Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 1, No. 1, 2023

A study of the price situation of sailboats based on a comprehensive evaluation model

Zhuoning Jin[1], Zhiwang Mao[1], Haocong Ding[1], Jingyan Cai[1], Jingbin Xue[1], Yuhan Zhang[2]

[1] School of physics, Hangzhou Normal University, Hangzhou, 311121, China
[2] School of international studies, Hangzhou Normal University, Hangzhou, 311121, China

Abstract: In order to optimize the price evaluation of used sailboats, this paper collects and analyzes a large amount of relevant data, and adopts principal component analysis, multiple linear regression and neural network model to establish the linkage between price and each variable. The non-numerical data are quantified and processed, and the comprehensive evaluation model is constructed by entropy value method and linear weighting method. The TOPSIS method was used to revalue the price and optimize the model. The geographic location is quantified by spherical coordinates and analyzed by clustering to build a geographic-price model. Finally, taking Hong Kong as an example, the time series and support vector machine models are used for pricing.

Keywords: Used Sailboat; Price; Comprehensive Evaluation Model; Hong Kong (SAR).

1. Introduction

The used sailboat market is growing rapidly and research on the used sailboat market is constantly evolving. The market for used sailboats is growing rapidly and research on the market is constantly evolving [1]. Due to the increasing number of used sailboats being sold, the price of used sailboats has become a major issue that could affect our sustainable way of living [2]. As the price of brand-new sailboats is unaffordable for most sailing enthusiasts, the size of the used sailboat market is rapidly expanding. However, the appraisal and evaluation standards for the price of used sailboats have not yet been perfected. The average age depreciation method and empirical appraisal, which are commonly used at present, lack scientific basis and are more difficult to convince many buyers. In order to solve these problems, we will study the evaluation methods in line with the market law, standardize the trade and appraisal of used sailboats, improve the trust of users and promote the development of the market.

2. Evaluation Model of Sailboat Prices

2.1. Pricing with a comprehensive evaluation model

2.1.1. Quantification of non-data-type data

In order to facilitate the establishment of the subsequent evaluation matrix, the various non-data attributes of the sailboat are quantified here in a more specific way around the impact of the attributes on the sailboat. By analyzing the sailboat data we found that the geographic environment and the development of the shipbuilding industry have a greater impact on the pricing situation of used sailboats, and the trend of the impact shows a trend of a greater impact and then a smaller impact. Accordingly, the expression for the final regional superiority class W is obtained:

$$W = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 \sqrt{x_3} + \alpha_4 \sqrt{x_4} \tag{1}$$

2.1.2. Sailboat Price Evaluation Model

In order to get the accurate evaluation value, it is necessary to establish a perfect price evaluation model, we are based on the entropy value method, TOPSIS method and other basic models to assess the impact of various indicators of the sailboat on the price, and analyze the linkage relationship that may exist between various variables.

Prior to the sailboat price assessment, the indicators were first preprocessed for consistency and dimension lessness. The entropy weighting method determines the weights by the magnitude of the variation in indicator differences[3].After that, the entropy value method is used to determine the weighting coefficients of each feature of the sailboat, and the information entropy of each feature of the sailboat is calculated through the input of a large amount of data, so as to get the weighting coefficients about each feature of the sailboat. We use to indicate the weight of the jth indicator in the ith sailboat, and the entropy value is calculated for the jth indicator:

$$e_j = -\frac{1}{\ln n}\sum_{i=1}^{n} p_{ij} \ln p_{ij} (j = 1,2,...,m) \tag{2}$$

After obtaining the weight coefficients of each feature of the sailboat, the sailboat is priced by a simple linear weighting model, let f denote the comprehensive evaluation value of the sailboat pricing, then there are:

$$f = \sum_{j=1}^{m} w_j * b_j \tag{3}$$

Where denotes w_j the extent to which sailboat characteristics affect pricing outcomes, f reflecting the size of pricing.

Through the above analysis, the approximate range of each feature for the sailboat with higher estimated price, according to which we can simulate to construct an optimal (i.e., the highest estimated price) and the least optimal (i.e., the lowest estimated price) sailboat, and roughly get the price of other sailboats by judging the difference in features between other objects and this optimal and least optimal sailboat. Therefore, we build an optimization model based on the TOPSIS method to revalue the prices of sailboats. For the ith ship, there are:

$$f_i^{'} = \frac{s_i^{-}}{s_i^{-} + s_i^{+}} (i = 1,2,...,n) \quad (4)$$

included among these

$$s_i^{-} = \sqrt{\sum_{j=1}^{m} (b_{ij} - c_j^{-})^2} (i = 1,2,...,n)$$

$$s_i^{+} = \sqrt{\sum_{j=1}^{j=m} (b_{ij} - c_j^{+})^2} (i = 1,2,...,n) \quad (5)$$

C^{+} for each feature of the highest priced sailboat, C^{-} Characteristics of the lowest priced sailboat.

2.2. Regional-Price Relationship Model Building

2.2.1. Quantifying geographic location in spherical coordinates

The principal component analysis above shows that geographic region (REGION) is a key factor influencing the pricing of used sailboats, which are prevalent in coastal areas such as Europe, USA and the Caribbean. In order to perform the distance operation, the location of the region needs to be expressed in latitude and longitude (u, v), and a three-dimensional coordinate system is established with the center of the earth O as the origin and the equatorial plane as the xOy plane, and the radius of the earth is set to be R=6370km.

$$\begin{cases} x = R\cos u \cos v \\ y = R\sin u \cos v \\ z = R\sin v \end{cases} \quad (6)$$

After geometric operations, it can be obtained that any two points A (u_1, v_1) ,B (u_2, v_2) ,let the length of OA be r_1 ,the length of OB be r_2 ,and the distance between the two A, B be:

$$d = R\arccos\left(\frac{\vec{r}_1 \cdot \vec{r}_2}{|r_1| \cdot |r_2|}\right) \quad (7)$$

Simplifying the above equation, we get:

$$d = R\arccos[\cos(u_1 - u_2)\cos v_1 \cos v_2 + \sin v_1 \sin v_2] \quad (8)$$

2.2.2. Sailboat Price-Territory Relationship Model

After calculating the coordinates (x, y, z) of each region using latitude and longitude information, cluster analysis was performed according to forumla 6.K-means algorithm is a classical clustering algorithm based on partition.[4] Considering the denseness and dispersion of the clusters, the number of clusters k is determined using the contour coefficient method to ensure that the samples within the clusters are dense and the samples between the clusters are dispersed. Then using K-mean clustering, the KMeans function of the CLUSTER submodule is invoked to segment the data set, and appropriate classification results are obtained as shown in the right figure. In geographic region clustering, cluster analysis is performed directly using (x, y, z) as input to subdivide the regions. It should be noted that for the region must be subdivided, otherwise obviously Europe and North America will be divided into two types of geographic regions as a whole, so simple clustering is performed first by macro features such as continents, and then detailed clustering is performed for each region. Take Europe as an example, the figure below shows the distribution of European countries obtained from our clustering, with different colors indicating different categories. Interval pricing of sailboat prices, the 2000 sets of data in the price of sailboats is divided into lowest, lower, medium, higher, highest five categories, the division results are shown in Figure1, Figure2 and Figure3.

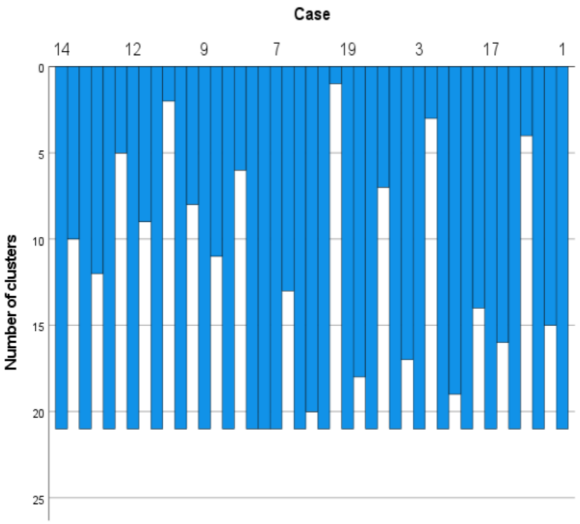

Figure 1. K-value clustering results

Figure 2. Distribution of countries in Europe

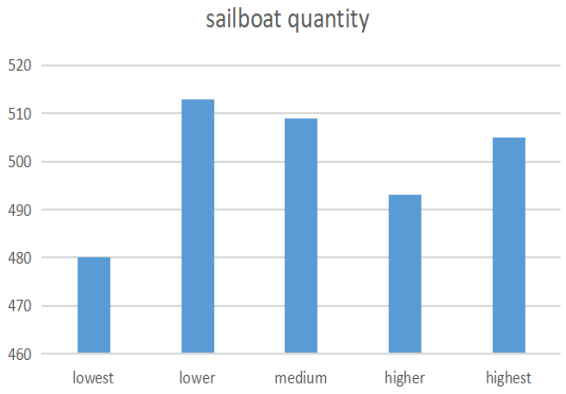

Figure 3. Price Classification

In statistics, the t-test is commonly used to compare the differences between two groups of data. The t test is the most common statistical test used in scientific publications. Compared with the conventional Neyman - Pearson approach, the evidential approach requires larger sample sizes [5]. In this study, we innovatively apply the t-test to the correspondence between geographic location and price. First, the categorized geographic locations and prices are combined

into dataset A, and then the t-test operation is performed with dataset B of the five price categories. If the number of data is inconsistent, the most representative data in dataset B is taken. If the largest p-value is obtained, it means that the price of sailboats in that geographic area is the least different from that price category, and therefore the sailboats in that geographic area can be positioned in that price category. For example, the results of the t-tests performed for the five price categories in the geographic area of the red dot in the chart of Upper Europe are shown in the table1 below:

Table 1.T-test results

category	lowest	lower	medium	higher	highest
p	0.01	0.03	0.03	0.12	0.05

Then it means that the prices in this geographic area are very significantly different for all categories of prices except for higher category prices, so the prices in this geographic area are considered to be within the higher division.

2.3. Preprocessing of data

2.3.1. Detection and handling of outliers

Inevitably, there are missing values and outliers in the searched data. For missing values, it is sufficient to use the Pandas library in python to detect them. For outliers, there are two ways of detecting them. Since the data does not follow normal distribution, here we use the box-and-line plot method instead of the standard deviation method to detect them, and the results are shown in Figure 4 and table2:

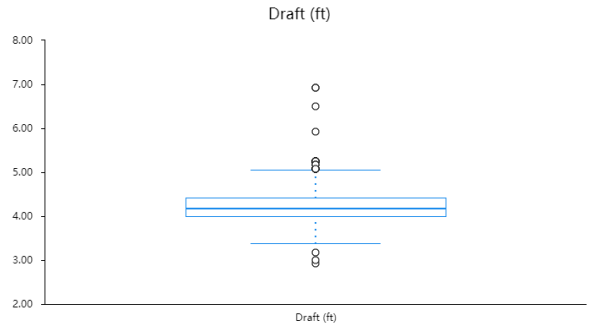

Figure 4. Draft

Table 2. Outliers summary (Draft (ft))

items	IQR value	Q1-1.5IQR	Q3+1.5IQR	Number of outliers	Specific outlier figures
Draft (ft)	0.420	3.370	5.050	19	2.92,3,3.17,5.08,5.08,5.08, 5.08,5.08,5.17,5.17,5.25, 5.25,5.25,5.25,5.25,5.92, 6.5,6.92,6.92

When dealing with outliers, because the various data of used sailboats are too cumbersome and complex, we chose to simplify and analyze the data, using the form of interpolation to deal with the outliers in each variable in turn, and use Lagrangian interpolation to interpolate data such as listed price, draught depth, and discharge volume.

Take the example of prices where there are a large number of outliers. y denotes the price of a used sailboat and x denotes the quantized used sailboat brand.

The Lagrange interpolation polynomials are shown in formula 9.

$$P(x)=\sum_{i=0}^{n}\left(\prod_{j=0,j\neq i}^{n}\frac{x-x_j}{x_i-x_j}\right)y_i \tag{9}$$

where the Lagrangian interpolation basis function is shown in formula 10:

$$l_i(x)=\prod_{j=0,j\neq i}^{n}\frac{x-x_j}{x_i-x_j} \tag{10}$$

Write a program to implement Lagrange interpolation as shown in Figure 5:

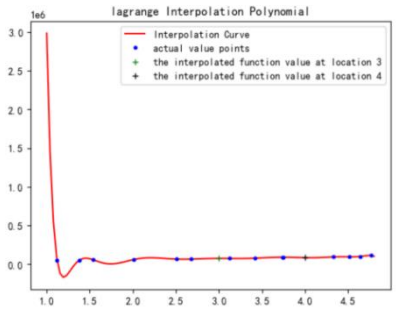

Figure 5. Lagrarge Interpolation Polynomial

As can be seen from the figure, a reasonable interpolation curve is obtained after the Lagrangian interpolation of the quantized sailboat brands and prices. The brand pricing data with outliers can be well represented on the interpolation curve. This indicates that the use of Lagrange interpolation is appropriate for outliers in used sailboat pricing. Similarly, similar Lagrangian interpolation is performed on the collected data,which can also obtain complete results.

Deep screening of variables

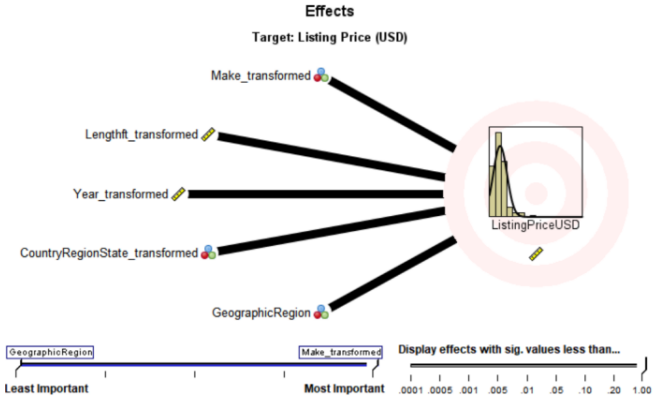

Figure 6. Effect

After the initial cleaning of the variables, an in-depth screening of the variables will be done using principal component analysis. For the 15 variables that can be found, model, length, region, year (manufacture), breadth, draft, displacement, rigging, sail area, hull material, engine hours, and headroom, are set to be X_1, X_2, X_3, ……, X_{13} ,find C_{i1}, C_{i2}, C_{i3}, ……, C_{i12} ,under the premise that

$\sum_{k=1}^{12} c_{ij}^2 = 1$,to make the value of variance $Var\left(c_{i1}x_1 + c_{i2}x_2 + c_{i3}x_3 + \ldots\ldots + c_{i12}x_{12}\right)$ is maximized. For this, a Python program was written to process it and the resultant graph was obtained in Figure6 and Figure7

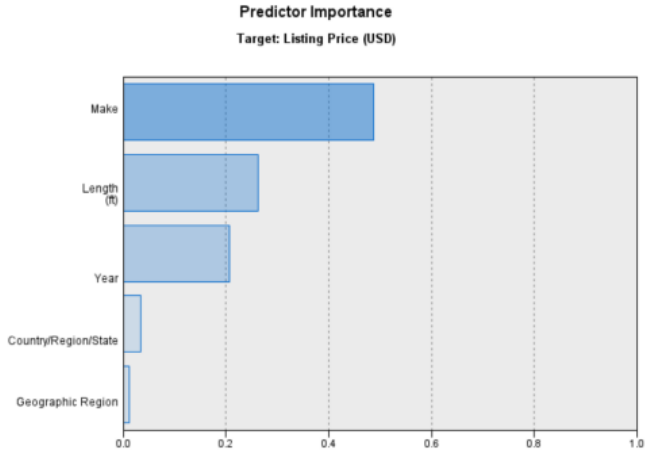

Figure 7. Pradietar importance

Based on the results of the program run, we can conclude that the six variables of LENGTH, REGION, MODEL, ENGINE HOURS, DRAINAGE, and SAIL AREA have a large impact on the listed price.

3. Results

3.1. Results of the integrated evaluation model

3.1.1. Quantification of data

Taking the economic parameters x_1 as an example, we use the expressions and constraints and combine them with a large amount of data for linear regression and least squares estimation.

The final relationship is obtained as

$$x_1 = 0.32P + 0.51\frac{1}{E} + 0.17\frac{1}{G} \tag{11}$$

For the regional merit class W, a nonlinear regression is computed on the model obtained from the operation to obtain the final quantitative expression.

$$W = 0.12x_1 + 0.05x_2 + 0.43\sqrt{x_3} + 0.4\sqrt{x_4} \tag{12}$$

By quantifying the data using the methods described above (some of the data is shown in the figure below), you can see that non-numerical data that is difficult to compare has been quantified into data that is comparable and can be calculated. The results are partly shown in Table3.

Table 3. Post-quantitative data

number	length	region	displacement	sail area	engine hours
1	1.515730303	0.141853385	0.077981321	1.425121988	0.046658722
2	0.580848957	1.912219285	1.115482937	0.938414985	1.210971132
3	1.311268866	1.223092458	1.405204212	0.397545218	1.462363548
4	1.152567372	1.55779692	0.202631571	1.813480745	0.825947597
5	0.199118493	0.282665187	0.995457866	1.53875451	1.982535447

3.1.2. Results of the sailboat price evaluation model

For the weighting coefficients of each indicator of the sailboat, we get the results shown in Table4 after programming based on the model using python.

Table 4. Weighting factors for indicators

sailboat features	variant	length	region	displacement	sail area	engine hours
coefficient	0.93	0.32	0.52	0.07	0.37	0.64

Afterwards, we applied these coefficients in the linear weighting method calculations and used polynomial regression to regress the metrics to the specific pricing of the used sailboat, and polynomial fitting with matlab software revealed that the cubic curves fit the pricing P(price) (in millions of dollars) vs. f better. The results are shown in Figure8.

$$P = 9.5f^3 - 14.1f^2 + 14.9f + 20.2 \tag{13}$$

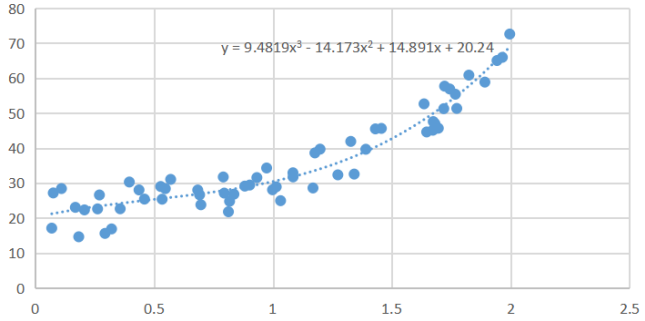

Figure 8. Function P

Finally, to test the accuracy of the price revaluation model, we do the following error analysis comparing $f_i^{'}$ vs f_i:

$$\sigma_{f^{'}} = \sqrt{\frac{\sum_{i=1}^{m}(f_i - f_i^{'})^2}{m*(m-1)}}, \quad \eta_{f_i^{'}} = \frac{\sum_{i=1}^{m}\frac{\sigma_{f^{'}}}{f_i}}{m} \tag{14}$$

It is obtained to be about 4.4%, which conforms well to the theoretical value within the error margin of 5%. Therefore, the above model can be used to optimize the evaluation matrix model within a certain time period.

3.2. Model prediction results

After categorizing the two, each geographical area type is corresponded to each price type. In practice, for an area with an unknown price, the distance between the area and the known area can be calculated and matched with the above

results, making it possible to roughly infer the pricing of a used sailboat in this geographic area when the geographic type is known. The specific pricing method is as follows: find the most representative location data in each sailboat geographic tendency classification (in the actual programming process, mostly using the average value), calculate the distance between the area that needs to be valued and each representative data, when the minimum value is taken, then classify the area, and the price corresponds to the level of the area can be. The following is a comparison of the theoretical and actual values of 50 groups of samples of monohulls and catamarans obtained through the stock price in the above way (since it is not convenient to express the categories directly in the charts and graphs, it is considered that a random value is taken to characterize the category in the corresponding category).

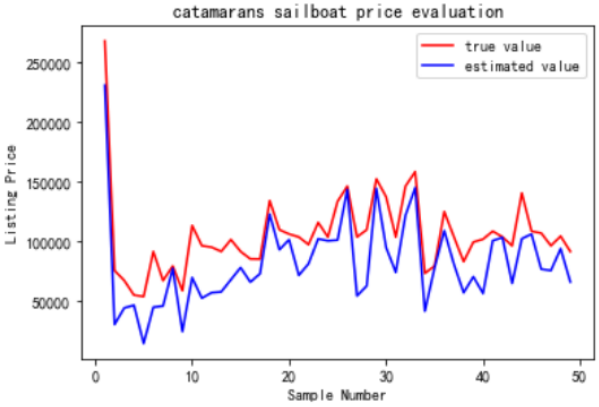

Figure 9. Catamarans sailboat price evaluation

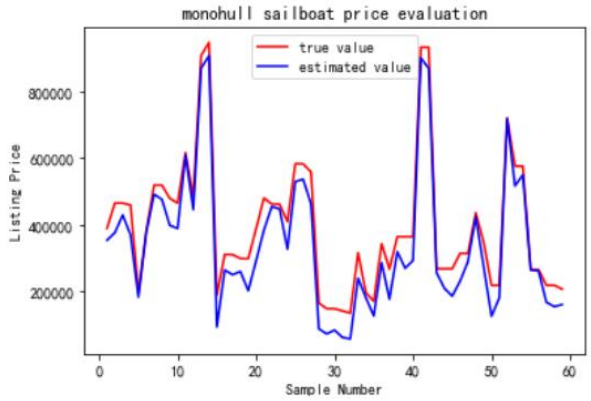

Figure 10. Monohull sailboat price evaluation

As can be seen from Figure 9 and Figure 10, the prices of sailboats predicted by this method are in a general trend with the theoretical values.

3.3. Analysis of application result

The data of second-hand sailboats in Hong Kong is selected for the application of the model. We can see that the data spans a small-time horizon, which is very unfavorable for the determination of the coefficients of the model evaluation matrix as well as for simulating the stability of the model over a long period of time. Therefore, we decided to first use the time series model to simulate the value patterns of different types of monohulls and catamarans in the price range of 10-50w over a larger time span, respectively, in order to facilitate the subsequent application of the model and analysis of the results.

In addition, in the process of data collection, we found that the types of second-hand sailboats in Hong Kong are very small, mainly limited to eleven models, but the quantification of models we did previously was based on a large number of sailboat types, so the quantification of international sailboat types established in the Hong Kong model is not applicable. Therefore, we had to reclassify used sailboats. We use the Support Vector Machine (SVM) model to finely classify the sailboat types by the parameters of hull materials, rigging, and sleeping capacity of the sailboat, and assign the classes separately. The classification results are obtained as in Figure 11:

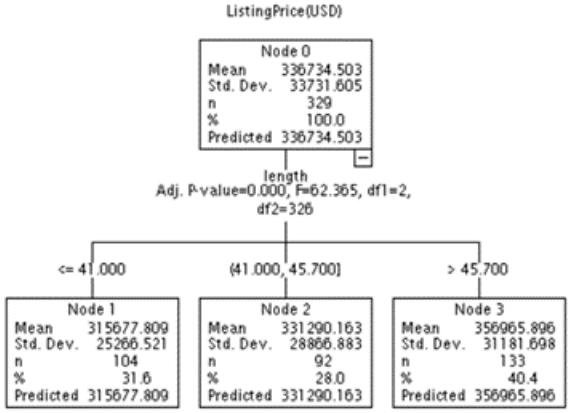

Figure 11. Listing Price

After completing the prediction and categorization work, we use a comprehensive evaluation model on the processed Hong Kong second-hand sailboat data to obtain the price prediction of Hong Kong sailboats, and display the estimated and actual prices of monohulls and catamarans in Figures 12 and 13.

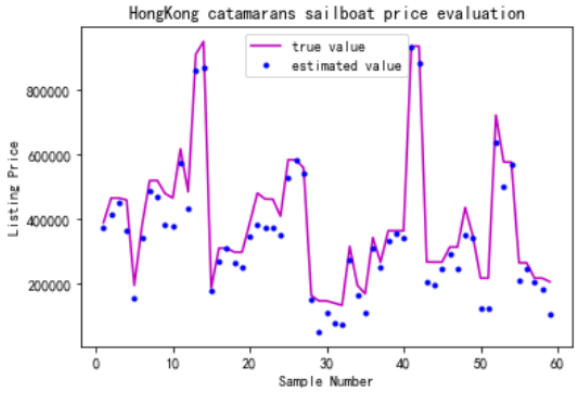

Figure 12. Listing Price

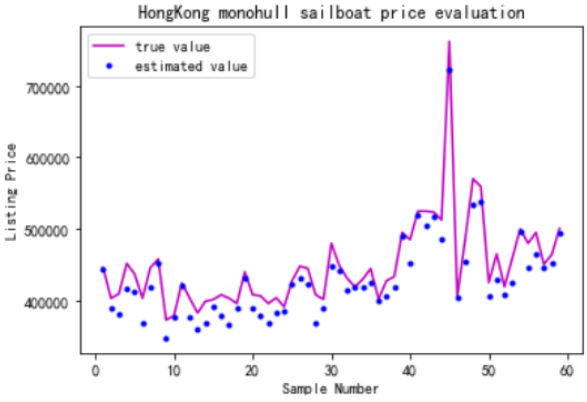

Figure 13. Hongkong monohull sailboat price evaluation

From the results in Figures 12 and 13, it is clear that our integrated evaluation model has roughly the same impact on monohulls and catamarans, and that the integrated evaluation model provides a better valuation and a roughly correct depiction of the direction and trend of the price estimates.

Among them, the measured values of both monohulls and catamarans are lower than the theoretical values as a whole, which is because it is difficult to be comprehensive in the elements considered by the comprehensive evaluation model, and the elements that are not considered are not taken into account as part of the price, which leads to the overall low price. If a suitable constant is added to the measured value as a whole, the error of our experiment can be no more than 5%, and our comprehensive evaluation model is more accurate compared with the neural network model and the multiple linear regression model.

4. Conclusions

Used sailboat pricing is affected by length, region, model, engine hours, displacement, and sail area. Research is conducted using multiple linear regression and machine learning models. Determine feature weights by entropy method, linear weighting method for pricing, and TOPSIS method for revaluation to build a more accurate comprehensive evaluation model. Validating the model, the results are able to explain the influence of region on listing price well. In Hong Kong SAR, the SVM model is used to subdivide the sailboat types and construct the region and price model. The model is applied to the valuation of used sailboats in Hong Kong SAR and performs well.

References

[1] Xi Yuan. Used Car Market Analysis and Price Evaluation---Take Jinan City as an Example [D]. Shandong Normal University, 2020.

[2] Alhakamy A'aeshah et al. Are Used Cars More Sustainable? Price Prediction Based on Linear Regression[J]. Sustainability, 2023, 15(2): 911-911.

[3] LI Fang, LI Dongping. A portfolio evaluation model based on entropy weight method[J]. Information Technology and Informatization, 2021, (09): 148-150.

[4] JIA Ruiyu, LI Yugong. K-means algorithm for self-determination of the number of class clusters and initial centroids[J]. Computer Engineering and Applications, 2018, 54 (07): 152-158.

[5] M.B. P C, E. S M . Estimating sample sizes for evidential; t; tests[J]. Research in Mathematics, 2022, 9(1).