Research and Analysis Based on Pricing of Used Sailboats in the World and Hong Kong

Yi Ding 1, *, Zhiji Wang 2, Jiacheng Sun 2

1 School of Economics and Management, Dalian University of Technology, Dalian, China, 116024
2 Faculty of Electronic Information & Electrical Engineering, Dalian University of Technology, Dalian, China, 116024

* Corresponding Author Email: ddddd_999@163.com

Abstract. In order to better complete the evaluation of used sailboat prices, this article succeeded in having a system that can predict the prices of used sailboats and combined various analytical methods to realize the prediction of Hong Kong sailboat prices. This article focuses on the following issues: identifying mathematical models for rational pricing, explaining the effect of regional variables on prices, and simulating the regional impact of sailboat prices. This article consists mainly of the following: First, neural networks and deep learning related models are developed to measure the impact of sailboat characteristics on their pricing from multiple perspectives, and predictions are made for the used sailboats based on these models. Second, a variety of geographical factors are taken into account to analyze the correlation of regions, and regional variables with strong correlation analysis are added and combined with BiLSTM-AT model to explain the influence of regions on listing prices. Then, data on the corresponding variables in Hong Kong were collected, and cluster analysis was performed on regionally relevant factors to construct a multi-regional cluster price model. Last, sensitivity analysis and robustness analysis are performed on the completed model.

Keywords: Used Sailboats; BiLSTM-AT; Correlation Analysis; Cluster Analysis.

1. Introduction

1.1. Background

Sailboats were originally used by the ancient inhabitants of the sea area and were once one of the main means of transportation at sea, using the Bernoulli Effect to move forward. Nowadays, with the development of the technological era, sailing ships, which are mainly powered by wind, are gradually replaced by ships using oil and speedboats using electricity, etc.

At this stage, the use of sailboats for leisure and recreation is often recognized as a high-end activity. His ownership is very low, and most of the owners are highspending people, making it an expensive luxury.

Since the price of a sailboat is influenced by many factors, brokers are often unable to assess the price of a used boat very accurately, so a model is needed to fully consider various factors and form a report to help brokers accurately assess the price of a used boat.

1.2. Literature Review

From the used sailboat market, we can associate the used car market, which is more typical of the used market. In conducting the literature reading related to the used car market, we categorize the literature therein in two parts:

The first part is about the model construction and learning using artificial neural networks and machine learning related content, and the following conclusions were drawn: some literatures used grey relational analysis combined with the particle swarm optimization algorithm to optimize the traditional BP neural network, and put forward the used car price prediction method based on PSO-GRA-BPNN (Liu et al. 2022) [1]; Some literatures constructed an ANN artificial neural network by using Keras regression algorithm and other machine learning algorithms (Varshitha et al. 2022, January) [2]; some literature preprocesses the dataset through Python's Pycaret package and compares...
the performance of each algorithm through the machine learning algorithms comparison function (Wang et al. 2021, December) [3]; some articles use machine learning algorithms to build a statistical model based on the provided data with a given set of attributes (Sharma and Sharma 2020) [4]; and others propose a price prediction based on the Bayesian neural network with latent factors on a cloud platform model (Huang et al. 2023) [5].

The second part is about using other models to explore and analyze the pricing factors in the used car market, and we summarize that: some literatures propose an iterative framework combining XGBoost and LightGBM (Cui et al. 2022) [6]; some literatures utilizes the random forest method to reduce the information asymmetry in the used car market (Bies et al. 2021, September) [7]; and some literatures have proposed random forest prediction model, GBDT prediction model and SVM prediction model that predict and compare used car prices (Xu et al. 2023, March) [8]; some literature uses python, flask, and HTML as well as linear regression and lasso regression to create models that utilize machine learning to predict used car prices (Mukharjee et al. 2023) [9]; and others identify the best predictive model through heuristic algorithms (Bilen 2021) [10].

From the above study, we can construct relevant artificial neural networks, choose BP neural networks, LSTM models and other relevant algorithms, and use deep learning to explore the pricing factors of used sailboats, and at the same time derive them against the relevant pricing factors in the used car market.

2. Assumptions

Considering that practical problems always contain many complex factors, first of all, we need to make reasonable assumptions to simplify the model:

(1) It is assumed that the samples with more missing feature data provide less information on the transaction price of used sailboats and will not affect the modeling.

(2) All transaction prices change only because of the model, local geography, affordability, etc. and are not affected by human factors.

(3) The data after random forest and linear interpolation are valid and credible, and can reflect its basic laws.

(4) Objective data such as the model parameters of the ship do not change due to the region.

Additional assumptions are made to simplify analysis for individual sections. These assumptions will be discussed at the appropriate locations.

3. Notation

Some important mathematical notations used in this article are listed in Table 1.

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Description</th>
<th>Symbols</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega$</td>
<td>Weights</td>
<td>$t$</td>
<td>Vector of candidate values</td>
</tr>
<tr>
<td>$f_1$</td>
<td>The number of the broods</td>
<td>$c_{i-1}$</td>
<td>Last cell status i</td>
</tr>
<tr>
<td>$f_2$</td>
<td>BP output layer function</td>
<td>$b$</td>
<td>Bias term for cell state</td>
</tr>
<tr>
<td>$\theta_{ij}$</td>
<td>Threshold</td>
<td>$b_o$</td>
<td>Bias term for the output gate</td>
</tr>
<tr>
<td>$c_i$</td>
<td>Memory cells</td>
<td>$f_i$</td>
<td>The value of forgotten door</td>
</tr>
<tr>
<td>$x_t$</td>
<td>Present moment input</td>
<td>$b_f$</td>
<td>Bias term of forgotten door</td>
</tr>
<tr>
<td>$\overline{h_t}$</td>
<td>hidden states of encoder</td>
<td>$h_t$</td>
<td>Current implied layer state</td>
</tr>
</tbody>
</table>

Note: There are some variables that are not listed here and will be discussed in detail in each section.
4. Model Preparation

4.1. Data Source


From the above website we obtain the reference price corresponding to the conditions of origin, manufacturer, trafficking area, etc. of the sailboat deal, etc. In the end, a total of 8802 items of sailboat information and the corresponding prices were compiled.

4.2. Digitization of category variables

We use one-hot coding to digitize Make, Variant, GeographicRegion, District and MakeVariant to transform the category variables into a form that can be easily utilized by machine learning algorithms.

4.3. Data supplementation

(1) We obtained detailed data on sailboatdata.com by means of a web crawler including eleven metrics. Meanwhile, we use linear interpolation to realize the interpolation.

(2) Considering the practical applicability and completeness of the model, we introduced a multiple substitution approach to deal with missing values.

(3) To evaluate the validity of the model, we chose commonly used performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE) applied to continuous variables and Accuracy applied to discrete variables.

5. Model I: LSTM model establishment and result

We use deep learning algorithms for learning a large amount of data and compare the fitting results of multiple algorithms: BP neural network, LSTM, BiLSTM, BiLSTM-AT and WOA-BiLSTM.

5.1. Model I Establishment

(1) BP Neural Network: BP neural network is a multilayer feedforward network trained by error back propagation algorithm, consisting of input layer, intermediate layer and output layer. Using the fastest descent method, the network weights and thresholds are adjusted by back propagation. The output value corresponding to the minimum mean squared deviation of the network is the predicted value.

(2) Long Short-Term Memory networks: Long Short-Term Memory networks (LSTM) is a structure and extension of recurrent neural network (RNN), which can selectively forget or memorize information by memory gates for each memory cell, making it have better memorability, avoiding the difficult problems of long time dependence, gradient explosion and gradient disappearance. Its main structures include: forgetting gate, memory gate and output gate. The main formulas are:

\[ f_t = \sigma(\omega_f [h_{t-1}, x_t] + b_f), \quad i_t = \sigma(\omega_i [h_{t-1}, x_t] + b_i) \]  

\[ c_t = f_t * c_{t-1} + i_t * \tilde{c}_t, \quad \tilde{c}_t = \tanh(\omega_c [h_{t-1}, x_t] + b_c) \]  

\[ h_t = o_t \tanh(c_t), \quad o_t = \sigma(\omega_o [h_{t-1}, x_t] + b_o) \]

(3) BiLSTM: BiLSTM can capture dynamic information from the front and back segments sequentially. BiLSTM hidden layer structure consists of forward LSTM layer and backward LSTM layer, such a bi-directional structure provides complete past and future contextual information to each node in the input sequence of the output layer, and uses both information simultaneously to obtain the output. The main formulas are:
\[ h_i = f(\omega_1 x_i + \omega_2 h_{t-1}), \quad h'_i = f(\omega_3 x_i + \omega_4 h_{t-1}') \]

\[ o_i = g(\omega_4 x_i + \omega_5 h'), \quad (4) \]

\[ h_{t+1} = g(\omega_6 x_i + \omega_7 h_{t-1}'), \quad (5) \]

\( (4) \) BiLSTM-Attention: We further extend the BiLSTM model by combining the BiLSTM model with the Attention mechanism to construct the BiLSTM-Attention model.

### 5.2. Model I Solving and Result

After establishing the model, we trained the above BP neural network with LSTM, BiLSTM, BiLSTM-AT, and WOA-BiLSTM, and for data, we selected the data ensemble generated after preprocessing all monohull and catamaran mergers for training to obtain its prediction results, and also used mae, rmse, mape. Figure 1 is the result.

![Figure 1. Training.](image-url)
Using prediction plots and loss function images, we set the number of iterations of the bp neural network and LSTM to twenty and compare their error metrics and get Figure 2.

A comparative analysis was performed through the histogram, and it was found that the best learning effect in the model building process, was the BiLSTM-AT method with the highest fit, so in this prediction model, BiLSTM-AT was chosen to build the price prediction model.

6. Model II: Correlation Analysis

Next we analyze the impact of region on the listed price, where we choose a variety of different indicators to measure the re-gional factor and filter them by their different effects on the correlation, leaving five final the final five indicators were selected.

6.1. Model II Establishment

We use the KMO test, Bartlett’s sphericity test and Variance explanation rate for correlation analysis.

6.2. Model II Solving and Result

The data were subjected to KMO test and Bartlett test:

<table>
<thead>
<tr>
<th>KMO test and Bartlett’s test</th>
<th>KMO value</th>
<th>Bartlett’s sphericity test</th>
<th>Approximate cardianlity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.837</td>
<td></td>
<td>42844.922</td>
</tr>
</tbody>
</table>

From Figure 3, the results of the KMO test showed that the value of KMO was 0.837, while the results of the Bartlett’s spherical test showed that the significant obtained a p-value of 0.000***, which showed significance at the level and rejected the original hypothesis that there is a correlation between the variables.

Also for each factor on the absolute value of the price correlation coefficient, Also in combination with the rolling stone figure:
From Figure 4, it was found that it entered the smooth period after eight factors entered the smooth period, and considering that the eighth factor birth rate was less relevant to the listing prize, the first seven factors were finally determined as the optimal number of elements.

7. Model III: Cluster Analysis

For the regional characteristics of Hong Kong, we have done a cluster analysis based on the geography and environment of Hong Kong, and divided Hong Kong into ninety other countries and regions, so as to obtain the influence of geography and environment on used sailing boats in similar regions of Hong Kong.

7.1. Model III Establishment

We choose K-means cluster analysis to achieve our purpose.

7.2. Model III Solving and Result

We selected District, whether it is near the sea, average daily minimum temperature, average daily maximum temperature, and total national GDP per capita GDP as variables for cluster analysis.

The common selection method is elbow method, which is a method to confirm the optimal K value using the relationship graph of SSE value and K value. We drew Figure 5, which is the following comparison graph of the number of clusters by the elbow rule:

![Figure 5. Process of random forest method.](image)

All the levels present significance and rejects the original hypothesis, indicating that the variables have significance differences between the categories divided by cluster analysis.

We then summarized, resulting in a clustering scatter plot of Figure 6:

![Figure 6. Scatter plots.](image)
For the clustering analysis, it is obtained that Hong Kong and France, Florida, Germany and other eighteen countries have similarity in regional environment, and their similarity reacts to the same pricing method.

8. Sensitivity and Robustness Analysis

8.1. Sensitivity Analysis

We performed sensitivity analysis by varying the learning rate and data size. In the machine learning process of BiSLTM-AT, the team set the learning rate to 0.01 and varied it from -0.1 to +0.15. We found the model has good stability.

8.2. Robustness Analysis

It was found that the corresponding error changes were small with good robustness when the number of iterations was changed to a large extent.

9. Conclusion

First, we found that the BiLSTM-AT model has the smallest error, which is in line with expectations, and the BiLSTM-AT model is selected as the model basis for price prediction. Then, it is suitable for factor analysis and shows significant correlation and we screen 7 factors as the total number of indicators. What’s more, we plot the relationship between SSE and K-value to confirm that the optimal K-value is 5. Last, we found the model has good stability and robustness.

Our model can both determine the price range of a specific ship type in the corresponding region when the conditions are ambiguous, and also directly forecast the price when the indicators of all factors of the ship type are complete. The model performs cluster analysis for different impact methods, so that the same pricing strategy can be used for a collection of regions with similar pricing impact methods, which is convenient for brokers to handle. The weakness is that there is no combination of real-time market supply and demand factors, which can be combined with good buyer strategies or seller strategies in the subsequent modeling of combined market conditions.

References


