Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.10, No.5, 2020

Dimensionality Reduction (Linear Technique Using California Housing Data)

Lyn Mushanyuki[1], Ren Dongxiao[2*]
School of Sciences,
Zhejiang University of Science and Technology, Hangzhou-China.
rendx29@163.com*

Abstract
Most automatic pricing systems for existing homes are based solely on specific text data, such as neighborhood and number of rooms. The final price will be determined by a human agent who visits the house and visually evaluates. In this article, we suggest removing the visual features of the images from the house and merging them with the text information of the house. The data set consists of 535 model houses from the state of California in the United States. Our experiments showed that adding visual features increased and reduced the average error by an order of magnitude compared to plain text properties. Linear techniques for reducing dimensionality, especially the analysis of key components, are often used in data analysis to interpret high-dimensional datasets. These linear methods may be appropriate for analyzing nonlinear process data in the housing system. Recently, a lot of techniques have been developed to reduce nonlinear dimensionality, which can be a potentially useful tool for identifying small varieties in climatic data sets due to nonlinear dynamics. In this paper, I used linear techniques to do data analysis for prediction.
Keywords: *Dimensionality Reduction, Statistics, Housing Data, Predictions, Linear and Nonlinear*

1.0 Introduction

Dimensionality Reduction: There have been techniques for reducing dimensions for more than a century. They are developed in all statistical field such as, machine learning areas and fields used to analyze high size data. High-dimensional data refers to data that requires more than two or three dimensions to represent, and this sort of data can be difficult to interpret. Dimensionality reduction is therefore, the process of decreasing the total of random variables in respect to locating a set of key variables.

Most of the problems related with information processing require a form of dimensionality reduction. Researchers working in various fields, including computer science, astronomy, real estate, bioinformatics, remote earth exploration, economy and facial recognition, are trying to reduce the number of data variables. It will be very interesting to see that reducing the dimensions of your data is a big problem in many areas. The reduction of the dimensionality can be linear and nonlinear, depending on the technique used. If the original large dataset contains nonlinear relationships, nonlinear approaches may be more suitable.

The housing market plays an important role in shaping the economy. Home and housing renovations drive the economy by increasing sales of housing, employment and spending. Demand from other relevant areas, such as building needs a sustainable housing property [1] is also surprising. The value of the asset portfolio for households whose homes are the largest asset is strongly influenced by rise and fall in property prices. Recent studies show that the internal market affects the effectiveness of financial institutions, which in turn influences the surrounding financial system. In addition, the housing sector is an important indicator of asset prices that can be included in both the real estate sector and production prediction [1].

The traditional pricing process is based on price comparison and cost comparison, which is unreliable and is not accepted by the standard and certification process [2]. Therefore, accurate automatic housing price prediction are needed to help policy makers shape strategies and manage inflation and help people with smart investment plans [1].

Housing price predicting is a very difficult task as a result of both physically and geographically factors in the housing market. There is also a slight synergy between the price of the house and several other macroeconomic factors, which makes the predicting process very difficult.

Previous studies have been conducted to investigate the main factors influencing the price of housing. All works so far are based on the text characteristics of the house [3]. Second [4] the price of the house has influenced several factors such as silk, rank, number of bedrooms and bathrooms. The more bedrooms and bathrooms the

house has, the higher the price. We therefore rely on these factors together with the look of the houses to estimate the price.

1.1 Motivation for the Study

The goal of the paper is to predict average housing prices in California, as many properties are located in those districts. The project also aims to create a California housing price model using California Census data. Data include statistics such as population, median income, average house price, etc. for each block group in California using Linear Regression Techniques. This model should learn from the data and be able to predict the average price of housing in each neighborhood, given all other statistics.

2.0 Literature Review

During the last decade, some work has been done to automate the real estate price evaluation process. The emphasis is on acquisition properties such as property quality, property prices, environment, and location. Compared to different methods, we found that the above method can be divided into two main categories: models that qualify data and models based on data integration. However, data integration models estimate their value depending on all the features of the home, such as neural networks and backward models. For example, the hedonic price theory, where real estate prices are a function of its qualities, is an example of the data breakdown model. The attributes associated with a property determine a set of obvious prices. The problem with this method is that the difference between the different characteristics of the same geographical area is not considered. So, it is considered unrealistic [5] tries to estimate the best value of a property by comparing the data to the aggregate and the results of the disappointment. They found that the results of the merger were more accurate. They also found that the hedonic value for some partners is not stable for some quality because it depends on the location, age and type of property. Therefore, they realize that hedonic analysis can be effective in analyzing these changes, but not guessing the value based on each feature. They also found that the geographic location of the property plays an important role in influencing property prices for data integration models; neural networks are the most common models. [6] Compares neural network performance with multiline background (MLR). NN achieved a higher R standard and less than MLR. Compared to the results of the hedonic model compared to neural network models, the neural network surpassed the hedonic model reaching a higher R2 standard of 45.348% and 48.8441%. Lack of information in the hedonic model can lead to poor results. However, this model has some limitations as the estimated price is not the actual price, but it is real. This is due to the difficulty of collecting actual market data. In addition, neural networks cannot automatically verify whether the time effect plays an important role in guesswork. This means that property prices are influenced by many other economic factors that are difficult to include in the speculative process.

A housing market analysis and housing price assessment literature reveals two key trends in research: the use of the eon regression approach and artificial intelligence techniques to develop models for predicting home prices. For decades, several eonic methods have been used to identify the relationship between property prices and housing characteristics [7]. [8] Developed it on regression approaches to fundamentally assess the market impact on property price dynamics.

In his 2004 study, [9] examined the heteroscedasticity in hedonic models of home prices based on the average age of households in Boston, Massachusetts, as data. The results supported heteroscedasteation tests compared to age at home in previous results. He evaluated an economic pricing function with semi-parameter regression and compared price prediction performance with conventional parameter models.

The results showed that linear regression performed better in both sample price prediction and sample prices and could be used to measure and predict property prices. [10] Examined family-level data on building models to quantify the implicit heterogeneity of prices in terms of family type, age, educational level, income and former buyer properties for both first-time buyers and previous owners.

However, heterogenic methods have potential limitations in the assumptions and estimates of the base model. These include the determination of supply and demand, market imbalance, selection of independent variables, choice of functional form of hedonic comparison and market fragmentation [11], [12]. Latest studies have concentrated on the outcomes of price prediction.

2.1 Dimensionality Reduction

Dimensionality reduction describes the process of reducing the number of variables in a dataset to eliminate unnecessary variables and dependencies [13]. This usually results in a smaller integration of the data, which still records a large part of the structure of the original data. A smaller view of the data offers a number of advantages. For example, we can store data more efficiently because we can remove unnecessary variables and still count

most of the data with the remaining variables. More importantly, reducing dimensionality can help you understand the inherent structure of a dataset. For example, if you take a layer in 2D space and twist it in the form of "JA" (now shown in 3D), this does not change the fact that the original structure is two-dimensional. The resulting 3D object has an inherent dimensionality of 2, rotation is not useful for explaining the data in general, and can be ignored. Good dimension reduction algorithms determine that the underlying structure of the object is a 2D plane.

3. Methodology (Design Approach)

This section provides a methodical view of the dimensional reduction techniques we use in our analysis. As the preponderance of these technique increases, it is important that researchers and end users have a good understanding of their underlying technical framework. With summaries of the following technique, we hope to achieve this goal.

A. Linear Regression Technique:

Linear regression is a very powerful statistical technique. Many people have some knowledge of regression just by reading the news, where the graphics are covered with straight lines in the trash plots. Linear models can be used to predict or evaluate whether there is a linear relationship between two numeric variables. Linear regression statistics allow us to summarize and explore the relationship between two continuous variables. A variable named x is considered a prediction indicator, an explanation, or a separate variable.
The second variable called as a response, result, or dependent variable.

B. Multiple Regression Analysis

Multiple regression analysis is used to verify a statistically noticeable association in the middle of variable sets. It is used to detect patterns in the information specified in the individual. Numerous relapse studies will have almost the same basic straight relapse in the same way. The main distinction in the middle of the equal direct relapse is also used in the number of predictors ("variable x") within these relapses. The direct reversion control uses an absolute x variable for each subordinate variable" and "y". Case in location: (x1, Y1). Numerous spare consumption variables for each free variable: ((x1)1, (x2)1, (x3)1, Y1). Either way, you'll be intrigued by how different types of notifications offers affect relapse. You set your X1 as a specific type to sell, your X2 seems to transact, etc.
Linear regression uses a single predictor variable to explain a dependent variable. A simple linear regression equation is as follows:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Where:
y = dependent variable
β = regression coefficient
α = intercept (expected mean value of housing prices when our independent variable is zero)
x = predictor (or independent) variable used to predict Y
ϵ = the error term, which accounts for the randomness that our model can't explain.

4.0 Experiment and Analysis of Data

The data refers to households in a particular area of California and aggregate statistics from the 1990 census. So, while it may not help predict current home prices, such as Zillow's estimation record, it provides a first cost-effective dataset to teach people the basics of machine learning and data science.

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanielton, WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Raymond\nFPO AE 09386

Figure 1: Dataset

The columns are as follows; their names are self-explanatory.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 7 columns):
Avg. Area Income                5000 non-null float64
Avg. Area House Age             5000 non-null float64
Avg. Area Number of Rooms       5000 non-null float64
Avg. Area Number of Bedrooms    5000 non-null float64
Area Population                 5000 non-null float64
Price                           5000 non-null float64
Address                         5000 non-null object
dtypes: float64(6), object(1)
memory usage: 273.6+ KB
```

Figure 2: Attributes of Dataset

There are 5000 instances in the data set, but the "**avg_bedrooms**" attribute has only 500 values .The type is an object so it can contain any type of python object, but because the data is retrieved from the CSV file, we know that it must be a text attribute. The summary of all numerical attributes:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+03
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.232073e+06
std	10657.991214	0.991456	1.005833	1.234137	9925.650114	3.531176e+05
min	17796.631190	2.644304	3.236194	2.000000	172.610686	1.593866e+04
25%	61480.562388	5.322283	6.299250	3.140000	29403.928702	9.975771e+05
50%	68804.286404	5.970429	7.002902	4.050000	36199.406689	1.232669e+06
75%	75783.338666	6.650808	7.665871	4.490000	42861.290769	1.471210e+06
max	107701.748378	9.519088	10.759588	6.500000	69621.713378	2.469066e+06

Figure3: Numerical Attributes

A brief look at the distribution of all numeric attributes shows that avg_income is a very important feature in the prediction. Therefore, you must ensure that the test set represents the different revenue categories in the dataset. Now that support is a continuous numerical feature, you must first create a feature for the sales category. It is important to have quite a few issues with each team's files. In other words, it's hard to be able to afford too many shifts, and all layers must be big enough.

Linear Regression Model:

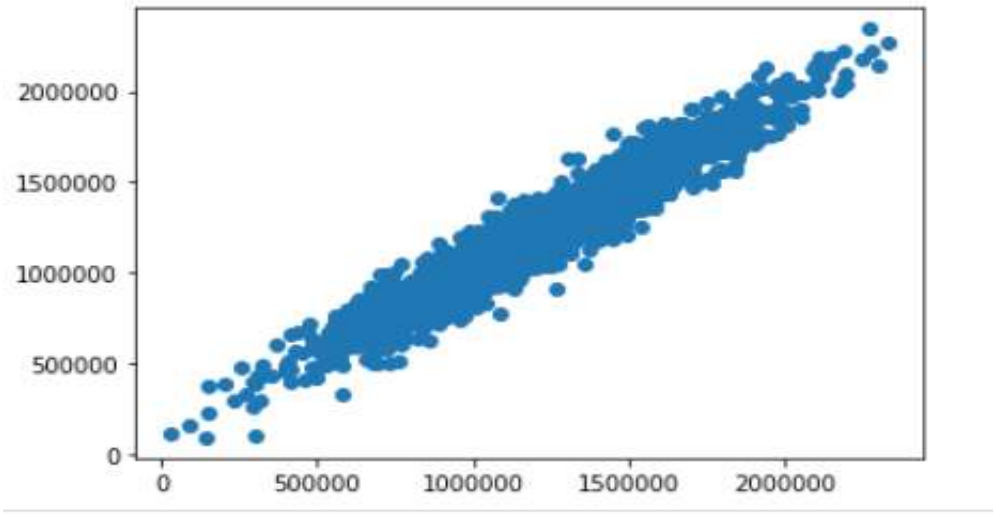
To use linear regression, the quadrant attributes are assigned to the X axis and the y-axis values. For each property, linear regression is achieved once. The independent X axis is the option available to the user.

```
print(lm.intercept_)
```

```
-2640159.796851911
```

```
coeff_df = pd.DataFrame(lm.coef_,X.columns,columns=['Coefficient'])
coeff_df
```

	Coefficient
Avg. Area Income	21.528276
Avg. Area House Age	164883.282027
Avg. Area Number of Rooms	122368.678027
Avg. Area Number of Bedrooms	2233.801864
Area Population	15.150420




```
X = HouseDF[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',
             'Avg. Area Number of Bedrooms', 'Area Population']]

y = HouseDF['Price']
```

Figure 4(a): Outcome of the Model Prediction

In the above scatter plot, we see data is in line shape, which means our model has done good predictions

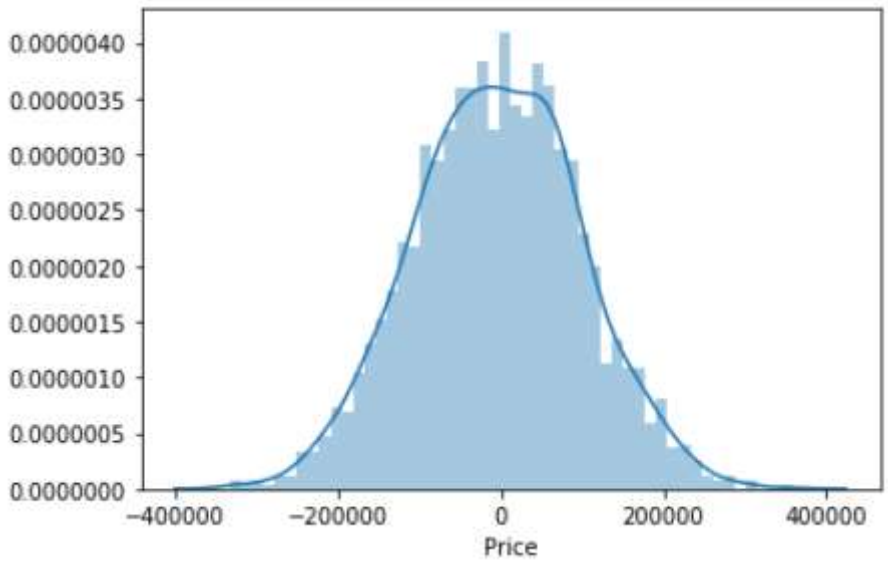


Figure 4(b): Outcome of the Model Prediction

In the above histogram plot, out data is in bell shape (Normally Distributed), which means our model has done good predictions. This means that the features provide enough information to make enough predictions, or the model is strong enough. The main way to resolve nonconformance is to choose a more powerful model and provide better functionality to training algorithms or reduce model limitations. Due to the simplicity of linear algorithm and its suitability for the linear system, it is widely used in industrial sectors and other fields.

Evaluation Metrics

```
from sklearn import metrics
```

```
print('MAE:', metrics.mean_absolute_error(y_test, predictions))
print('MSE:', metrics.mean_squared_error(y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predictions)))
```

```
MAE: 82288.22251914957
MSE: 10460958907.209501
RMSE: 102278.82922291153
```

A perfect model that can predict the exact value of an apartment is very difficult, if not impossible - to achieve because it covers all aspects of what makes the apartment valuable. Trying to achieve such a model will be very complicated, with variables that are difficult to measure and can differ between people. This is why we believe in a simpler model that is easy to understand and usable, which at the same time largely predicts the value of the apartment. This model will provide a good illustration of the importance of the various factors in assessing the opportunity.

5.1. Conclusion

This paper provides a linear regression for assessing housing prices due to macroeconomic factors and how to assess the quality of the underlying linear regression model. There is no doubt that clarifying house prices is a complex problem. There are many others that can be used. And even more complex, these variables can affect

each other at the same time. Data scientists are urged to delve into the data and adapt to the model by adding and removing the variables, recalling the importance of Ordinary least squares (OLS) regression assumptions and regression results.

References

[1] Li, Y., Leatham, D. J., et al. (2011). Forecasting Housing Prices: Dynamic Factor Model versus LBVAR Model.

[2] Khamis and Kamarudin (2014). Comparative Study on Estimate House Price Using Statistical and Neural Network Model

[3] Ng, A. And Deisenroth, M. (2015). Machine Learning for a London Housing Price Prediction Mobile Application.

[4] Limsombunc, V., Gan, C., and Lee, M. (2004). House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. American Journal of Applied Sciences, 1(3):193–201.

[5] Fletcher, M., Gallimore, P., and Mangan, J. (2000). The Modelling of Housing Submarkets. Journal of Property Investment & Finance, 18(4):473–487

[6] Bin, O. (2004). A Prediction Comparison of Housing Sales Prices by Parametric Versus Semi-Parametric Regressions .Journal of Housing Economics, 13, 68–84.

[7] Adair, A., Berry, J., & McGreal, W. (1996). Hedonic Modeling, Housing Submarkets and Residential Valuation. Journal of Property Research, 13(1), 67–83.

[8] Meese, R., & Wallace, N. (2003). House price dynamics and market fundamentals: The Parisian Housing Market. Urban Studies, 40(5–6), 1027–1045.

[9] Stevenson, S. (2004). New Empirical Evidence on Heteroscedasticity in Hedonic Housing Models. Journal of Housing Economics, 13, 136–153

[10] Kestens, Y., Theriault, M., & Rosier, F. D. (2006). Heterogeneity in Hedonic Modelling Of House Prices: Looking At Buyers' Household Profiles. Journal of Geographical Systems, 8(1), 61–96.

[11] Fan, G., Ong, Z. S. E., & Koh, H. C. (2006). Determinants of House Price. A Decision Tree Approach. Urban Studies, 43(12), 2301–2315.

[12] Schulz, R., & Werwatz, A. (2004). A state space model for Berlin House prices: Estimation and Economic Interpretation. Journal of Real Estate Finance and Economics, 28(1), 37–57.

[13] J. A. Lee, M. Verleysen. Nonlinear Dimensionality Reduction. Springer Science & Business Media, 2007 (cit. on pp. 17–19, 21

Funding:

This work was supported in part by the **Cooperative Education Project of the Ministry of Education** under **Grant 201901020031, 201901108008 And 201901167001 and Graduate Teaching Reform Research Project of Zhejiang University of Science And Technology Under Grant 2019yjsjg03.**