Data Research Based on the Hong Kong Sailing Market

Yangting Liu

18954559536@stu.xjtu.edu.cn

Xi'an Jiaotong University Xi'an, China

Abstract—This article mainly conducted data research on the Hong Kong sailing market. First, we divided the factors that affect the price of sailboats into two categories. The first category is macro factors such as time and region, and the second category is micro factors that are specific to the attributes of a sailboat. For the former, we built an ARIMA model of time series distribution to predict the time forecast of the average price of sailboats, and for the latter, we built a Bayesian optimization XGBoost model based on machine learning to predict the price of a sailboat under different factors. Then, we apply the established model to Hong Kong and give the forecast of the Hong Kong market. At the same time, in view of the poor information obtained from the data, we evaluated a variety of prediction algorithms and gave the optimal solution. Next, we summarize innovative conclusions from natural conditions, economics, and sailboat design. Finally, we performed a sensitivity analysis and strengths and weaknesses analysis of the model, and evaluated the rationality of the model used at each step.

Keywords—ARIMA model, XGBoost model, k-means clustering, grid search algorithm

1. INTRODUCTION

Unlike ordinary daily necessities, sailboat trading is a commodity or "luxury" transaction. Its expensive price will inevitably affect its audience and market. Like other luxury items, the transaction price of a used sailboat varies over time and by region. It is mainly reflected in the fact that aging will be exacerbated by the passage of time, and different regions will affect the local market and purchasing power. In order to open up the Hong Kong trading market, offer reasonable prices for sale of sailboats, predict future price trends in the sailing market based on available data, and predict the market transaction price of Hong Kong sailing boats according to the factors influencing sailing boats.

2. MODEL BUILDING

2.1 Prediction of time series of second-hand sailboat prices based on ARIMA model

Before the model is established, we pre-process the data, and the processing results are shown in Table 1.

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334411

TABLE 1 IMPORTANCE TABLE

variable	length	Beam width	Draft	standard displacement	*sail area*	Material stiffness
Importance	0.3478	0.2564	0.3215	0.3475	0.2479	0.1026
variable	engine time	Net space area	Make	Variant	Geography region	Country/state / region
Importance	0.1624	0.2854	0.2457	0.1343	0.0179	0.0055

ARIMA model is a commonly used time series forecasting model, which can be used to forecast time series data[1][2]. ARIMA model is composed of autoregressive model (AR), difference model (I) and moving average model (MA), so it is also called ARIMA(p, d, q) model.

$$y_t = c + a_1 y_t + a_2 y_{t-2} + \cdots + a_p y_{t-p} + e_1 \quad (1)$$

$$\Delta y_t = y_t - y_{t-1} \quad (2)$$

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (3)$$

Among them, y_t represents the value of the dependent variable at time t, y_{t-1}represents the value of the independent variable at time t, a(1)~a(p) are model parameters, c is a constant term, e (t) is the random error. μ is the mean value, ε_t is the error term of the time series at t time, θ_1,θ_2 ...θ_q are the parameters of the model, which means that q moments ago The influence of the error term on the current.

2.2 Using ADF to Test Time Series Stationarity ADF test

Using the ADF model to test the stationarity of the second-hand sailboat price, the results are as shown in Table 2

TABLE 2 THE RESULTS OF THE STATIONARITY TEST

Statistics	cValue	*h*	pValue	stat
Result	-1.9416	1	1×10^{-3}	7.6489

Where: **h** is a Boolean value indicating whether the null hypothesis is rejected. If the null hypothesis is rejected, the price data has a unit root, i.e. is non-stationary pValue is a scalar representing the p-value of the null hypothesis. **stat** is a scalar representing the value of the test statistic. cValue is a vector of length 3 representing the critical values of the ADF test, which are related to the confidence.

According to our results, the value of pValue is much smaller than the confidence interval 0.05, and h=1 indicates that the null hypothesis is rejected, so our data is stable.

2.3 Draw ACF and PACF plots

According to the previous analysis, the data time series of the price is stable, so the step of difference can be omitted, the parameter d is set to 1, and the values of the parameters p and q need to refer to the ACF and PACF figure as follows：

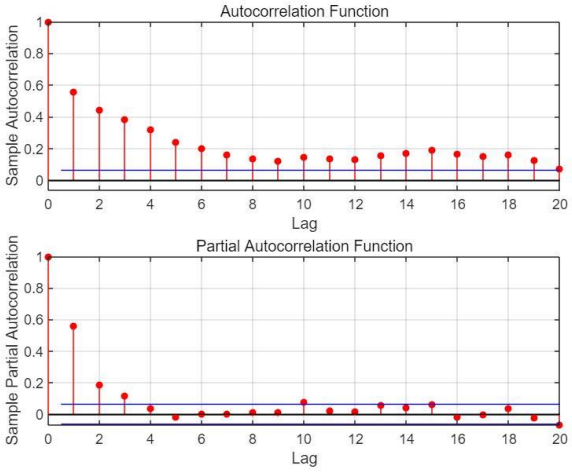


Fig. 1. ACF and PACF

According to the above Fig.1., there is only one truncation in the ACF graph, so the parameter q can take a truncated value: 0.079; there are two truncations in the PACF graph, so the larger parameter p is 0.037, and neither of these two values exceeds 5. Since these two truncations are between 0-1, ARMA (Autoregressive Moving Average Model) is selected to estimate the values of parameters p and q. The ARMA(p, q) model consists of a p-order autoregressive process and a q-order movement The linear combination composition of the averaging process is expressed as:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \qquad (4)$$

Where y_t represents the response variable at time t, c represents the intercept of the model, ϕ_i represents the i-th autoregressive coefficient, y_{t-i}represents the response variable at time t-i, θ_i represents the i-th moving average coefficient, ε_{t-i} represents the error term at time t-i, and ε_t represents the error term at time t.It can be seen from this that p=q=0.

2.4 Price prediction under the influence of other relevant factors

The Bayesian XGBoost model is an XGBoost model based on Bayesian optimization, and its purpose is to automatically optimize hyperparameters in the XGBoost model to achieve better performance[3].

In the Bayesian XGBoost model, we first need to define a parameter space, which includes all possible hyperparameters and their value ranges. We then use a Bayesian optimization algorithm to automatically select hyperparameters to minimize the loss function on the validation set. At each iteration, the Bayesian optimization algorithm uses previous results to build a Gaussian process model that represents the relationship between hyperparameters and the corresponding loss function. The algorithm then uses this model to choose the next best hyperparameters. This process continues until convergence.

The Bayesian optimization algorithm uses the Bayesian formula：

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (5)$$

to update the Gaussian process model. Among them, A and B are events, P(A|B) represents the probability that A occurs under the condition that B is known to occur, P(B|A) represents the probability that B occurs under the condition that A is known to occur, P(A) and P(B) are the prior probabilities of A and B, respectively. In the XGBoost model, we use a picture to represent the common hyperparameters that need to be adjusted. The implementation of the Bayesian XGBoost model is as shown in Table 3.

Table 3 BAYESIAN XGEBOOST MODL ALGORITHM

Input:
training data (X_train, y_train)
search space of hyperparameters (hyperparameters)
number of iterations (num_iterations)
objective function (objective_function)
acquisition function (acquisition_function)
Output:
best hyperparameters (best_hyperparameters)
Initialize the best hyperparameters and best objective function value to None.
For each iteration:
a. Sample a set of hyperparameters from the search space.
b. Train an XGBoost model using the sampled hyperparameters and the training data.
c. Evaluate the objective function using the trained model.
d. **If** the objective function value is better than the best one so far, update the best hyperparameters and best objective function value.
e. Calculate the acquisition function value for the current set of hyperparameters.
f. **If** the acquisition function value is greater than a pre-defined threshold, select the current set of hyperparameters as the new starting point for the next iteration.
g. Otherwise, sample a new set of hyperparameters from the search space and repeat the above steps.
Return the best hyperparameters

Fig.2. use the queried data to make predictions using the model. Both the feature quantity and the target variable in this data set are a data set of one thousand data, all of which are used as a training set and then used as a test set, and the predicted results and actual results are drawn in a chart and the correlation factors are calculated. In order to avoid the randomness of the data being regarded as abnormal data, the prices are processed in ascending order before processing. The results of the predicted value and the actual value are shown in the Fig.2.

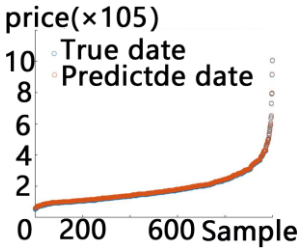


Fig. 2. Comparison chart of predicted value and actual value result

The calculated value of its fitting degree is 0.8866498, indicating that our prediction is very successful under a considerable number of factors.

3 Practical Application of Sailing Price Forecast in Hong Kong

The model we have established realizes the prediction of time series and multiple influencing factors under different regions. But now we need to use geographic information from other regions to train data in the absence of sufficient data in Hong Kong. Therefore, implicit geographic information such as country/state, manufacturer, and brand need to be taken into account as independent variables. This makes it necessary to evaluate whether the model used before is still reasonable.

3.1 Screen the best predictive model

Here we have chosen two alternatives, XGBoost model and grid search model. The pseudocode of the grid search model is shown in Table4[4].

Table 4. Grid search model algorithm

Set up hyperparameters with a grid of values to be tested
Create empty list or array to store evaluation metrics (e.g., accuracy, F1 score)
For each combination of hyperparameters:
Train a model using the current hyperparameters
Evaluate the model using a validation set and calculate the desired evaluation metric
Store the evaluation metric in the list or array from step 2
Find the hyperparameters that resulted in the highest evaluation metric
Train a final model using the best hyperparameters on the entire dataset
Test the final model on a separate test set to evaluate its performance

Both XGBoost and grid search algorithms are optimization algorithms used in machine learning. They each have pros and cons. The advantage of the XGBoost algorithm lies in its high efficiency: it adopts the method of parallel computing, which can quickly process massive data sets and shorten the training time; when processing data, it can automatically learn the characteristics of

the data, which has strong robustness; When dealing with large-scale data, it can achieve high accuracy. The disadvantage is that when the data is not fitted, overfitting may occur. In the case of a large data set, the training speed will become very slow.

In contrast, the advantage of the grid search algorithm is that the search space is wide: the grid search algorithm can search all possible combinations of hyperparameters, including some unconventional combinations, so it is possible to find a better model; the search result is A set of hyperparameter combinations can help us understand the interaction between various parameters. It does not depend on a specific model and is applicable to various types of models. However, it has a large amount of calculation and needs to search all possible parameter combinations, so when the search space is large, it takes a long time to search for the optimal solution. There may also be a local optimal solution that cannot guarantee the search for a global optimal solution, and may fall into a local optimal solution and miss a better solution.

The following is a set of directly queried sailing data in Hong Kong, and the prediction results under two different models：

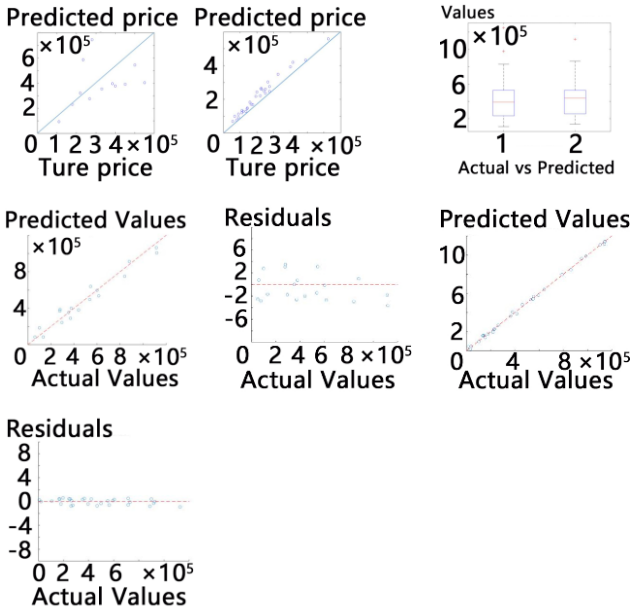


Fig. 3. Sales Statistics Chart

Fig.3. Forecast and statistics for Monohulled Sailboats and Catamarans sales in Hong Kong. Figure (a) is the result of using the CGBoost model, and Figure (b) is the result of using the grid search model. Each image contains prediction set results for 20 Monohulled Sailboats and 30 Catamarans.

The horizontal axis of the picture represents the actual value, and the vertical axis represents the predicted value. The closer to the y=x line, the better the prediction effect. From the drawn boxplot and residual plot, it can be seen that the prediction effect of grid search optimization is slightly better than that of XGBoost model. This is because the XGBoost model directly queries

too little Hong Kong data, which will affect the training effect. The grid optimization search smoothes this defect to a certain extent at the cost of time. But when we use a large amount of data in the outer region to verify Hong Kong, the error between CGBoost and grid optimization search is reduced, and the high time cost of the latter should be discarded.

3.2 Predict sailboat prices in Hong Kong using the trained model

Now we need to substitute the characteristics of Hong Kong into the previously trained model. Note that in the previous model, region was treated as an unimportant feature, but now it should be considered as an independent variable. The data is first aggregated using clustering.

The sailboat price data is clustered using the K-Means clustering algorithm[5]. The core of the clustering algorithm includes the following aspects:

3.2.1 Distance measures: Clustering algorithms usually use distance measures such as Euclidean distance, Manhattan distance, and cosine similarity to measure the similarity between data points.

3.2.2 Centroid of a cluster: The centroid of a cluster is the mean of all data points within that cluster.

3.2.3 Distance function: The distance function can be single link, full link, average link, etc. Single linkage indicates the distance between the two closest data points between two clusters, full linkage indicates the distance between the farthest two data points between two clusters, and average linkage indicates all data points between two clusters the average distance between.

3.2.4 Objective function: The goal of a clustering algorithm is to minimize the distance between all data points within a cluster and the centroid of the cluster and maximize the distance between different clusters. Here we use the k-means clustering algorithm to achieve.

In the k-means algorithm, the objective function is to minimize the sum of squared distances between each data point and the centroid of the cluster it belongs to, that is：

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_j} ||x_j - \mu_i||^2 \tag{6}$$

First, cluster the data of the previous training set according to the types and regions of sailing ships.

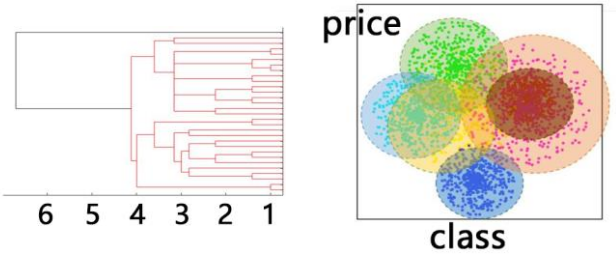


Fig. 4. The visualization effect diagram of clustering

The visualization effect diagram of clustering as shown in Fig.4., where the number of clusters k=6, represents the price distribution of two different sailboats in three different geographic regions, and the valid data is summarized in the right figure.

Next, the data are re-cleaned and these dummy variables are coded. The training variables should be Contains only a few items of the original data table. Using the XG-Boost model, the price forecast for Hong Kong is as follows:

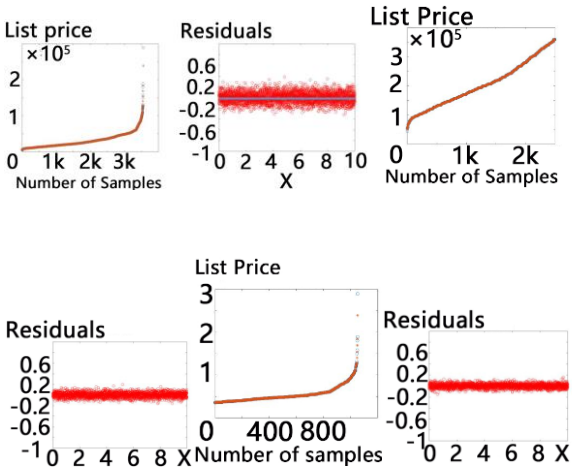


Fig. 5. Results of predicting prices in Hong Kong

Results of predicting prices in Hong Kong using the preceding model. The three sets of graphs in the Fig.5. correspond to the full data, the prediction result graphs of Monohulled Sailboats and Catamarans and their deviation graphs. It is noted that the deviation in the drawing group of all data reaches 0.2, which is twice that of each sailboat's individual drawing, so it can be explained that for different sailboats, the predicted results are not the same.

3.3 Analysis of Sailing Ship Types and Regional Differences

Aggregate the price data for the two sailboats and plot it as a boxplot. After superimposing these two boxplots and comparing them, in addition to the difference in the median, upper 3/4 quantile and lower 1/4 quantile of the two sailing boats, it can also be seen that the overall distribution of Catamarans' deviation data The value is more. Here, these data cannot be simply classified as outlier data like other statistical problems, but it shows that the price of Catamarans is greatly affected by non-quantitative factors such as geography, manufacturers, social factors and economic factors. Therefore, a further conclusion can be drawn by comparing this picture: various influencing factors have different impacts on the prices of these two sailboats. Fig.6. shows our comparison results.

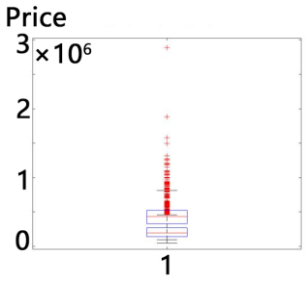


Fig. 6. Two kinds of sailboat price boxplots

Fig.7., on the other hand, shows a boxplot of sailboat prices in different geographic regions. By comparing this boxplot, we can observe the price distribution, deviation data, etc. in different geographical regions, and analyze the impact of differences in various factors in geographical regions on price distribution.

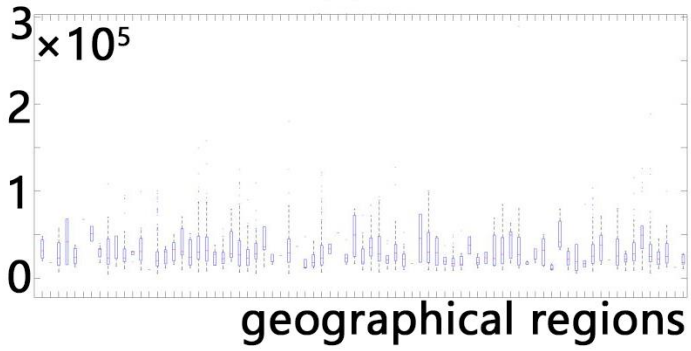


Fig. 7. Box plot of sailing price in different regions

In Figure 7, from left to right The geographical locations are, in order of preference: Aiabama, Antlgua and Bartuda, Aruba, Bahamas, Barbedos, Belgium, Belize, British Virgin blan, Bulgaria, California, Cayman islands, Connecticut , Cork, Crcatia,Cyprus,Denmark, Dominican Republic, Estonia,Fiolin,France,Georgia,Gaudry,Gbraltar,Greece,Grenada,Gueoupe,Guaterraia,Hawajl,Honduras,Hong Kong,Hungary,Minals,Meland, Massachusetts, Netherland , Netherlands Antilles, New Jerseyn, New York, North Carolina, Norway, ohio,Oregon, panama, Portugal, Puerto Rico, Rhodes Island, Romaria, Saint Kitts, Saint Lucia, Grenada, Saint-Martin, Sint Maarten, Slovenia, South Carolina, Spain,Swedn, Switzerland ,Texas, Trinidad,Turkey, US Virgin Island, United kingdom,Virginia, Washington, West Indies, Wisconsin.

3.4 Evaluation of Forecasting Methods

Since the application of the XGBoost model is subject to the following restrictions:

3.4.1 High computing resource requirements: XGBoost requires a lot of computing resources, especially when training large-scale datasets.

3.4.2 Complex hyperparameter adjustment: XGBoost has many hyperparameters to adjust, which requires users to have certain experience and skills, and requires more computing resources.

3.4.3 Sensitive to data quality: XGBoost has relatively high data quality requirements, and requires data sets to have certain standardization and denoising processing.

So we have to compare it with other schemes to confirm the optimal scheme. For common forecasting algorithms, we choose multiple regression models, XGBoost models, and LSTM models[6]. The LGB model, the one-dimensional convolutional CNN model and the CatBoost model are used as a comparison. After training with the relevant data sets at the same time, the mean square error, root mean square error, mean absolute error and coefficient of determination statistics are as shown in Fig.8.

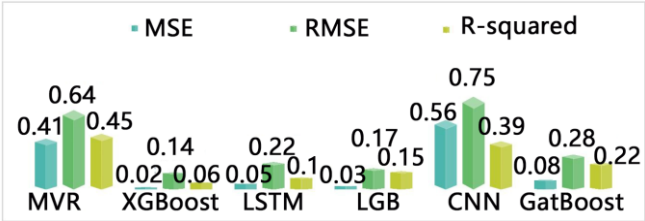


Fig. 8. Statistical comparison chart of different algorithms

The comparison chart of the time complexity of different algorithms and the size affected by the number of data is shown in Fig.9.

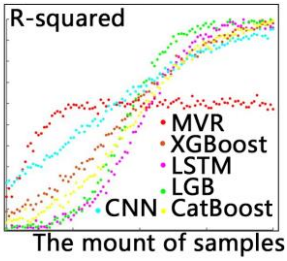


Fig. 9. The comparison chart of the time complexity of different algorithms

Figure 9 shows our tests of six different methods using the Monohulled Sailboats dataset of over 2300 entries. The index used is R-squared, and the closer the index is to 1, the better the prediction effect is.

There are many factors to consider in choosing the optimal solution. For example, some data in Fig.8. basically reflect the accuracy. Under such complex problems, MVR, CNN and CatBoost are obviously too low in accuracy and are not applicable. At the same time, since we do not have enough data on listed sailboats, the accuracy changes with the data is also one of the factors we have to consider. In addition, time costs must also be considered. In fact, sensitivity should also be considered when evaluating the model we use. This part is carried out in the sensitive analysis section.

The three indicators of sensitivity and accuracy, time complexity and dependence on data (quantified by growth rate) are used as parameter factors, and the entropy weight method is used to evaluate their importance, and finally the evaluation indicators of different algorithms are given.

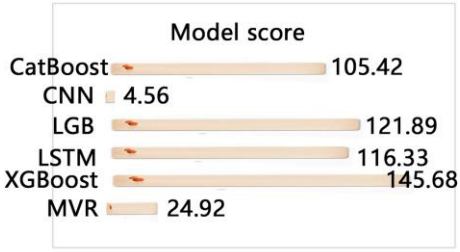


Fig. 10. Model score

The data in the Fig.10. has no dimension, but a "score value" obtained after weighting according to the results of the entropy weight method. The higher the score, the more accurate the model is. It can be seen from the results that under the existing conditions, XGBoost, LSTM, and LGB can all get good results. Therefore, when the time cost is not very high, we choose the XGBoost model with higher accuracy is more reasonable.

4 DISCOVERY OF INNOVATIVE CONCLUSIONS

4.1 Conclusions that can be drawn from an economic point of view

The Caribbean is a place where sailing activities are popular, there will be no shortage of sellers, and the local sailing prices are indeed the lowest among the three places. In terms of time, sailing events in the northern hemisphere winter are the most intensive. Therefore, for those customers with tight budgets, avoiding these time slots and choosing the Caribbean is a very cost-effective choice.

In recent years, sailing competitions have gradually flourished, and sailing enthusiasts have also begun to pursue longer and more exciting races. Therefore, the trading volume of large-scale, high-cost sailing boats has increased sharply. Market operators and manufacturers can combine the predictions in the model According to the consumption level and market demand, reasonably manufacture the type of sailboat that can better cater to customers, so as to maximize profits.

As a non-essential item, sailboats have extremely poor adaptability to emergencies. For example, the transaction volume and transaction price of the two sailboats were greatly affected by the economic crisis around 2008, so it is not difficult to predict that in 2020, affected by the global epidemic, the price of sailboats will usher in another round valley value.

4.2 How does a sailboat increase its selling price?

Shown in Fig.11. are the factors that may affect the price of sailing boats analyzed based on our model forecast data. Customers can choose the option that suits them best based on the value of

a boat, avoiding flashiness. At the same time, manufacturers should also focus on designs that are practical and cost-effective.

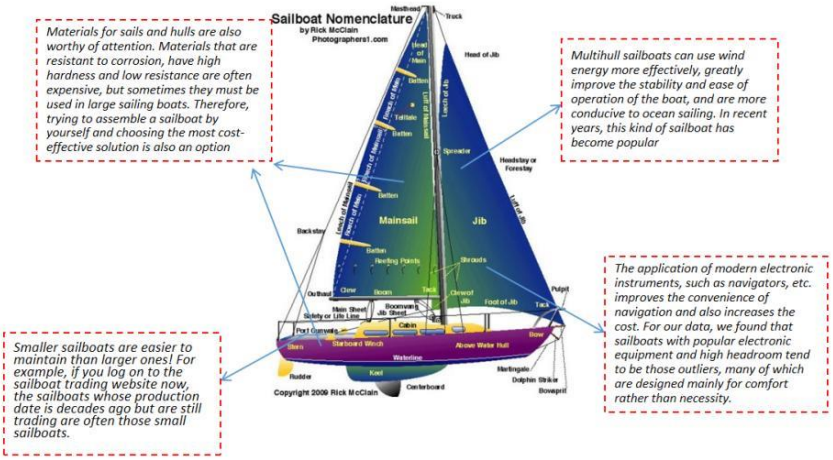


Fig. 11. Properties and prices of sailboats

5 MODEL ANALYSIS

5.1 Strengths and Weaknesses

5.1.1 Strengths

1） In terms of data processing, we conducted data cluster analysis and evaluated the importance, discarding unimportant factors to avoid interference with predictions.

2） In the process of forecasting, considering that region and time are macro factors that are different from the price of a single sailboat, separating the time series forecast from the forecast of other related factors can improve the accuracy of the forecast.

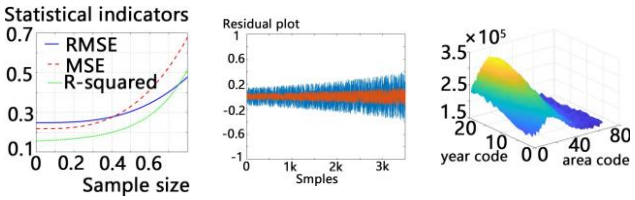
3） For different forecasting algorithms, we evaluated their reliability and time cost in predicting the results, and comprehensively considered all aspects to give an optimal algorithm model.

5.1.2 Weakness

1） The training of forecasting algorithms requires a large amount of data support, and for additional information, especially the lack of data information in Hong Kong, the forecast in Hong Kong has to rely on data from other regions.

2） Unable to consider unstable factors such as modern technology, which will cause some data to be considered as abnormal data and cleaned.

5.2 Sensitivity Analysis



(a) (b) (c)

Fig. 12. Picture of sensitivity analysis

Fig.12.(a) is the change diagram of the three statistical indicators MSE, RMSE and R-squared of the XGBoost model as the amount of data changes. The graph line is the result of fitting according to the data points. The unit of the abscissa is thousands. It can be seen from the figure that within the range of 0-10000 data, the accuracy of model prediction is very sensitive to the data. Data reading, the more accurate the prediction.

Fig.12.(b) is a residual line graph before and after adding singular disturbances. According to the model, although some excessive prices can be identified by the box plot, we cannot conclude that these are outliers. It can be seen from the figure that after adding singular disturbances, the prediction accuracy of the XGBoost model is greatly reduced, but the change is relatively stable. Note that for the sake of intuition, the results have been processed in ascending order before visualizing the results.

Region and time are treated as dummy variables in the forecasting process and exist in coded form. But the problem is that the change of coding is uniform, but the change of region and time is uneven for economic development, market and other factors. It can be seen from the figure that the price in the same area does not change much over time, basically changing evenly. Prices vary dramatically across regions. Here, in order to avoid the shock of the three-dimensional map being too violent, the local annual gross GDP value is used when coding the region.

5.3 Conclusion

Nowadays, with the development of the economy and the pursuit of a higher quality of life, sailing is becoming more and more popular. Therefore, how to choose a suitable sailboat has become a concern of sailboat enthusiasts and manufacturers. The prediction of the sailboat market is of great significance to consumers' purchasing strategies and the direction of future sailboat manufacturing and design.

In this project, we first carefully screened the data set, and carried out the construction of new features in feature engineering and discarded unimportant features. We then use the time-step difference equation and the XGBoost model to predict changes in market prices over time series and other single sailboat factors, and apply the model to the sailboat market in Hong Kong. Finally, we build an evaluation system to evaluate the optimal prediction schemes under different data sets, and give other innovative and instructive conclusions discovered during the research.

REFERENCES

[1] Meng Yingzhu, Sun Shengnan. Real estate price forecasting based on ARIMA time series model--shenyang city as an example[J]. Neijiang Science and Technology,2021,42(05):65+74.

[2] Ge Na, Sun Lianying, Zhao et al. Analysis of sales volume forecasting based on ARIMA time series model[J]. Journal of Beijing Union University,2018,32(04):27-33.DOI:10.16255/j.cnki.ldxbz.2018.04.006.

[3] Xiao, N., Kan, I., Li, C. F.. Parameter selection of Xgboost based on local search Bayesian algorithm[J]. Journal of Zhongnan University for Nationalities (Natural Science Edition),2023,42(02):201-207.DOI:10.20056/j.cnki.ZNMDZK.20230209.

[4] Chi, Xinyang. Research on car sales forecasting based on web search index and PSO-SVR model [D]. Northeast University of Finance and Economics, 2021. DOI:10.27006/d.cnki.gdbcu.2021.000137.

[5] Li Zhaobin, Ye Jun, Zhou Haoyan et al. A rough K-means clustering algorithm for variational firefly optimization [J/OL]. Journal of Shandong University (Engineering Edition):1-10 [2023-06-09]. http://kns.cnki.net/kcms/detail/37.1391.T.20230606.1640.006.html.

[6] Li-min L ,Chao-yang W ,Zong-zhou W , et al.Landslide displacement prediction based on the ICEEMDAN, ApEn and the CNN-LSTM models[J].Journal of Mountain Science,2023,20(05):1220-1231.