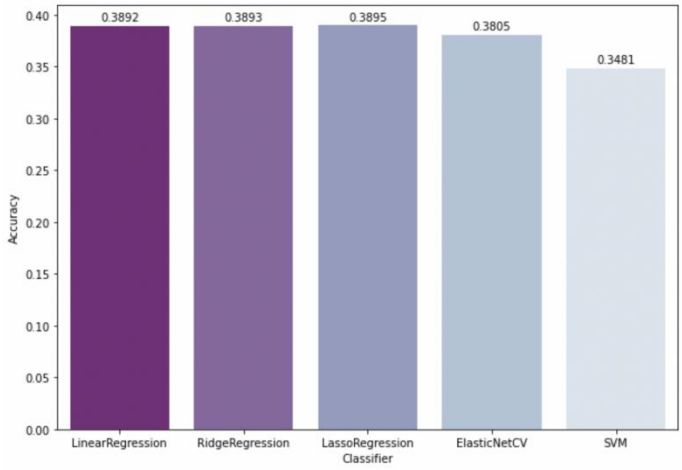
Training set score	0.39
Test set score	0.30
average cross validation score	0.3481

IV. MODELS SUMMARY

Model	**MAE**	**MSE**	**R^2**
Linear Regression	0.624411	0.645175	0.307683
Ridge Regression	0.624410	0.645178	0.307680
Lasso Regression	0.624327	0.644935	0.307941
ElasticNetCV	0.622171	0.639239	0.314053
SVM	0.606329	0.644385	0.308532

From the table above, the ElasticNetCV Model has the highest R^2, which means that the ElasticNetCV Model fits the data better than the other models.

Then average cross validation scores of each model are plotted in the following figure.



From the plot above, the Lasso Regression Model has the highest CV scores. Therefore, the Lasso Regression Model is chosen to be final model to predict the actual house price and find the top 10 undervalued projects.

V. FIND UNDERVALED PROJECTS

- Step 1: Conduct the Regression Model on the testing set
- Step 2: Make prediction on actual house price
- Step 3: Use predicted house price minus real house price to get the difference
- Step 4: Sort the difference from largest to smallest and output the top 10 projects

Top 10 undervalued projects are identified by the Lasso Regression Model.

	ID	SQFTmain	Bedrooms	Bathrooms	Units	PropertyType	GeneralUseType	isTaxableParcel	AdministrativeRegion	netTaxableValue	EffectiveYear
0	996035	4088.0	5	3	1	SFR	Residential	Y	07	15085235.0	52.0
1	940303	5777.0	4	8	1	SFR	Residential	Y	07	8251730.0	64.0
2	938997	2393.0	5	2	1	SFR	Residential	Y	07	2325724.0	84.0
3	890481	976.0	2	1	1	SFR	Residential	Y	07	1202328.0	94.0
4	2424277	2171.0	3	2	1	SFR	Residential	Y	09	2040000.0	91.0
5	976560	8633.0	6	7	1	SFR	Residential	Y	07	10298569.0	14.0
6	2451092	100.0	0	0	1	C/I	Commercial	Y	25	268490.0	22.0
7	957070	2970.0	4	4	1	SFR	Residential	Y	07	2988030.0	51.0
8	982428	2602.0	4	3	1	SFR	Residential	Y	07	2490158.0	47.0
9	925110	1488.0	2	3	1	SFR	Residential	Y	07	1433433.0	92.0

Based on the output above, we can conclude that the property type of highly undervalued projects is very likely to be SFR, which means Single Family Residence. And the general use type is very likely to be residential. And highly undervalued projects seem to have more than 2 bedrooms and bathrooms. And these projects are basically more than 50 years old.

VI. REFERENCES

[1] Gloria Russell (2021) 10 Important Features to Consider When Buying a House. https://homeia.com/10-important-features-to-consider-when-buying-a-house/

[2] Nelson Lau (2020) 5 Ways to Apply Data Science to Real Estate. Towards Data Science. https://medium.com/towards-data-science/5-ways-to-apply-data-science-to-real-estate-e18cdcd0c1a6

[3] Xian Guang LI, Qi Ming LI (2006) THE APPLICATION OF DATA MINING TECHNOLOGY IN REAL ESTATE MARKET PREDICTION. The CRIOCM 2006 International Symposium on "Advancement of Construction Management and Real Estate". https://www.irbnet.de/daten/iconda/CIB5807.pdf