*Price Prediction of Diamonds

Xiran Wen, Qiqi Xu, Zirui Su, Jiayi Fang
The Chinese University of Hong Kong, Shenzhen

Email: 119010333@link.cuhk.edu.cn

Abstract

The experiment aimed at price prediction based on diamond dataset which contains 53940 rows of information. The model is constructed based on linear regression model with the lowest estimated test error among all methods including tree and nonlinear models. The experimental results show that the mean square error for the training dataset and validation dataset are 592182.6 and 603833.2 respectively, and the R^2 reached 98%. The test MSE is 631947. The proposed model can well predict diamond prices.

Keywords: Price prediction, Linear regression, Mean square error, Cross Validation

1. Introduction

The classic ***Diamonds*** dataset contains the prices and 10 attributes of 53940 diamonds. It's a typical dataset of linear regression to analyze and visualize data.[1]

The price of a commodity is determined by its value, so as diamonds. Precisely estimating the price of diamonds could help merchants and customers make transactions properly. As the price of diamonds is significantly affected by the data of diamonds' attributes, we train a model on price of diamonds. The internal relationship of the dataset will be converted into the model with predictors. When we input the predictors, the model will automatically output a predicted price. And the goal of the model is to precisely predict the price according to their weights, color, quality of cut, measurement of clarity, width, length, depth, total depth percentage and width of top of diamond relative to widest point.

2. Research Material and Methods

2.1. *Data description*

The dataset contains 53940 rows and 10 columns. The price column is treated as the output y. Other variables and their meanings are as follows: (1)price: price in US dollars; (2)carat: weight of the diamond; (3)cut: quality of the cut (Fair, Good, Very Good, Premium, Ideal); (4)color: diamond color, from J (worst) to D (best); (5)clarity: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)); (6)x: length in mm; (7)y: width in mm; (8)z: depth in mm; (9)depth: total depth percentage = z / mean(x, y) = 2 * z / (x + y); (10)table: width of top of diamond relative to widest point.

2.2. *Methods*

In this paper, several regression analysis methods are used to determine the main factor that influence the price and study the relationship between diamonds' characteristics and prices. Linear regression model is used when the relationship between input and output is linear.[2]

*

Several non-linear regression models including polynomial, regression spline, natural spline regression and tree, are used to fit the data. Each model is estimated in two ways.[2] First, for the validation set approach, calculate training MSE based on training data set and estimated test MSE based on validation set. Second, for 10-fold cross validation approach, the estimated test MSE is calculated based on the training dataset only.[2]

Since, the training error of the model is approximately equivalent to its estimated test error, according to the Variance-Bias Trade off, which implies that the model has not yet over-fitted the data. At the same time, training error of the model is small enough when most of data is well explained which is the model with high R^2. However, smallest training error will lead to poor performance on test data that is over-fit. Therefore, the model with the high R^2 and the training MSE similar to estimated test MSE is chosen to be the final combination of predictors.

The final model is fitted based on training and validation dataset and test MSE is obtained by test dataset.

3. Experiments and Results

3.1. *Data processing*

The dataset is checked whether there is any invalid data. Rows contain NA are removed. The data is randomly split into three parts, 50% for training data, 20% for validation data and 30% for test data.3 category variables, cut, color and clarity, are transformed into 18 dummy variables using R.

3.2. *Data visualization*

Weight is generally considered as the critical factor of diamond price. So let x axis be the carat and y axis be price and randomly draw 500 data to plot.

Diamonds in different colors, cuts and clarity indicate linear relationship between price and carat as well. (Fig.1) Therefore, our first assumption of the model is a linear model with all 10 predictors.

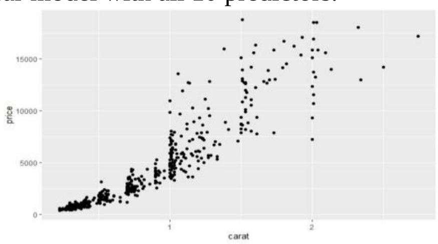


Fig.1 The relationship between the price and carat

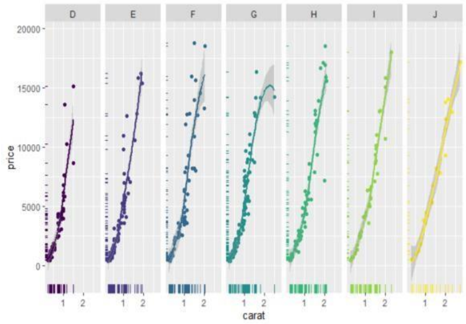


Fig.2 The price vs carat relationship based on diamond colors

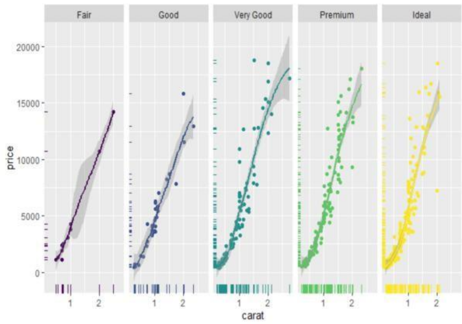


Fig.3 The price vs carat relationship based on cut quality

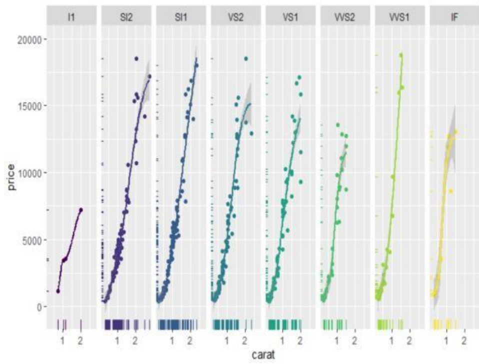


Fig.4 The price vs carat relationship based on clarity

All figures above show a nearly linear relationship. Then assumption is made that it is a linear model.

3.3. *Linear models*

From the data visualization, linear model is a good choice. The initial model is composed by all 10 predictors with 1301920 training MSE and estimated test MSE 1219288. Nearly 92% training data can be explained by this model which is not good enough.

After checking GVIF^(1/(2*Df)), it is found that the dimension of diamonds and their weight have value over 2 which indicates collinearity between them.[3] Through correlation map, it can also be seen that predictor selection is needed. Different methods are used to select significant variables.

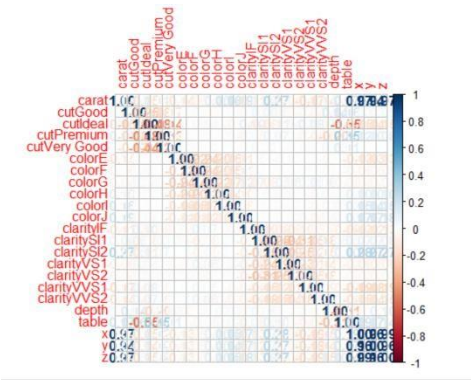
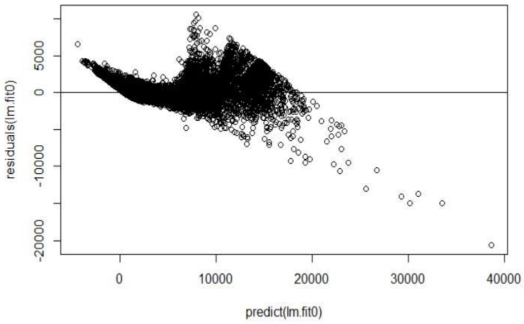


Fig.5 Correlation between single predictors in initial model

The residual plot of training data shows non-linearity in data as Fig.6 shown.



Shrinkage methods, Lasso and Ridge regression does not work well since no parameter shrink small enough to ignore the corresponding predictors.[4] Subset methods, best subset selection, forward and backward stepwise selection, are used with selection criteria of CP, BIC and R^2 to choose the best fit subset of predictors to remain in the model. And under the inspiration of data visualization plot, intercepts are removed since the line nearly cross the original point. It leads to the model without y and z predictors and intercept which gives 1302139 of training MSE and 1216096 of estimated test MSE. However, the R^2 increases from 92% to 95.85% which improves dramatically. This model is called Model21.

Since MSEs are not improved, interaction terms are added separately into the model of carat or x with other predictors. The p-value of the added interaction terms and the R^2 for the models are shown as follows to fix the non-linear problem shown in residual plot. Interaction terms with above 96% and a p-value smaller than 0.05 are added to the current model.

Table 1 R2 above 96% and p-value for interaction term

Interaction term added in the current model	R^2	p-value
carat:cut	0. 9611	< 2e-16
carat:color	0. 9639	Some levels < 2e-16
carat:clarity	0. 9717	< 2e-16
x:cut	0. 9605	< 2e-16
x:color	0. 9633	Some levels < 2e-16
x:clarity	0. 9702	< 2e-16

Since carat and x has a high correlation, this interaction term is also added to the Model21 as well as above terms. This model is called Model17. It decreases the MSEs by half that is 595552.7 for training data, 593042.4 for validation data and 610405.1 for C.V. estimated test MSE. R^2 reaches 98.1%. Since the training MSE and estimated test MSE are still similar to each other, it is not overfit so that qualifies the model selection criteria.

3.4. *Nonlinear models*

The residual plot of Model17 is shown as follow.

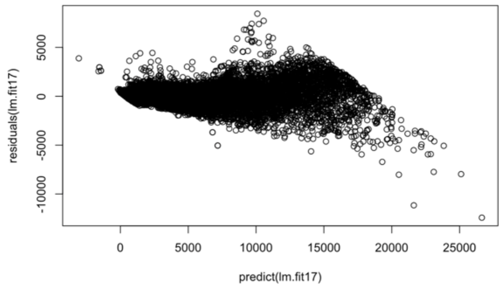


Fig.7 Residual plot of Model17

It still shows some outliers and non-linear relationship since the points do not form a horizontal band around zero. Therefore, nonlinear models are constructed.

The pruned tree with 6 terminal nodes has the training MSE of 1948964 and estimated test MSE of 1948964.[5]

Polynomial regression is applied with degrees from 2 to 5 based on Model21. The training MSE and the estimated test MSE is shown as Table 2.

Table 2 Polynomial regression with different degree

Degree	Training MSE	Estimated test MSE

2	1162500	1221956
3	1087814	1174913
4	1063071	1134305
5	1072774	1550505

According to the results, the lowest estimated test MSE is still higher than the linear model. Thus, the polynomial regression is not a good fit.

Spline regressions are applied with degree of freedom from 4 to 10 based on Model21 and spline regression with degree 10 gives the lowest training MSE of 1062536 and estimated test MSE of 1154814.[6]

Natural spline regressions with degree of freedom from 2 to 9 based on Model 21 are applied and the one with 9 degree gives the lowest training MSE of 1049669 and estimated test MSE of 1116726.

To wrap up, the nonlinear models are not suitable for the dataset as the estimated test MSE is too high, implying the low predictive accuracy of the nonlinear models.

4. Result and Discussion

The final model is Model17 we get in the linear part. Then we use tr_va data to fit this model and calculate the true test MSE on the test data. The result is encouraging as the R^2 on the test data achieves 0.9811 and difference between test MSE of 603833.2 and training MSE of 592182.6 is acceptable.

The final model is Price = carat + cut + color + clarity + depth + table + x + carat:cut + carat:color + carat:clarity +x:cut +x:color +x:clarity + carat:x with parameters shown in the table.

The model indicates that the price is affected by carat, cut, color, clarity, depth, table and x. Considering both the single term and the interaction term for each attribute, larger carat, better cut, better clarity, better color, smaller depth, smaller table and smaller x can lead to higher price, which conforms to the common sense.

However, from data visualization plot, as the weight increase, the data points are more scattered which indicates a quadratic formation. But polynomial regression did not give a satisfying result. The reason might be that only a small part of the dataset shows the scatter pattern compared with the whole dataset of more than fifty thousand diamonds. Most diamonds are not so heavy and therefore the data will mostly distributed on range of smaller carat. The scatter plot on the small-carat section indicates an obviously linear relationship. Thus, the linear model fits better than the nonlinear model on the whole dataset.

Table 3 Parameters of the final model

Predictor	Para.	Predictor	Para.
carat	-4039.014	cutFair	6804.717
cutGood	6487.471	cutIdeal	7428.431
Cutpremiun	8080.426	cutVery Good	8239.953
colorE	1263.565	colorF	671.327
colorG	575.086	colorH	1249.292
colorI	1417.310	colorJ	2592.538
clarityIF	-1899.250	claritySI1	5131.024
claritySI2	3302.777	clarityVS1	3124.533
clarityVS2	3831.450	clarityVVS1	3034.499
clarityVVS2	2599.145	depth	-66.094
table	-27.084	x	-27.281
carat:cutGood	391.478	carat:cutIdeal	1647.206
carat: cutpremiun	1378.577	carat: cutVery Good	1895.888
carat: colorE	596.943	carat: colorF	163.011
carat: colorG	-751.679	carat: colorH	-1428.383
carat: colorI	-2090.482	carat: colorJ	-3028.556
carat: clarityIF	9573.599	carat: claritySI1	9376.079
carat: claritySI2	6665.108	carat: clarityVS1	10051.921
carat: clarityVS2	9771.941	carat: clarityVVS1	12065.551
carat: clarityVVS2	11029.471	cutGood:x	30.490
cutIdeal:x	-252.080	cutpremiun:x	-351.244
cutVery Good:x	-457.012	colorE:x	-334.640
colorF:x	-193.257	colorG:x	-106.895
colorH:x	-200.700	colorI:x	-213.227
colorJ:x	-352.775	clarityIF:x	-338.959
claritySI1:x	-1981.432	claritySI2:x	-1393.376
clarityVS1:x	-1571.735	clarityVS2:x	-1711.464
clarityVVS1:x	-1654.501	clarityVVS2:x	-1495.811
carat:x	1001.630		

5. References

1. Shivam, A. (2017). Kaggle, "Diamonds".
2. James, G., Written, D., Hastie, T. & Tibshirani, R. (2021). An introduction to statistical learning (2nd ed.), New York, NY: Springer.
3. VIF in car package of R language, "Variance Inflation Factors".

4. Glmnet of R language, "Cross-validation for glmnet".
5. Tree of R language, "cv.tree", "prune.tree".
6. Splines2 of R language, "Regression spline functions and classes".

Authors Introduction

Miss. Xiran Wen

She is an undergraduate student majored in Science Data in The Chinese University of Hong Kong (Shenzhen). She is interested in data analysis based on statistics.

Mr. Zirui Su

He is an undergraduate student majored in Financial Engineering in The Chinese University of Hong Kong (Shenzhen). He is interested in financial analysis.

Miss. Qiqi Xu

She is an undergraduate student majored in Science Data in The Chinese University of Hong Kong (Shenzhen). She is interested in data analysis based on statistics.

Miss. Jiayi Fang

She is an undergraduate student majored in Financial Statistics in The Chinese University of Hong Kong (Shenzhen). She is interested in Financial analysis.