Strategies for Estimating Used Ship Prices by PSO-Lightgbm-Catboost

Zhengchen Li [1, #, *], Tianye Lei [1, #], Donghan Li [2, #]

[1] School of Electronic and information Engineering, South China University of Technology, Guangzhou, China, 510641

[2] School of Mechanical &Automotive Engineering, South China University of Technology, Guangzhou, China, 510641

* Corresponding Author Email: 202030241099@mail.scut.edu.cn

[#]These authors contributed equally

Abstract. With the growing popularity of water sports, the consumer market for sailing and surfing, a core sport in water sports, has attracted attention in recent years. In this study, the relationship between the conventional factors affecting the price of used sailing boats and the market price of used sailing boats is mathematically modeled and predicted. After outlier and missing value processing and visualization analysis of the collected and retrieved data, nine core indicators including sailboat performance, age of sailboat use, and the level of comprehensive regional development were selected to have a critical impact on the prices of both monohull and catamaran used sailboats. Using a series of algorithms such as LightGBM, PSO, and Catboost, a unique PSO-LightGBM-Catboost weighted fusion model was established to explore the role and influence of each factor on the prices of the two used sailboats. With this weighted fusion model, the R^2 values of the multi-factor fit were as high as expected, and the R^2 values of the predicted results were also at the expected level. Finally, it was found that the price of used sailboats is more significantly influenced by factors related to the age of the sailboat, the performance of the sailboat itself, and the economic level of the region.

Keywords: LightGBM, PSO, Catboost, Used Sailboat Prices.

1. Introduction

1.1. Problem Background

With the overall improvement of people's material life in the world, people's pursuit of sports has become more colorful. Water sports have received a lot of attention and love from sports enthusiasts in recent years. As one of the core sports of water sports, sailing has received a lot of attention in recent years. As a relative luxury consumer, the second-hand market of sailing has received much more attention than the new market. In this context, the market price of used sailing boats will be a decisive factor for successful transactions between buyers and sellers in the face of the huge potential of the used sailing boat market. This study will be based on this background to make relevant analysis and predictions.

1.2. The work of this paper

In this paper, a model framework as shown in Figure 1 is proposed to predict the price of used sailboats.

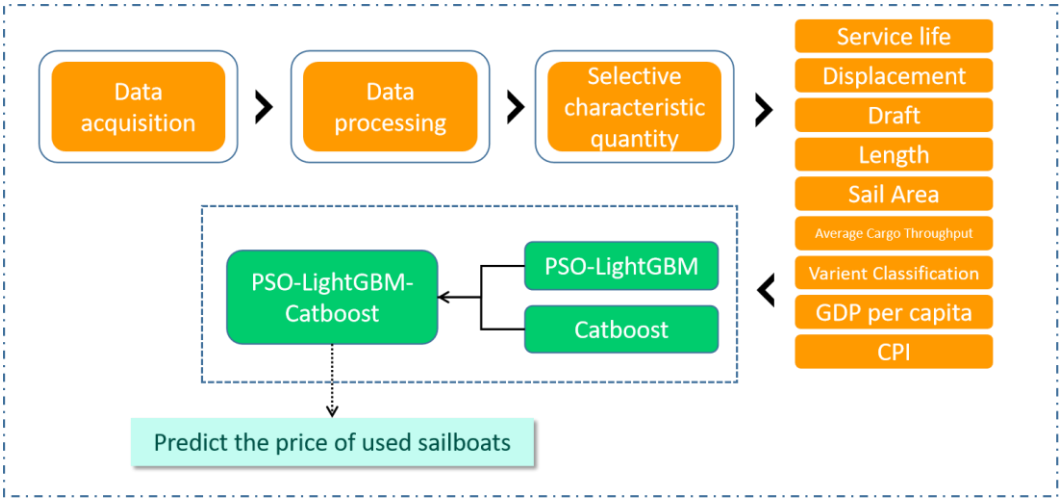

Figure 1 Flow chart of the model framework

This article was downloaded from http://www.sailboatdata.com/ The data were obtained from the World Bank, International Freight and Trade Association and World Economic Forum, and then the data were pre-processed and analyzed to select the nine characteristic factors most likely to affect the price of used sailboats, and then the LightGBM and Catboost models were fitted respectively, where the LightGBM was optimized by PSO parameters to determine its optimal parameters, and then the training set effect was used to The best-fit weights were determined for fusion to obtain the final used sailboat valuation model.

2. Assumptions and Rationale

In this section, the following global assumptions are presented; the specific assumptions for each model are described and proved in the corresponding model introduction and setup.

Assumption 1 The sailboats in the original data are all non-competitive, non-collectible sailboats.

This paper argues that certain special uses of sailboats can greatly affect the listing price of sailboats, and that these effects may come from certain other factors that are difficult to verify. For example, a competitive sailboat may be sought after for certain special advantages and its price should move closer to that of a comparable competitive sailboat, and then again, a collector's sailboat has a much greater collector's value than its use.

Assumption 2 It is assumed that all economic data are macroeconomic data, none of which originate from the effects of economic crisis, inflation, etc. in certain years.

The economic crisis, financial bubble and other factors will affect the prices of raw materials for boat building, oil prices, etc., while the manufacturing cost of sailboats will still have some impact on the selling price of sailboats.

Assumption 3 It is assumed that region-related economic data, such as GDP and port cargo throughput, are counted in the same way for each country, region, and state.

There are some differences between countries in the statistics of economic data such as GDP, and these differences are often difficult to assess. This paper can only minimize the impact of these factors as much as possible.

3. Data Exploration

3.1. Analysis of factors of price differences of used sailboats

A sailboat as a commodity has the attributes of both a recreational sport and luxury, and is a boat type of transportation in itself. In this case, the price of a used sailboat is influenced not only by its own parameters, but also by some external factors.

Sailboat parameters refer to the characteristics of the sailboat itself, mainly including the brand and the sailboat configuration. Due to the different brands, the materials and hull structure used in the sailboat will be different, and there will be price differences between different brands due to the difference in manufacturing cost and reputation. The configuration of the sailboat is mainly reflected in the parameters of the sailboat such as length, width, height and displacement, and different parameters will lead to different prices[1] . The age of the sailboat also has an impact on the price. It is easy to understand that the longer a sailboat is used, or the longer the engine is used, the more it wears out and the lower its value.

External factors mainly refer to the influence of the external environment, such as regional GDP per capita, regional maritime cargo throughput, etc. These factors may affect the price of local used sailboats from the side.

The sailing data covered in this paper are from http://www.sailboatdata.com/, economic data and throughput data from the World Bank, the International Freight and Trade Association and the World Economic Forum, to analyze the factors affecting the price of used sailboats by analyzing their parameters, age and market factors (e.g. GDP/GDP per capita/maritime cargo throughput of the region where the exchange is located).

In this paper, we extracted the data obtained from the sailboat's brand, brand variant, skipper, geographic region, price tag and year of manufacture, sailboat's width, draft depth, displacement, sail area, regional average cargo throughput, regional GDP per capita, etc., and organized them into an EXCEL table named Supplementary_data.xlsx.

3.2. Data pre-processing

Sailboats are divided into monohull and catamaran. In order to build a more accurate price prediction model, this paper extracts the price data of monohull and catamaran sailboats, and then draws a line graph to observe the relationship between them.

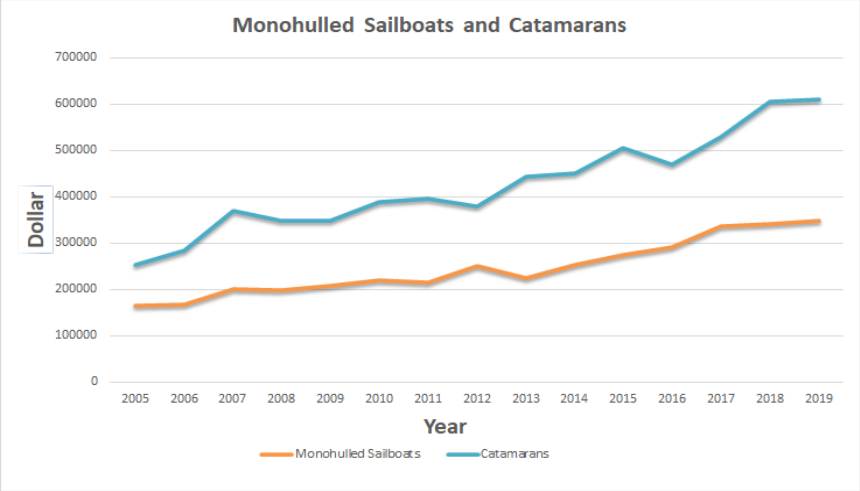

Figure 2 Used sailboat price line graph

The pre-processed used sailboat price line graph is shown in Figure 2. It is easy to see that the price of catamaran sailboats is significantly higher than that of monohull sailboats, so this study further split the data in the Supplementary_data.xlsx file into two sheets for monohull and catamaran sailboats.

Use python to examine the data file Supplementary_data.xlsx, visualize the missing values, and later exclude the null values.

Next, the data are pre-processed by:

- For brands and brand variants, the study concluded that, to some extent, the number of times an item is sold reflects the degree of fluctuation in the value of the item. Therefore, brands and variants are counted and sorted according to the number of times they appear in the file , and then the different brands and variants are coded and digitized.

- Subtract the year of manufacture from the current year to get the age of the sailboat.
- For the listing price, first, find the data of sailboats of the same make and model, sold to the same region, and average the data of that category, then, compare the data of that category with the average: if the absolute value of the difference between a data and the average exceeds 20% of the average, the data of that category is considered as abnormal. If the difference between that category of data and the average exceeds 20% of the average, the category of data is considered abnormal. It is removed and the other data in that category are averaged again, and finally the abnormal data are classified as average.

Finally, in this paper, a total of nine variables, namely service life, displacement, draft, length, sail area, average cargo throughput, variable classification, GDP per capita, and CPI, were selected as the characteristic variables of the used sailing ship price prediction model.

3.3. Introduction to the method

3.3.1. LightGBM algorithm

The LightGBM algorithm was proposed mainly to solve the problem of slow running and memory consumption of GBDT when encountering large amount of data, and its solution is to use two new techniques, namely gradient-based one-sided sampling (GOSS) and mutually exclusive feature bundling (EFB).LightGBM is a framework for implementing GBDT, which has the advantages of fast training, efficient parallel training, distributed support for fast processing of large amounts of data, low memory consumption, and high accuracy.[2] .

3.3.2. Catboost algorithm

CatBoost's base learner , a GBDT framework based on symmetric decision trees, supports categorical variables with few parameters and high accuracy, mainly solves the problem of poor processing of categorical features, but also solves the problem of prediction bias and gradient bias, reduces overfitting, and improves the generalization ability and accuracy of the algorithm[3] . Catboost also combines categorical features. Catboost also solves the problem of prediction bias by using the connection between features, enriching the feature dimension, and using ranking boosting to reduce the noise in the training set and avoid the bias in gradient estimation.

Catboost has the following five advantages[4] :

●High model quality without parameter adjustment and good results with default parameters

●Supports categorical variables, eliminating the need to pre-process non-numeric features

●Rapid prediction, allowing models to be deployed quickly and efficiently, even for latency-demanding tasks;

●Superior performance: comparable in performance to any advanced machine learning algorithm;

●A new gradient boosting mechanism can be used to construct the model to reduce overfitting and improve accuracy [5] .

3.4. Introduction of model evaluation indicators

(1) Mean Absolute Percentage Error (MAPE)[6]

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \quad (1)$$

where yi^ is the predicted value, yi is the true value, and n is the number of samples in the Verification set. When the value of MAPE is smaller, it means that the error between the predicted value of the number of reporters and the actual observed value of the number of reporters is smaller, and furthermore, the performance of the prediction model used is better.

(2) Judgment factor (R^2)[6]

$$R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST} \quad (2)$$

a) SSR: sum of squares of the regression, i.e., the sum of squares of the difference between the predicted data and the mean of the original data

b) SST: Sum of Total Squares, i.e., the sum of squares of the differences between the original data and the mean

4. PSO-Catboost-LightGBM Price Forecasting Model

4.1. LightGBM-based Valuation Model for Used Sailboats

4.1.1. PSO-LightGBM parameter optimization

The sailboat dataset used was processed and a total of 2987 data were available for training, and the hyperparametric optimization heuristic PSO was used to find its optimal combination of parameters before constructing LightGBM with Spss[7] . The LightGBM parameters are introduced as shown in Table 1.

Table 1. Introduction of important parameters of LightGBM

Parameter Name	Meaning
Basic Learners	Used to specify the type of weak learner LightGBM uses decision trees as the base learner
num_boost_round	The number of trees in LightGBM. The higher the number, the higher the complexity of the model, but it may also lead to over-fitting.
Learning Rate	The parameter used to control the degree of contribution of each tree, usually ranging from 0 to 1.
lambda_l1	The multiplicative term of the sum of the absolute values of the weights in the model is used to reduce the complexity of the model and to prevent over-fitting.
lambda_l2	The multiplicative term of the sum of squares of the weights in the model is used to reduce the complexity of the model and prevent over-fitting.
Bagging rate	The proportion of training samples used to construct the tree in each iteration
Feature_Score	The proportion of features used to construct the tree in each iteration.
Minimum gain	A threshold value used to control the growth of the tree. When the gain of the samples in the split nodes is greater than this threshold, the split is performed.
min_sum_hessian_in_leaf	denotes the minimum value of the sum of Hessian matrices (i.e., the sum of sample weights) of the samples on the leaf nodes.
max_depth	Number of nodes in the longest path from the root node to the leaf node
min_data_in_leaf	denotes the minimum number of samples that must be available at each leaf node.

The steps of the basic particle swarm algorithm are relatively simple. The particle swarm optimization algorithm consists of a set of particles moving in the search space subject to their own best past position pbest and the best past position gbest of the entire swarm or nearest neighbor. the dth dimensional velocity update of particle i in each iteration is given by the following equation[8] . Among them, the parameters of the particle swarm optimization algorithm formula are shown in Table 2.

$$v_{id}^{k+1} = \omega v_{id}^{k} + c_1 r_1 \left(p_{id,\text{pbest}}^{k} - x_{id}^{k}\right) + c_2 r_2 \left(p_{d,\text{gbest}}^{k} - x_{id}^{k}\right) \quad (3)$$

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \quad (4)$$

Table 2. Introduction of parameters of particle swarm optimization algorithm formula [9]

Symbols	Description
N	Particle swarm size
$i, i = 1,2, \ldots, N$	Number of particles
r_1, r_2	Interval $[0,1]$ Random numbers in the interval to increase the randomness of the search
D	Particle size
$d, d = 1,2, \ldots, D$	Particle size serial number
k	Number of iterations
ω	Inertia weights
c_1	Personal Learning Factors
c_2	Group learning factors
v_{id}^{k}	Particle i In the first k iteration of the d dimensional velocity vector
x_{id}^{k}	Particle i In the first k iteration of the d dimensional position vector
$p_{id,\,\text{pbest}}^{k}$	Particle i In the first k iteration of the d The best position in the dimensional history
$p_{d,\,\text{gbest}}^{k}$	Population in the first iteration k iteration of the d The historical optimal position of the dimension

The velocity and position of each particle are initialized in a random manner. Then, the particles approach to the global optimum and individual optimum[5] . For the parameters of the PSO algorithm, the recommended value of the basic algorithm, i.e., N = 50, is chosen here in this paper $k = 150$,$\omega = 0.9$, and$c_1 = 2$, $c_2 = 2$ [10].

The final optimal parameter results are: base_learner-"GBDT", num_boost_round=50, learning_rate=0.13, lambda_l1=1, lambda_l2=1, bagging_ fraction=1, feature_fraction=0.5, min_ gain_to_split= 0.656924099690, min_sum_hessian_in_leaf=0.5, max_depth=2, min_data_in_leaf=1.

4.1.2. Appraisal results

The optimal combination of parameters was used to build the used PSO-LightGBM model for sailboat valuation. In order to make the results more specific and convincing and to avoid bias in the selected validation set, this paper uses a five-fold cross-validation method, setting MAPE and R^2 as evaluation metrics, and printing MAPE and R^2 for both the training and Verification sets in each round of cross-validation.

The prediction results for each round of the training and Verification sets were collated and the final bimodal evaluation results were obtained by taking the average of five rounds of evaluation metrics:

For a single sailboat: the MAPE on the training set is 10.59575, the MAPE on the Verification set is 17.301, the R^2 on the training set is 0.951, and the R^2 on the Verification set is 0.75875, implying that the variance explained by the explanatory variables can account for about 75% of the total variance and the model has a good fit.

For the double-hulled sailboat, the MAPE is 5.833 for the training set and 9.55575 for the Verification set. R^2 is 0.95425 on the training set and R^2 is 0.84625 on the Verification set, implying that the explanatory variables explain about 84% of the total variance and the model has a good fit.

This model was constructed relatively successfully, and the evaluation results need to be compared and validated with other models in subsequent comprehensive comparisons.

4.2. CatBoost-based valuation model for used sailboats

4.2.1. Catboost model parameters

Since the Catboost model does not need to be tuned to obtain high model quality, the model in this paper is not tuned and the default parameters are used.

4.2.2. Appraisal results

Again, in order to make the results more specific and convincing and to avoid bias in the selected validation set, the method in 5.2.3 is still used here to evaluate the results. The prediction results of the training and validation sets are collated for each round, and the final bimodal evaluation results are obtained by taking the average of the evaluation metrics of the five rounds.

For a single sailboat: the MAPE on the training set was 11.7245, the MAPE on the Verification set was 17.356, the R^2 on the training set was 0.95325, and the R^2 on the Verification set was 0.755, implying that the variance explained by the explanatory variables could account for about 75% of the total variance and the model had a good fit.

For the double-hulled sailboat, the MAPE was 8.28 on the training set and 8.28 on the Verification set, and R^2 was 0.932825 on the training set and R^2 was 0.8445 on the Verification set, implying that the explanatory variables explained about 84% of the total variance and the model had a good fit.

This model was constructed relatively successfully, which means that both models have better predictions, and subsequently the two models were fused.

4.3. PSO-lightGBM and Catboost fusion valuation model

4.3.1. Weighted Fusion

The sailboat valuation models used by PSO-lightGBM and Catboost were developed separately, and both models were found to be effective and each model had its own advantages. In this section, the two models are fused using the weighted fusion method and the fused models are tested using MAPE and R^2 as evaluation metrics.

The results obtained from the Verification set were compared with the PSO-lightGBM and Catboost models since the training set has limited comparative effect and the Verification set is more effective in demonstrating the fusion model.

In the weighted fusion model, for the single ship, the MAPE and R^2 of the verification set were 16.892 and 0.7723 respectively. For catamarans: MAPE for verification set is 8.015, R^2 is 0.8572. Compared with PSO-lightGBM and Catboost models, it can be concluded that the predictive performance of weighted fusion model is better and stronger.

Overall, the optimization of the two metrics was combined, and the weighted fused PSO-LightGBM-Catboost model was successfully constructed to better evaluate the value of used sailboats.

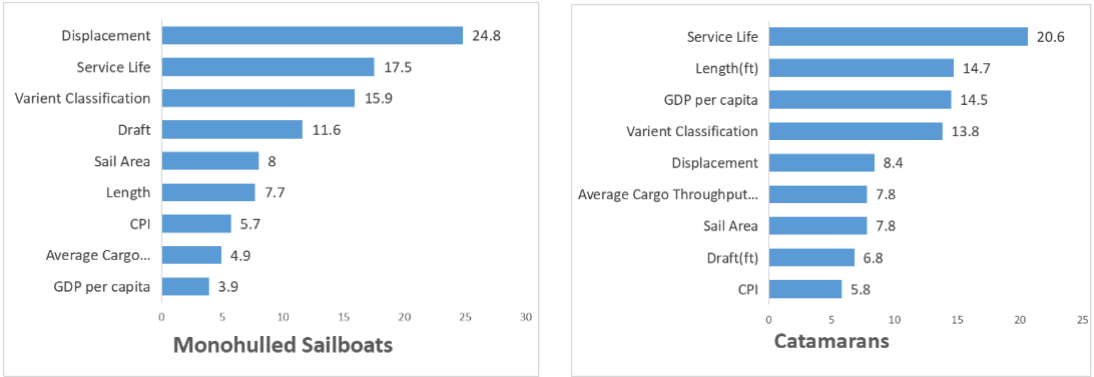

Figure 3 Feature importance analysis of PSO-LightGBM-Catboost model

The characteristic importance analysis of the PSO-LightGBM and Catboost models is shown in Figure 3. According to the analysis of the model, the factors that affect the price of monohull and catamaran sailboats are different. Figure 1 on the left shows the order of importance of nine factors on the price of sailboats.

This paper can clearly conclude that

●**Monohull sailboats The** most important factor in water displacement is 24%, followed by longevity and brand type.

●**Catamarans** Unlike monohulls, the age of the boat is the most important factor affecting the price of a catamaran, accounting for 20.6%, followed by the length of the hull and the GDP per capita of the region.

The mean absolute percentage error (MAPE) of this model is 15.79% for monohulls and 7.22% for catamarans, which means that this paper is 84.21% confident that the model is fine for monohulls. For catamarans, this paper has 92.78% confidence that the model is not a problem.

From this, this paper can argue that displacement, age and brand variants are more important for monohull sailboats, while for catamarans, the effect of GDP per capita becomes important.

In summary, this paper relies on the PSO algorithm to optimize the LightGBM parameters, and then adopts a weighted fusion method to integrate the two models through a comparative analysis with Catboost, and obtains a more excellent PSO-LGBM-Catboost price prediction model for sailboats, and passes the accuracy analysis.

5. Conclusions

In this paper, we rely on the PSO algorithm to optimize the LightGBM parameters, and then through a comparative analysis with Catboost, the two algorithms are integrated using a weighted fusion method according to the different advantages of the corresponding algorithms to obtain a PSO-LightGBm-Catboost fusion model with significant effects for sailboat price prediction model. In the process, this study found 9 factors that have a strong influence on the price of used sailboats through exploratory analysis at . Based on the commonality of the 9 factors and their own characteristics, this study conducted the corresponding process of visual analysis, data fitting and model construction. The final P-L-C fusion model was constructed. The verification set is used for testing and good results are obtained. The successful analysis of second-hand sailboat prices affected by various factors.

References

[1] Liu Xiaoqiang. Research on Investment Risk Assessment of Sanya Luhuitou Bay Sailing Port Construction Project [D]. Ocean University of China, 2012.

[2] Cui Baoyang, Ye Zhonglin, Zhao Haixing, Ren Qingzhomei, Meng Lei, Yang Yanlin. Used car price prediction based on XGBoost+LightGBM iterative framework [J]. Electronics, 2022, 11(18).

[3] Niu Li. Credit risk assessment and model research based on CatBoost fusion algorithm [D]. Taiyuan University of Technology, 2021. DOI: 10.27352/d.cnki.gylgu.2021.001059.

[4] Data whale. An exhaustive series of CatBoost [EB/OL]. https://blog.csdn.net/Datawhale/article/details/103193557, 2019-11-21.

[5] Xia Yisong, Jin Wenzhou. Short term bus passenger flow prediction and impact factor analysis based on S-Catboost algorithm [J]. Journal of Guangxi University (Natural Science Edition), 2021,46 (03): 747-763.

[6] Jesszen. [Statistical Learning 3] Linear regression: R-squared (R-squared) and Adjusted R-Squared[EB/OL]. https://blog.csdn.net/Jesszen/article/details/81017669, 2018-7-12.

[7] VariableX . LightGBM important parameters, methods, functions to understand and transfer reference ideas, grid search (with examples) [EB/OL]. https://blog.csdn.net/VariableX/article/details/107256149. 2020-07-10 17

[8] Luo Qing Yi. Parameter optimization based on improved particle swarm algorithm for support vector machines and its application [D]. Lanzhou Jiaotong University,2020.

[9] Wang JW, Wang DW. Experiments and analysis of inertia weights in particle swarm algorithms [J]. Journal of Systems Engineering, 2005, 20(2):194-198.

[10] Shi Y, Eberhart R C. Empirical study of particle swarm optimization [C]//Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406). IEEE, 1999, 3: 1945-1950.