Real estate value prediction using multivariate regression models

IOP Conf. Series: Materials Science and Engineering **263** (2017) 042098 doi:10.1088/1757-899X/263/4/042098

Real estate value prediction using multivariate regression models

R Manjula, Shubham Jain, Sharad Srivastava and Pranav Rajiv Kher
School of Computing Science and Engineering, VIT University, Vellore, Tamil Nadu – 632014, India

E-mail: rmanjula@vit.ac.in

Abstract. The real estate market is one of the most competitive in terms of pricing and the same tends to vary significantly based on a lot of factors, hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy. Therefore in this paper, we present various important features to use while predicting housing prices with good accuracy. We have described regression models, using various features to have lower Residual Sum of Squares error. While using features in a regression model some feature engineering is required for better prediction. Often a set of features (multiple regressions) or polynomial regression (applying a various set of powers in the features) is used for making better model fit. For these models are expected to be susceptible towards over fitting ridge regression is used to reduce it. This paper thus directs to the best application of regression models in addition to other techniques to optimize the result.

1. Introduction

Predicting housing prices has always been a challenge for many machine learning engineers. It has been hosted as part of the haggle competition. Several researchers have tried to come with a model to accurately predict housing prices with high accuracy and least error. These models are created using various features such as square feet of the house, number of bedrooms, ambiance etc. Some of the researchers have used techniques like clustering [2] for grouping same houses together and then estimating the price. So each of the features in our model is given certain weight and it determines how important is that feature towards our model prediction. This is called feature engineering. Most companies which do real estate business have probably a billion different features to choose from however one of the drawback of having a large number of feature involved is the heavy computations involved in making the regression model, and computing the gradient descent solution. Later we introduce to another algorithm called as coordinate descent algorithm which drastically reduces the computation workload and limits the number the features while selecting the only important ones. Companies like "Zillow.com","magicbricks.com", often have a large dataset of houses whose prices they predict using machine learning. One of the techniques they use is regression [3], deep learning [3] to learn the nature of models from the previous results (houses which were sold off previously which are used as training data).In this paper we have defined linear model data using only one feature, multivariate model, using several features as its input and polynomial model using the input as cubed or squared and hence calculated the root mean squared error (RMS value) for the model.

Sometimes the surrounding conditions of a locality determines what kind of price we can expect for different kind of houses, [4] presented a predictor using nonlinear Support Vector Regression showing relationship between visuals of some cities and non-visuals attributes(crime stats, population density,

etc.), their research also presented few prototype application based on the same predictor. Other researchers showed a different mechanism to predict house prices by focusing on Multiple Listings Data [7], it showed how different correlations can be referred while we estimate regression coefficient for predicting the price of a house. They determined the coefficient by using the concepts of maximum likelihood and discussed kirging, a technique that can be used to merge spatial correlation for predictions. Housing prices can also be predicted using semi-parametric regression models or nonparametric model because of their better results as compared to parametric models. Non-parametric models allow the regressions to belong to a class of function, semi-parametric models incorporate the advantages of other models by allowing the function to be linear, convex, etc. based on which function provides the best predictions [8].

2. Related Work

Housing prices are based on several factors. For creating the model we can use several features, features can also be extracted from several sources. One the most notable work I find in feature extraction is that of "City Forensics: Using Visual Elements to Predict Non-Visual City Attributes" [4], which used visual features to predict the housing prices. Clustering has been used in [2] to cluster houses of same features and prices.

3. Methodologies Used

We have used several models and have calculated the root mean squared error for each. Graphs have been plotted for each model. The data-set we have used is a group of houses of King County region in Seattle. The size of the dataset is of 21,000 houses which are divided into training data and testing data in the ratio 80:20.The number of features present in our data is square feet, price, date sold, a number of bedrooms, floors, water-front there or not, floors, year built square feet above, zip code, latitude and longitude. For each of the model, we calculated the root mean square error (RMS value) as proposed in [5].With this we have applied the following methodologies:

3.1 Simple Linear Regression:

In this methodology, we have used just 1 feature- square feet of the house versus the price of the house to train our model. The basic equation of the fit of our model looks like:

$$F(x) = w0+w1x$$

Where: x=square feet, F(x) =price w0=intercept term, w1=coefficient of square feet being simple linear regression we have just used one feature which is square feet.
The sample data set of each of the models is shown in Table 1 below.

Table 1: Sample data set

id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront
7129300520	2014-10-13 00:00:00+00:00	221900.0	3.0	1.0	1180.0	5650	1	0
6414100192	2014-12-09 00:00:00+00:00	538000.0	3.0	2.25	2570.0	7242	2	0
5631500400	2015-02-25 00:00:00+00:00	180000.0	2.0	1.0	770.0	10000	1	0
2487200875	2014-12-09 00:00:00+00:00	604000.0	4.0	3.0	1960.0	5000	1	0
1954400510	2015-02-18 00:00:00+00:00	510000.0	3.0	2.0	1680.0	8080	1	0
7237550310	2014-05-12 00:00:00+00:00	1225000.0	4.0	4.5	5420.0	101930	1	0
1321400060	2014-06-27 00:00:00+00:00	257500.0	3.0	2.25	1715.0	6819	2	0
2008000270	2015-01-15 00:00:00+00:00	291850.0	3.0	1.5	1060.0	9711	1	0
2414600126	2015-04-15 00:00:00+00:00	229500.0	3.0	1.0	1780.0	7470	1	0
3793500160	2015-03-12 00:00:00+00:00	323000.0	3.0	2.5	1890.0	6560	2	0

view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat
0	3	7	1180	0	1955	0	98178	47.51123398
0	3	7	2170	400	1951	1991	98125	47.72102274
0	3	6	770	0	1933	0	98028	47.73792661
0	5	7	1050	910	1965	0	98136	47.52082
0	3	8	1680	0	1987	0	98074	47.61681228

For the above model, the residual sum of squares is calculated as 1.20191835632 e+15.

Since this a very high error we will reduce it by bringing more models.

3.2 Multivariate regression models:
In multivariate models instead of 1 feature, we use several features as proposed by [6].

The below models were trained using the given features:

Model1 = Regression trained using [square feet,bedrooms, bathrooms]

Table 2. Coefficients obtained by training the above model:

name	index	value	stderr
(intercept)	None	87910.0724924	7873.3381434
sqft_living	None	315.403440552	3.45570032585
bedrooms	None	-65080.2155528	2717.45685442
bathrooms	None	6944.02019265	3923.11493144

[4 rows x 4 columns]

Model 2 = Regression trained using [square living, bedrooms, bathrooms, latitude, longitude]

Coefficients obtained by training the above model is given in table 3.

Table 3. Coefficients obtained by regression.

name	index	value	stderr
(intercept)	None	-56140675.7444	1649985.42028
sqft_living	None	310.263325778	3.18882960408
bedrooms	None	-59577.1160682	2487.27977322
bathrooms	None	13811.8405418	3593.54213297
lat	None	629865.789485	13120.7100323
long	None	-214790.285186	13284.2851607

[6 rows x 4 columns]

Model 3=Regression trained using [sqft_living, bedrooms, bathrooms, latitude, longitude, bed_bath_rooms]

Bed_bath_rooms = number of bedrooms * number of bathrooms.

This is done because any trained model is biased towards bedrooms,the bathrooms usually are drowned out or neglected because heavy weight assigned to the coefficient of square feet and bedrooms as shown in table 4.

Table 4. Coefficients trained by model 3

name	index	value	stderr
(intercept)	None	-54410676.1152	1650405.16541
sqft_living	None	304.449298057	3.20217535637
bedrooms	None	-116366.043231	4805.54966546
bathrooms	None	-77972.3305135	7565.05991091
lat	None	625433.834953	13058.3530972
long	None	-203958.60296	13268.1283711
bed_bath_rooms	None	26961.6249092	1956.36561555

[7 rows x 4 columns]

Residual sum of squares error for each of the above models:

Model 1: $ 16,545,470

Model 2: $ 31,166,139

Model 3: $ 31,009,548
It shows that model 1 is doing well as our fit according to multivariate fit in our model

Table 5. Coefficients obtained by training the above model:

name	index	value	stderr
(intercept)	None	-56140675.7444	1649985.42028
sqft_living	None	310.263325778	3.18882960408
bedrooms	None	-59577.1160682	2487.27977322
bathrooms	None	13811.8405418	3593.54213297
lat	None	629865.789485	13120.7100323
long	None	-214790.285186	13284.2851607

[6 rows x 4 columns]

Model 3=Regression trained using [sqft_living, bedrooms, bathrooms, latitude, longitude, bed_bath_rooms]

Bed_bath_rooms=number of bedrooms * a number of bathrooms.

This is done because any trained model is biased towards bedrooms,the bathrooms usually are drowned out or neglected because heavyweight assigned to the coefficient of square feet and bedrooms.

Residual sum of squares error for each of the above models:
Model 1: $ 16,545,470
Model 2: $ 31,166,139
Model 3: $ 31,009,548
It shows that model 1 is doing well as our fit according to multivariate fit in our model

3.3 Polynomial Regression

To demonstrate polynomial regression we use features which have been multiplied to several powers (up to 15) in our test.
The linear fit looks like below: X axis – square feet, Y axis – price (in $)

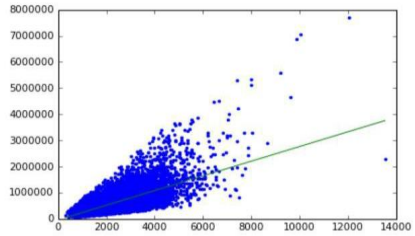

Figure 1. Polynomial regression

In Figure 1, the green line indicates our fit and the blue dots are our data points

For power = 2 or where we had used feature = square_feet * square_feet we get the fit like below:

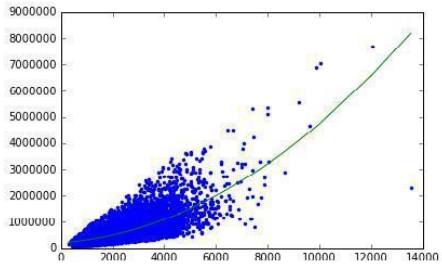

Figure 2. The fit function for power 2.

From the figure 2, we observe that as we increase the feature complexity, the model is tending towards overfitting.

For power =15 we get the fit:

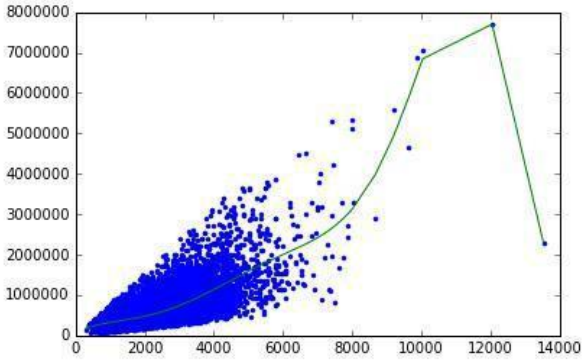

Figure 3.power 15 model

The power 15 model is extremely overfit as shown in figure 3, where the curve in our model touches the data points further away. For each of the models with features (power 1 to power 15), we get the following residual sum of squares as shown in the table below.

Table 6. Residual sum of squares

11359157.9835
11188838.9573
11222275.0371
11275979.7364
11271101.5527
11203987.5869
11162830.0001
11156117.8075
11159090.4547
11162568.185
11164306.2366
11164502.1527
11163879.6932
11163000.4206
11162163.935

4. Conclusion

We have defined several models with various features and various model complexities. We realized we need to use a mix of these models a linear model gives a high bias (under fit) whereas a high model complexity based model gives a high variance (overfit).Data Scientist tends to overfit their models which can be reduced by ridge regression and LASSO. Housing price prediction also uses K Nearest Neighbour search which tends to cluster various houses of the same genre together to cluster houses of same price and genre [2].

References

[1] Yeonjong Shin and Dong bin Xiu 2016, Near optimal sampling strategy for least squares polynomial regression.*Journal of Computational Physics* vol. no.326 pp 931-946

[2] Yang Li, Quan Pan ,Tao Yang and Lantian Guee 2016 Reasonable Price Recommendation on Airbnb Using Multi-Scale Clustering *Proceedings of the 35th Chinese Control Conference* pp27-29

[3] Mansural Bhuiyan and Mohammad Al Hasan (2016) "Waiting to be Sold: Prediction of Time-Dependent House Selling Probability" IEEE *International Conference on Data Science and Advanced Analytics* pp468-477

[4] Sean Arietta, Alexa, A Efros , Ravi Ramamoorthi, and Maneesh Agarwal City Forensics: Using attributes to predict non visual Attributes visual elements for *prediction IEEE Transactions on Visualization and Computer Graphics* vol. 20(12) pp 2624-2633

[5] Samuel Hameau, Tri Kaviniawan Wijaya, Matteo Vasirani, Karl Aberer 2013 Electricity load Forecasting for Residential customer : Exploiting Aggregation and Correlation between Households profiles. *Sustainable Internet and ICT for Sustainability IEEE*

[6] Byeonghwa Park and Jae Kwon Bae 2015 Using machine learning algorithms for housing Price prediction : The Case of Fair fax country Virginia housing data *Expert Systems with Applications* vol. 42(6) pp 2928-2934

[7]Robin A Dubin 1998 Predicting Housing Prices using Multiple Listings Data *Journal of Real Estate Finance and Economics* vol.17(1) pp 35–59
[8] JooyongShim ,OkmyunBin and Changha Hwang 2014 Semi-parametrics partial effects kernel minimum squared error model for prediciting housing sales price *Neuro computing* vol. 124, pp 81-88