Construction and analysis of prediction model for sales of child car seats

Yuhang Yang [1,*], Xin Ni [1], Hongda Yan [1], Haoyuan Li [1], Mengsha Hong [2]

[1] Ocean University of China, Qingdao, China

[2.]Jiangxi University of Finance and Economics, Nanchang, China

* Corresponding Author Email: yyh2469@stu.ouc.edu.cn

Abstract. Based on the Carseats data set in the ISLR package, this paper constructs a regression model of the data set and predicts some variables. In addition, it also includes the establishment of regression trees and other models for variables. The final task requirement is to construct the classification model of the data, divide the classification model into the training set and the test set, and construct the linear discriminant analysis, logistic regression, random forest and so on with another variable. By making some models, the prediction accuracy on the test set is relatively high, and the evaluation effect of the model is also very good, which can provide a sufficient theoretical basis for the prediction of the sales of children's car seats.

Keywords: multiple regression model, regression tree, classification model Introduction.

1. Introduction

Nowadays, with the continuous development of transportation and the high popularity of private cars, most families choose cars as the means of transportation. After the advent of child car seats and the national requirements, child seats have been widely used and saved many lives. A child car seat is specially designed for children of different weights (or ages). Children installed in the car can effectively improve the safety of children in the car seat. It can be fixed to a motor vehicle with a child safety protection system consisting of a seat belt assembly or flexible parts. In the case of a car crash or sudden deceleration, the impact pressure on the child can be reduced, and the movement of the child's body can be restricted to reduce injury to them.[1-3]

For a child seat company, under the premise of ensuring quality, the most important thing is to be able to understand the sales volume. However, the factors that affect sales are many and complex. Therefore, controlling and forecasting sales is crucial to the company's profitability.[4-5]

2. DATA Progressing

This paper uses the data set of Carseats in the ISLR package of the RStudio program.

This data contains 11 variable names shown in TABLE 1, including Sales, CompPrice, Income, Advertising, Population, Price, ShelveLoc, Age, Education, Urban, and the US, a total of 400 groups of data. There are two types of data: numeric and factor. This paper will use these 400 sets of data for forecasting, analysis, and modeling Of sales volume.

We need to use the three variables price, urban and the US to fit the multiple regression model for Sales and make predictions. At the same time, to better establish the model, we need to conduct statistical analysis and hypothesis tests on the fitting results.

Table 1. Table Type Styles

Sales(x_1)	Unit sales (in thousands) at each location
CompPrice(x_2)	Price charged by competitor at each location
Income(x_3)	Community income level (in thousands of dollars
Advertising(x_4)	Local advertising budget for company at each location (in thousands of dollars
Population(x_5)	Population size in region (in thousands)
Price(x_6)	Price company charges for car seats
ShelveLoc(x_7)	A factor with levels Bad, Good and Medium indicating the quality of the shelving location
Age(x_8)	Average age of the local population
Education(x_9)	Education level at each location
Urban(x_{10})	A factor with levels No and Yes to indicate whether the store is in an urban or rural location
US(x_{11})	A factor with levels No and Yes to indicate whether the store is in the US or not

3. Construct linear regression model

I selected the multivariate number set and defined the mathematical model. Then, I estimated the parameters, carried out significance test, residual analysis and outlier detection on the estimated parameters, and finally determined the regression model for model prediction. The following is a presentation of the results shown in TABLE 2.

Table 2. Each coefficient after lm regression

	Estimate	Std.Error	T value	Pr(> |t|)
(Intercept)	13.043469	0.651012	20.036	< 2e-16***
Price	-0.054459	0.005242	-10.389	< 2e-16***
UrbanYes	-0.021916	0.271650	-0.081	0.936
USYes	1.200573	0.259042	4.635	4.86e-06***

From the results, it is observed that:

F-statistic=41.52,p-value<2.2e-16

The influence of each variable on Sales:

Price: The price variable passes the significance test, and the correlation coefficient is negative, indicating that the higher the price, the lower the sales volume

Urban Yes: The variable did not pass the significance test, indicating that whether the city has no significant impact on the sales volume

US Yes: The variable passed the significance test, and the correlation coefficient was 1.20, which indicated that the sales volume in the United States would increase by 1201 units.

4. Construct regression tree model

4.1. Regression tree

General steps to build a regression tree:

Firstly, the RPART function of the RPART package is used to train the decision tree model. Formulas and data sets need to be specified, and the model is set as a regression model, that is, Method is set as ANOVA. Finally, the control parameters are mainly some pre-pruning parameters to control the growth of the decision tree. Then output the results of the previous model, namely the initial tree, as well as the complexity related tables and graphs. Based on these tables and figures, an optimal CP value can be determined for the next step of post-pruning. The pruned decision tree, which is the final

decision tree model, is used for subsequent output and prediction. Finally, the output decision tree related graph, including variable importance bar chart, tree chart. Regression tree is shown in Figure 1.

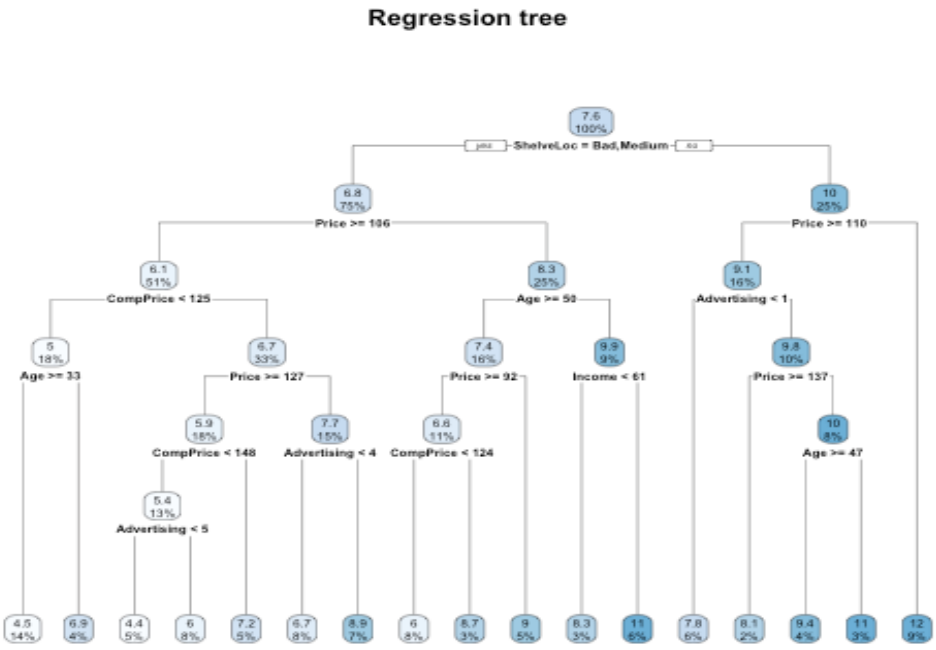


Figure 1. Regression Tree

4.2. Pruning of the regression tree

Next, I want to get the complexity of the tree by cross-validation. Pruning of the regression tree is shown in Figure 2.

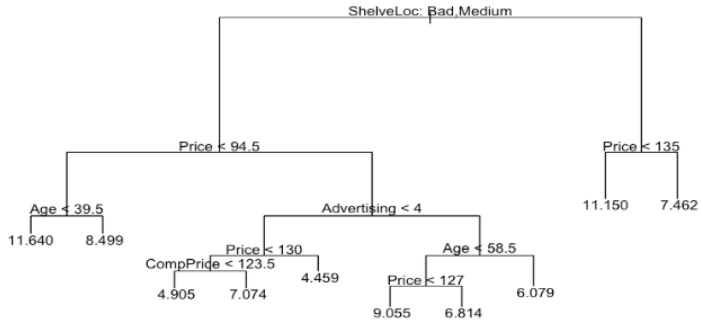


Figure 2. Pruning of the regression tree

At the bottom is a tree diagram for determining the best value. Determining the best value is shown in Figure 3.

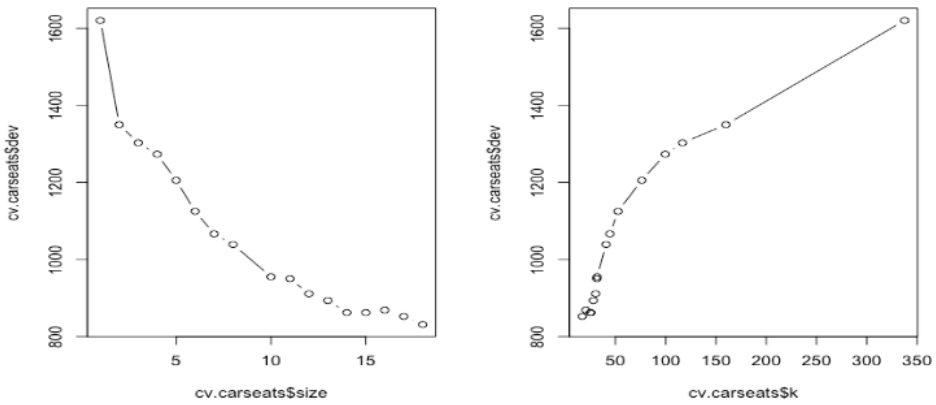


Figure 3. Determining the best value

In the first picture on the left, it is obvious that at the position of the number 9, the curve has a clear turn up and down, so it can be concluded that the best value is 9.

Pruning the tree in this case increases the test MSE to 4.99.

I found that the MSE after pruning was larger, that is, the error was larger.

Then determine the error rate of the test. I used the function to do some data processing and got the following result.

The error rate we tested using the important function is 2.87. The degree of importance of each variable is shown in TABLE 3:

Table 3. The Degree of importance of each variable

Importance()	IncMSE	IncNodePurity
ComPrice	17.40549458	158.95624
Income	1.67508456	112.47131
Advertising	10.21128321	104.79976
Population	-2.35224559	80.37306
Price	45.58923658	444.81480
ShelveLoc	41.37394274	335.45870
Age	16.07638139	170.69457
Education	1.99636706	55.93799
Urban	-0.02934908	11.71216
Us	4.70086680	27.17367

5. Constructing a classification model

In this question, we divide Sales into two parts: high Sales and low Sales, and the classification rule is whether it is less than or equal to 8. The classified Sales is taken as the new variable High. We divided the 400 sets of data into two parts, using random seeds to make 300 sets as the training set and 100 sets as the test set. We use linear discriminant analysis, logistic regression, classification tree, random forest and other classification models for training, and evaluate the effect of the model on the test set. In order to better construct the model, we need to make High into "Yes" and "No" type, while in logistic regression, we need to make High into "0" and "1" factor type.

5.1. Linear discriminant analysis

I take High as the dependent variable and other variables (excluding Sales) as the independent variable to obtain the linear discriminant model. At the same time, the model was put into the test set for prediction, and the confusion matrix with the original data was obtained as TABLE 4:

Table 4. confusion matrix

	newHigh	newHigh
High	No	Yes
No	47	3
Yes	16	34

5.2. logistic regression

We establish and obtain various statistics . The summary content can be summarized as follows through data analysis. The following table lists the coefficient estimates of the logistic regression model established with all predictor variables for high sales probability, including Comprice, Income and other variables. It can be seen that the P value is small, which means that these variables have an impact on the sales probability. Parameters of different variables is shown in TABLE 5:

Table 5. Parameters of different variables

Coefficients	Estimate	Std.Error	Z value	Pr(> |z|)
(Intercept)	-4.8364518	2.8489203	-1.698	0.089575
ComPrice	0.1703408	0.0268506	6.344	2.24e-10***
Income	0.0343774	0.0089050	3.860	0.000113***
Advertising	0.2992754	0.0580793	5.153	2.57e-07***
Population	-0.0002601	0.0016098	-0.162	0.871624
Price	-0.1709624	0.0232358	-7.358	1.87e-13***
ShelveLocGood	8.9341272	1.2584168	7.099	1.25e-12***
ShelveLocMedium	3.8568045	0.7807333	4.940	7.81e-07***
Age	-0.0836464	0.0171480	-4.878	1.07e-06***
Education	-0.0190338	0.0898153	-0.212	0.832168
UrbanYes	-0.7186646	0.5564617	-1.291	0.196534
UsYes	-0.8627300	0.6683812	-1.291	0.196782

5.3. Classification tree

The classification tree hopes to find the optimal branch structure through training, and the key to the construction of classification tree is tree generation

Long and pruning process. In the growing process of classification tree, the algorithm hopes to select the best feature continuously .Rows recurse and split the training set until there are no suitable features or all subsets are basically divided correctly Make the classification tree as large as possible. In order to prevent overfitting, it is also necessary to base the constructed classification tree on the loss function. Small principles for bottom-up pruning, select the best subtree. You can see decision tree in figure 4.

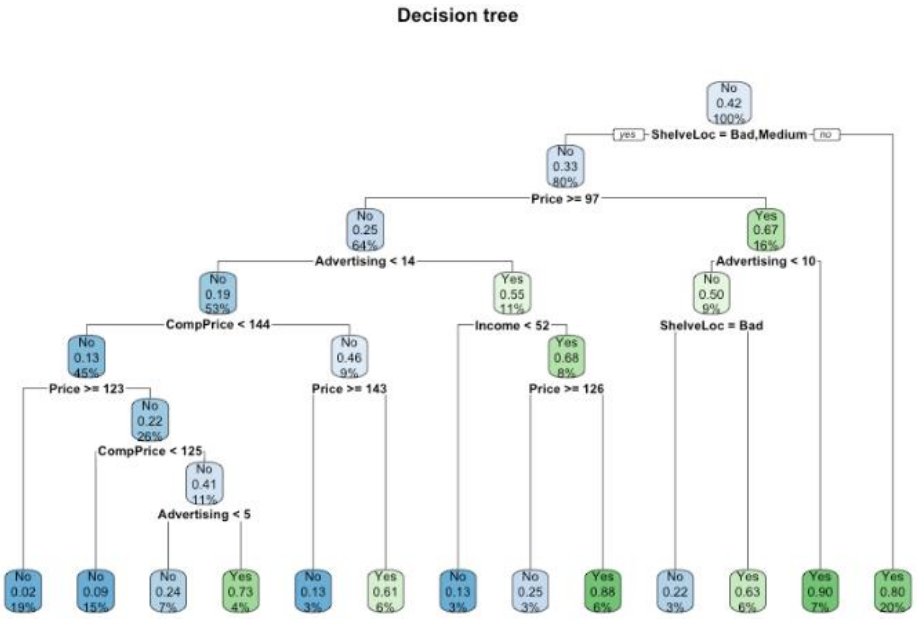


Figure 4. Decision tree

The precision of the decision tree in the test set is 0.770000.

5.4. Randomforest

Random Forest is ensemble learning based on decision trees. The idea is that if an input variable significantly affects. The prediction effect of the output variable will lead to the overall prediction accuracy after adding random noise to the out-of-pocket observed independent variables. A significant drop. The characteristic of random forest method is that the selection of training samples and features is random. Assuming that the sample. This set T has n samples and P indices.

Table 6. n samples and P indices

	0	1	Class. error
0	152	21	0.1213873
1	45	82	0.3543307

You can read from the graph above that the error rate is 22%.

The error rate is plotted against different numbers in the random forest is shown in Figure 5：

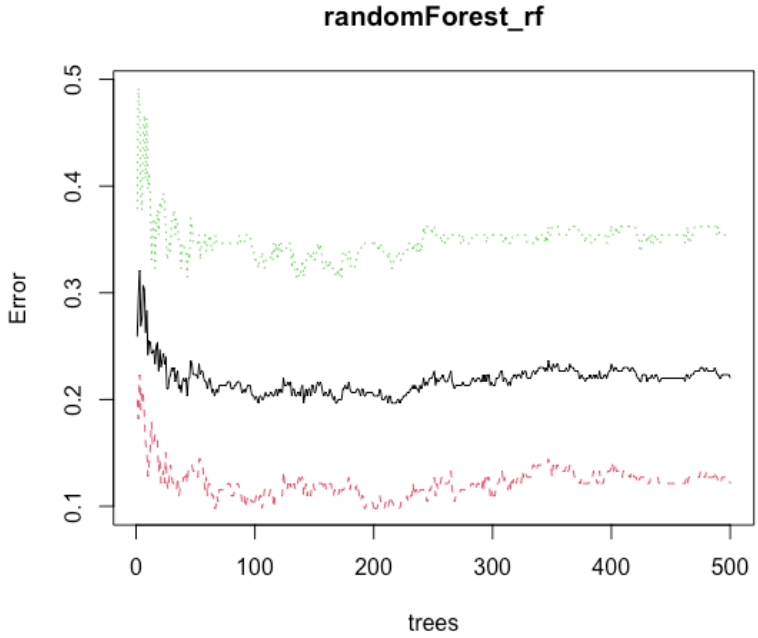


Figure 5. Random Forest

Draw a rating chart of importance shown in Figure 6:



Figure 6. The ranking of important values

6. Conlusion

The prediction accuracy, Kappa value and AUC value of each model were comprehensively compared and analyzed to obtain the best training effect

The models are Fisher discriminant model and SVM model, and the prediction accuracy of the two models on the test set is up to 100% and 98.28%, the evaluation effect of the model is also very good. In the above machine learning methods, for practical response Fisher discriminant model and SVM model are more likely to be used in the classification of child seats Forecast, for the actual child seat market sales activities to provide theoretical guidance.

From the perspective of algorithm, KNN classification algorithm is relatively simple; Fisher discriminant analysis was done by casting. The method of shadow can reduce the dimension and simplify the problem, and the model has good generalization. The classification results of the classification tree are visualized. The latter is intuitive and simple, but the generalization is not good in this paper. Random Forest is an integrated science based on decision trees

After selecting the optimal index number, the number of decision trees is further determined. SVM model determines classification efficiency by support vector. The support vector can be given to effectively deal with the data in the high dimensional space, and the generalization is good for the class. Non-overlapping, difficult to classify situations can be optimized by changing the weight of the categories.

References

[1] Liu Kaicheng, Guo Rui, Liu Rongchuan, Qiao Honglei. Optimization model of calcareous soil consolidation characteristics based on multiple regression model [J]. Science and Technology and Innovation, 2022(16): 67-70.DOI:10.15913/j. cnki. kjycx.2022.16.021.

[2] Wang Qian, Sun Daqiang. Prediction of Tianjin Health Resource Allocation Based on Multiple Regression-Gray Model [J]. China Chronic Disease Prevention and Control, 2022, 30(08): 592-596. DOI: 10.16386/j.cjpccd. issn.1004-6194.2022.08.007.

[3] Liu Yuyu, Zheng Qiwen, Li Chen. Analysis of Influencing Factors of Arch Dam Horizontal Displacement Based on Mathematical Statistics Multiple Regression Model [J]. Water Resources and Hydropower Express, 2022,43(07):101-105+128.DOI:10.15974/j. cnki. slsdkb.2022.07.017.

[4] Zhang Naifu, Guo Zhongtao. Research on grid infrastructure investment model based on weak learning regression tree [J]. Mechanical Design and Manufacturing Engineering, 2022,51(08):69-72.

[5] Wang Haoran, Xiao Jianhua, Ouyang Maolin, Gao Hongyan, Liu Yuxin, Gao Li, Gao Xiang, Wang Hongbin. Assessment of risk areas of African swine fever in China based on Bayesian cumulative regression tree [J]. Chinese Journal of Agricultural Engineering, 2022, 38 (11): 180-187.