Advances in Economics, Business and Management Research, volume 203
Proceedings of the 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021)

Business Analytics for Used Car Price Prediction with Statistical Models

Yufei Chen[1,†], Chenle Li[2,†], Minglu Xu[3,*,†]

[1]*College of liberal arts, Xi'an University, Xi'an, Shaanxi, 710075, China, 2198971840@qq.com*
[2]*Art & Science, University of Toronto, Nanjing, Jiangsu, 210000, China, joanne.lichenle@gmail.com*
[3]*School of business, Macao university of science and technology, Xu Zhou, Jiang Su, 221000, China, 1176789750@qq.com*
[*]*Corresponding author. Email: guanghua.ren@gecacademy.cn*
[†]*These authors contributed equally*

ABSTRACT
With the development of the used car market, the demand for a more accurate and scientific price prediction model of used cars becomes urgent. This paper uses multiple linear regression，decision tree and random forest to build up the automobile price forecasting model. We use means to cluster cars and find out that some factors like power, kilometers, gearbox have an influence on the price. According to the analysis, we find out that random forest has the best prediction performance，make sure R2 reaches 0.92 will be enough.

Keywords: *Car Price Prediction, Model, Statistics, Business Analytics, Clustering*

1. BIG DATA AND ITS USE

In essence, the value of big data for commercial projects is to organize the circulation of better products or services, solve most people's needs in the fastest and most efficient way, and achieve stable income. In fact, big data has already provided practical applications for the development of business. Here is the summary of the five fundamental values and application scenarios of big data commercial applications.

To begin with the tagging management. Big data can achieve a relatively fine division of users. [1]For example, the current SCRM system can automatically label different groups of people, continuously operate and calibrate user labels, which realizes the enrichment and improvement of each user's portrait, and finally realizes the accurate push and personalized service of the brand to different users.[2]

Next is the ARVR big data commercial advertising scene enhancement and simulation. Deeply integrate "big data" and analytical technologies for business marketing to successfully transform consumer operations and business models. Enabling brands to access more effective information in more interactive ways and store and model user behavior information and transaction information anytime, anywhere [3]. Any transaction process, product usage scenario and consumption behavior can all be managed by data and visualization.

Furthermore is the improvement on the return of sales investment[4]. Improving the return on investment of the whole marketing management and customer acquisition transformation based on the existing operation mode. With the help of big data capabilities, a company or brand customer can perform a comprehensive analysis of information from the cloud, the Internet, and local databases to form a good operational climate for the entire enterprise, ultimately outputting customer conversion and business profits [5].

Then is the interactive customer management. It can be understood as social CRM or interactive CRM management. According to different scenarios of users, basic information and behavioral catches of users are collected, and users are analyzed from different dimensions to comprehensively understand the preferences, habits, consumption tendencies, and consumption-ability of each user. And new customers are made through digital operation to enhance the attention of brand users, improve customer loyalty, and stimulate sustainable consumption of users[6].

Moreover, deliver personalized and accurate information. Accurate information push is now mainly used in information flow advertising and video. In fact,

for brands and enterprises, especially under the current trend of content operation, users do not lack content, but lack content that can quickly meet the needs and meet users' reading habits. For the commercial operation of users, big data can implement individual sample analysis with the characteristics of sub region association algorithm, semantic analysis, tagging, etc., realize directional push with multiple maintenances according to region, interest, crowd preference, etc., and solve the problem of users' selection of content[7].

Here are the basic steps for the business analysis. The first step to solve the problem must be the problem of the data source. Allen usually divides data into two categories. The first type is directly accessible data, usually internal data[8]. Nothing more than from the website background or their home database inside the guide. The second type is external data, which needs to be processed. Then we are going to do the data cleaning. The purpose of cleaning data (screening, clearing, supplementing, and correcting) is to extract and derive valuable and meaningful data from a large number of disordered and difficult to understand data. After cleaning, we save precious and organized data and reduce the obstacles to data analysis. The third step is to do the Data comparison-contrast is the starting point of data analysis. Because if there is no reference, there is no quantitative evaluation standard for the data. Generally speaking, we start from two points for comparative data analysis: Horizontal comparison and vertical comparison. For the horizontal comparison, we compare with the industry average data and the data of competitors[9]. While the vertical comparison compares with the historical data of their own products around the time axis. After the data comparison and anomalies appear, we certainly want to know what caused it. Here we need to use data subdivision, which is usually divided into latitude and then granularity. Classification by time is time latitude, classification by region is region latitude, classification by origin is source latitude, and classification by visited page is visited latitude. What is the granularity? Do you use time and latitude according to day or hour? This is the difference in granularity[10]. The latitude of your origin, the website of your origin, or the URL of your origin are the differences of granularity. By subdividing the latitude with granularity, you can gradually lock the difference value of the contrast to the problem area. It is easier to find out the cause of the problem. Usually, we can analyze the causes of most problems and deduce conclusions through data subdivision. But there are also special cases. Even if we specify the granularity, it cannot deliver a convincing conclusion. At this time, we can go further and find out the cause of the problem through data traceability. The latitude and granularity of the lock are the search criteria, querying the source logs and records involved and then analyzing and reflecting on the user's behavior based on this. There are often amazing discoveries.

1.1. Basics of statistics and data cleaning

Median

Median is the number in the middle of a set of data arranged sequentially.

Standard Deviation

$$\sigma(r) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - r)^2} \tag{1}$$

Skewness

$$S = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{X_i - \mu}{\sigma}\right)^3\right] \tag{2}$$

The process of data cleaning

First,choose one data subsetting and name it.

Choose one data column that needs to be analyzed in the dataset in order to avoid interference.

Second, delete duplicate values.

Third, deal with missing data.

There may be a lack of data values in the original data, so data-free data cells may be in the data set. Data analysis could influence the result. So we are supposed to complete the missing data.

There are four ways:

1.Replenish the data manually when there are less data.

2.Delete missing data directly.

3.Replace the missing data with average.

4.Replace the missing data with statistical values.

Fourth, Consistency treatment

It could happen that some data leads to inconsistent standard or naming rules in a dataset. So we ' re supposed to use the segmentation function to split the data values into inconsistent data columns.

Fifth, sort the cleaned data.

It is the application of functions such as filtering and sorting, ascending and descending one data column in the dataset[11].

Sixth, deal with an outlier.

The value that deviates more than twice the standard deviation from the average in the set of values is defined as an outlier.

1.1.1. Introduction of linear regression

Simple linear regression is a useful way to predict the response variable with a single predictor variable. In practice, however, there is often more than one predictor variable. A better method is to extend the simple linear regression model to include multiple predictor variables directly compared with building a simple linear regression model for each predictor variable separately. To this end, we can give each predictor a separate slope coefficient in a single model. In general, suppose there are p different predictors. Then the form of the multiple linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \tag{3}$$

Where represents the jth predictor variable and represents the association between the jth predictor variable and the response variable. It can be interpreted as the average effect of every additional unit of on Y while holding all other predictors fixed. It is the intercept, value of Y when = 0. is the slope, sensitivity coefficient of the ith factor.

The regression coefficients β_0 , β_1 , β_2 ,…, β_p unknown, so we need to predict.

The parameters of multiple linear regression are estimated by the least square method, ,, ,⋯, is selected to minimize the residual sum of squares(RSS):

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - y_i)^2 \tag{4}$$

The β_0, β_1, β_2,⋯, β_p that can minimize the RSS is the least square estimation of the multiple regression coefficients.

2. METHODS OF PREDICTING THE SECOND-HAND CAR PRICE

2.1. Assess the price of used cars based on neural networks

There exists asymmetric information between buyers and sellers in the second-hand car market. The artificial neural network is able to solve nonlinear problems well and has fault tolerance, generalization and adaptive capacity[12].

2.1.1. Steps:

Make price forecasts based on the basic conditions of models, driving miles, age and other basic information.

Adjust the price of the used cars according to the specific vehicle condition forecast based on the basic price.

2.1.2. Two stages of the research process:

Data from officially certified websites can serve as a basis for prediction and analysis of factors affecting second-hand car market prices. [13].

Modeling by artificial neural network method

2.2. Predict the value preservation rate of used vehicles based on cluster analysis

Aiming at a part of used car models that lack historical data, find a group of similar models according to various characteristics that affect the value preservation rate of second-hand vehicles.

Firstly, make the sample can optimize the classification according to the analysis of layered clustering[14].

After that, use the historical data of other models, which is similar to the second-hand cars with missing historical data in order to establish a multiple linear regression model taken as the preservation prediction model of this part of second-hand cars.

2.2.1. Steps:

Carry out dynamic cluster analysis

Establish a multiple regression prediction models

2.2.2. Analysis method

Stratified clustering: Decompose the collection of data objects hierarchically until certain conditions are achieved.

3. DECISION TREE MODEL AND RANDOM FOREST MODEL

Decision tree is an algorithm to solve classification problems. The decision tree algorithm uses a tree structure and uses layered inference to achieve the final classification. The decision tree is composed of the following elements:

1. Root node: contains the complete set of samples

2. Internal node: corresponding characteristic attribute test

3. Leaf node: represents the result of the decision[15]

The establishment of a decision tree is divided into three steps:

3.1. Feature selection

Feature selection determines which features are used to make judgments. In the training data set, there may be many attributes of each sample, and the effects of different attributes are different. Therefore, the function of feature selection is to filter out the features that are more relevant to the classification results, that is, the features with strong classification ability. Decision trees divide the predictor space (i.e. all the possible values for for X1,X2, ...,Xp) into distinct regions, say R1, R2, ..., Rk . For every xi that falls in a particular region (say Rj) we make the same prediction[16].

3.2. tree growing

The goal is to find regions R1, ..., Rk that minimize the RSS given by,

$$\sum_{j=1}^{k}\sum_{i\in R_j}(y_i - y\text{^}R_j)^2 \qquad (5)$$

where $\hat{y}_{R_j}$ is the mean response for the training observations within the jth region.

Given the predictor Xj and the cutpoint s, for any j and s, the predictor space is split into the following two regions[17]

$$R_1(j,s) = \{X|X_j < s\}, R_2(j,s) = \{X|X_j \geq s\} \qquad (6)$$

and we seek the value of j and s that minimize the equation

$$\sum_{i:x_i\in R_1(j,s)}(y_i - y\text{^}R_1)^2 + \sum_{i:x_i\in R_2(j,s)}(y_i - y\text{^}R_2)^2 \qquad (7)$$

We stop until our regions have too few observations to continue.

3.3. tree pruning

Rather than considering every possible subtree, we consider a sequence of trees indexed by a nonnegative tuning parameter α. For each value of there corresponds a subtree T ⊂ T0， such that

$$\sum_{i=1}^{|T|}\sum_{i:x_i\in R_j}(y_i - y\text{^}R_j)^2 + \alpha|T| \qquad (8)$$

is as small as possible. Here |T| indicates the number of terminal nodes of the tree T.

The main purpose of pruning is to combat "overfitting" and reduce the risk of overfitting by actively removing some branches.

Random Forest is an improvement over bagged trees. As in bagging, we build a number of decision trees on bootstrapped training samples. When building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. A fresh sample of m predictors is taken at each split, and typically we choose m ≈ √p.

R^2 is coefficient of determination.

$$R^2 = \frac{SSR}{SST} = \frac{Explained\ Variation}{Total\ Variation} \qquad (9)$$

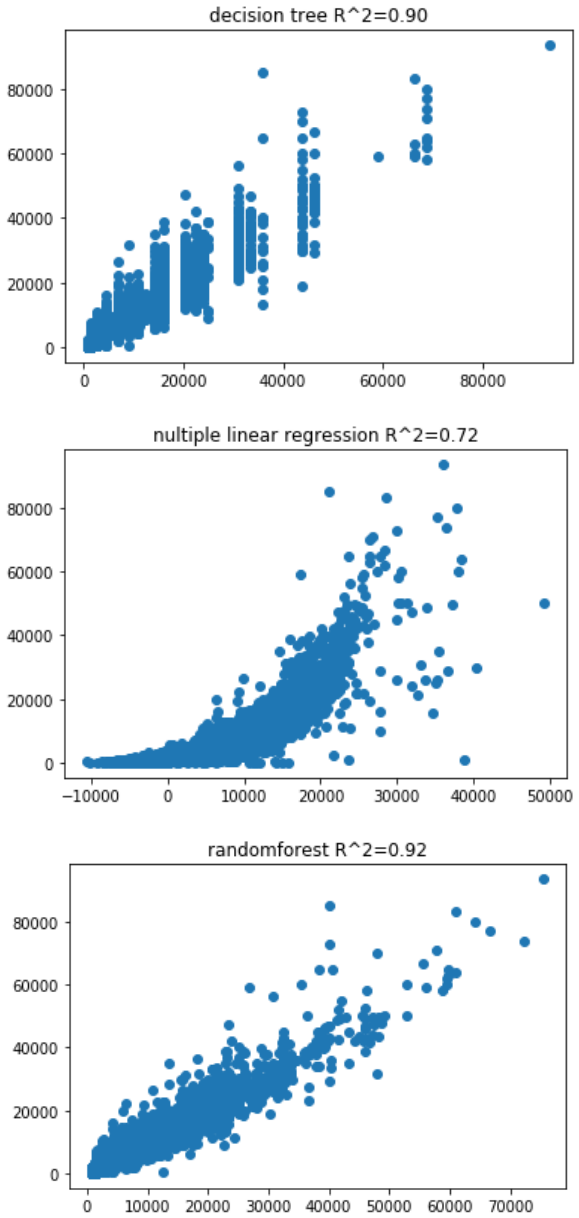
It’s the ratio pf variation explained by the regression. IT measures the prediction power of the fitted model. The larger R^2, the better.

The Pearson correlation coefficient is used to measure the degree of correlation between two variables X and Y. The Pearson correlation coefficient varies from -1 to 1. A coefficient value of 1 means X and Y can be well described by a linear equation, all data points fall on a very good straight on, and Y increases as X increases. The value of the coefficient −1 means that all data points fall on a straight line, and Y decreases as X increases. The value of the coefficient is 0 means that there is no linear relationship between the two variables.

More generally, we find that if and only if X i and Y i are on the same side of their respective mean values, then the value of (X i − X)(Y i − Y) is positive. That is, if the X- I and the Y I simultaneously tend to be greater than, less than, or at the same time tending to their respective mean value, the correlation coefficient is positive. If the X- I and the Y I tend to fall on the opposite side of their average, then the correlation coefficient is negative[18].

The Spearman correlation coefficient is a parameter-free (independent of distribution) test method used to measure the strength of the relationship between variables. In the absence of repeated data[19], if one variable is a strictly monotonic function of another variable, the Spearman rank correlation coefficient is +1 or -1, which means that the variable is completely Spearman rank correlation. Note the difference between this and Pearson correlation[20]. Only when the two variables have a linear relationship, the Pearson correlation coefficient is +1 or -1.

4. SCATTER DIAGRAMS OF PREDICTED PRICE AND REAL PRICE



These three scatter graph shows the actual price and predict price. Adjusted tested by random forest , decision tree and multiple linear regression are 0.92, 0.90, 0.72 respectively. Thus, random forecast is the most accurate method. The difference between predicted price and actual price is caused by non-linear exchange.

We clustered six types of cars using K-means clustering and then used PCA to visualize it. We can see from the graph that these six kinds of clustering are equally distributed.

5. CONCLUSION

We use multiple linear regression，decision tree and randomforest to build up automobile price forecasting model.We use kmeans to cluster cars and find out that there are some factors like power ,kilometers ,gear box have influence on the price.According to the analysis, we find out that randomforest has the best prediction performance，make sure R2 reaches 0.92 will be enough.

REFERENCES

[1] E.M. Clarke, E.A. Emerson, Design and synthesis of synchronization skeletons using branching time temporal logic, in: D. Kozen (Eds.), Workshop on Logics of Programs, Lecture Notes in Computer Science, vol. 131, Springer, Berlin, Heidelberg, 1981, pp. 52–71. DOI: https://doi.org/10.1007/BFb0025774

[2] J.P. Queille, J. Sifakis, Specification and verification of concurrent systems in CESAR, in: M. Dezani-Ciancaglini and U. Montanari (Eds.), Proceedings of the 5th International Symposium on Programming, Lecture Notes in Computer Science, vol. 137, Springer, Berlin, Heidelberg, 1982, pp. 337–351. DOI: https://doi.org/10.1007/3-540-11494-7_22

[3] C. Baier, J-P. Katoen, Principles of Model Checking, MIT Press, 2008.

[4] M. Kwiatkowska, G. Norman, D. Parker, Stochastic model checking, in: M. Bernardo, J. Hillston (Eds.), Proceedings of the Formal Methods for the Design of Computer, Communication and Software Systems: Performance Evaluation (SFM), Springer, Berlin, Heidelberg, 2007, pp. 220–270. DOI: https://doi.org/10.1007/978-3-540-72522-0_6

[5] V. Forejt, M. Kwiatkowska, G. Norman, D. Parker, Automated verification techniques for probabilistic systems, in: M. Bernardo, V. Issarny (Eds.), Proceedings of the Formal Methods for Eternal Networked Software Systems (SFM), Springer, Berlin, Heidelberg, 2011, pp. 53–113. DOI: https://doi.org/10.1007/978-3-642-21455-4_3

[6] G.D. Penna, B. Intrigila, I. Melatti, E. Tronci, M.V. Zilli, Bounded probabilistic model checking with the muralpha verifier, in: A.J. Hu, A.K. Martin (Eds.), Proceedings of the Formal Methods in Computer-Aided Design, Springer, Berlin, Heidelberg, 2004, pp. 214–229. DOI: https://doi.org/10.1007/978-3-540-30494-4_16

[7] E. Clarke, O. Grumberg, S. Jha, et al., Counterexample-guided abstraction refinement, in: E.A. Emerson, A.P. Sistla (Eds.), Computer Aided Verification, Springer, Berlin, Heidelberg, 2000, pp. 154–169. DOI: https://doi.org/10.1007/10722167_15

[8] H. Barringer, R. Kuiper, A. Pnueli, Now you may compose temporal logic specifications, in:

Proceedings of the Sixteenth Annual ACM Symposium on the Theory of Computing (STOC), ACM, 1984, pp. 51–63. DOI: https://doi.org/10.1145/800057.808665

[9] A. Pnueli, In transition from global to modular temporal reasoning about programs, in: K.R. Apt (Ed.), Logics and Models of Concurrent Systems, Springer, Berlin, Heidelberg, 1984, pp. 123–144. DOI: https://doi.org/10.1007/978-3-642-82453-1_5

[10] B. Meyer, Applying "Design by Contract", Computer 25(10) (1992) 40–51. DOI: https://doi.org/10.1109/2.161279

[11] S. Bensalem, M. Bogza, A. Legay, T.H. Nguyen, J. Sifakis, R. Yan, Incremental component-based construction and verification using invariants, in: Proceedings of the Conference on Formal Methods in Computer Aided Design (FMCAD), IEEE Press, Piscataway, NJ, 2010, pp. 257–256.

[12] H. Barringer, C.S. Pasareanu, D. Giannakopolou, Proof rules for automated compositional verification through learning, in Proc. of the 2nd International Workshop on Specification and Verification of Component Based Systems, 2003.

[13] M.G. Bobaru, C.S. Pasareanu, D. Giannakopoulou, Automated assume-guarantee reasoning by abstraction refinement, in: A. Gupta, S. Malik (Eds.), Proceedings of the Computer Aided Verification, Springer, Berlin, Heidelberg, 2008, pp. 135–148. DOI: https://doi.org/10.1007/978-3-540-70545-1_14

[14] Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). Data analysis: A model comparison approach. Routledge.

[15] Mannila, Heikki. "Data mining: machine learning, statistics, and databases." Proceedings of 8th International Conference on Scientific and Statistical Data Base Management. IEEE, 1996.

[16] Dangeti, P. (2017). Statistics for machine learning. Packt Publishing Ltd.

[17] Xu, R., & Wunsch, D. (2008). Clustering (Vol. 10). John Wiley & Sons.

[18] Ding, Chris, and Xiaofeng He. "K-means clustering via principal component analysis." Proceedings of the twenty-first international conference on Machine learning. 2004.

[19] Jafarzadegan, Mohammad, Faramarz Safi-Esfahani, and Zahra Beheshti. "Combining hierarchical clustering approaches using the PCA method." Expert Systems with Applications 137 (2019): 1-10.

[20] Arias-Castro, Ery, Gilad Lerman, and Teng Zhang. "Spectral clustering based on local PCA." The Journal of Machine Learning Research 18.1 (2017): 253-309.