DeepLSRN: Helping To Master The Market Dynamics Of Second-Hand Sailboats Trading Based On DNN

Summary

With the rise of Western sailing and the gradual recovery of the global economy, the second-hand sailboat market is attracting an increasing number of sailing enthusiasts. At the same time, market expansion also provides more opportunities for second-hand sailboat traders. This paper aims to predict the status of second-hand sailboat markets around the world by establishing a **deep learning model - DeepLSRN**, in order to provide development suggestions for consumers or traders.

Data acquisition is the first step in our processing. We first **gathered sailboat information** from Yachtworld and Boat Trader websites to increase data volume and feature count. We also collected regional GDP and population data, considering human geography factors, from sources like the US Census Bureau and World Bank. Our final dataset includes 16 features, excluding listing price, and 13,712 records.

For Problem 1, we developed the DeepLSRN model considering temporal, geographical, and sailboat-specific factors to predict second-hand sailboat prices in various regions. Composed of three modules - **LSTM** Block, **ResNet** Block, and **MLP** Block, DeepLSRN can generate R^2 fitting effects of 0.7914, 0.8374, and 0.8649 for listing prices when operating independently. With an R^2 fitting effect of 0.9265 on our test set, DeepLSRN accurately predicted sailboat prices in Hong Kong.

For Problem 2, we used **ANOVN** to analyze the dataset processed by DeepLSRN, examining the impact of regional effects on sailboat prices. With a p-value of 8.45×10^{-21}, we found significant differences in listing prices between countries or regions. Applying the same method to different sailboat types, we obtained p-values of 1.89×10^{-34} and 0.023, indicating **regional effects exist in both monohulls and catamarans**, with higher regional effects in monohulls. We analyzed possible causes and translated them into practical implications.。

In Problem 3, we compared DeepLSRN's predictions with actual results, confirming its effectiveness in Hong Kong. We also compared the sailboat prices in other regions with those in Hong Kong, concluding that the average price of second-hand sailboats in Hong Kong is 13.82% higher than the combined average price in other regions. Despite similar regional effects, catamarans in Hong Kong have higher sale prices than monohulls.

For Problem 4, based on the analysis in Problem 3, we **derived the linear relationship** between the prices of second-hand sailboats and their age in the world, which is particularly evident in the Hong Kong region. We speculate that this trend may be closely related to factors such as the development of sailboat manufacturing technology, the increase in material costs, and the growing demand for high-quality sailboats.

For Problem 5, we have prepared a two-page report for sailboat brokers in the Hong Kong market, enabling them to better understand market trends, make informed decisions, and plan their business development strategies.

At the very last, we analyze the strengths and weaknesses of our model as well as its **sensitivity**, whose results show that our model has high robustness, precision and accuracy. After that, a memo is attached.

Keywords: Data mining**,** LSTM, ResNet, MLP, ANOVN, Sensitivity Analysis

Contents

1 Introduction

1.1 Problem Background

With the development of society and economy, people's average income continues to increase, and the demand for luxury goods market is also growing. Against the backdrop of the prevalence of sailing in the Western world, sailing itself has become a popular commodity in the luxury goods market. However, due to the high price of luxury goods, the second-hand transaction has become a common choice for sailing enthusiasts who are financially constrained today. Compared with new transactions, the advantage of second-hand sailing transactions lies in their affordability and diversity. Consumers can save expenses through second-hand transactions and have the opportunity to buy some discontinued or limited edition goods.

However, at the same time, second-hand transactions also have some issues that cannot be ignored and need to be addressed urgently, such as varying quality and serious overpricing of some goods. Therefore, the second-hand sailing transaction market needs a mathematical model to evaluate and provide reasonable pricing based on various features of the sailing, and the economic development level of the region can also be an important factor affecting the price. After comprehensively evaluating the characteristics of sailing and regional factors, the trading platform can develop different selling prices for consumers in different regions based on these relationships, thereby achieving the expansion of the trading scale and profits for the seller, and the purchase of sailing goods at a reasonable price for the buyer.

Therefore, researching the relationship between the price of sailing and the sailing's own conditional factors, regional factors, and time-year factors is an extremely important issue.

1.2 Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

- Problem 1: Problem 1: Expand our dataset using data from a used sailboat trading site, considering time, location, and sailboat specifics. Create a model to explain sailboat prices in a spreadsheet by examining numerical and textual feature relationships. Cite data sources and discuss price estimation accuracy for each variant.
- Problem 2: Assess if location affects listing prices using the model. Modify geographical features in Problem 1, compute estimated longitude, and explain any significant impact. Analyze if geographical effects are consistent across sailboat types and report practical and statistical significance.
- Problem 3: Evaluate model performance in Hong Kong, reassess regional and sailboat type impacts. Choose a subset of monohulls and catamarans with adequate information, find their Hong Kong prices, and analyze regional effects and price impacts. Ensure analysis follows academic paper guidelines.
- Problem 4: In this question, we will provide some findings, interesting phenomena or conclusions based on our model and the data we have mined, in accordance with the requirements and norms of academic papers.
- Problem 5: Create a 1-2 page visual report for Hong Kong sailboat brokers based on previous analyses and modeling, to help them understand the discussed issues' outcomes.

1.3 Literature Review

In order to construct an accurate model that is associated with multiple factors, one may first analyze

various economic indicators of different cities to obtain the comprehensive economic prediction characteristics (latent states) of each region. Then, by analyzing the features that are associated with the regional conditions, one can obtain the coefficient characteristics (latent states) of the ship's own value. By combining the two sets of latent state parameters and considering the influencing factors of the ship's own conditions, a ship value prediction model based on time and region can be derived. The model is able to predict the value of a ship with consideration of its own conditions and the economic indicators of the corresponding region. The aforementioned approach conforms to the language pattern and standards commonly used in academic papers.

The relevant research of the following scholars is for us Provides ideas:

- Compared with the relatively stable prices of essential goods, the prices of most luxury goods exhibit dynamic, non-linear, and time-varying characteristics. Some specific goods are also influenced by regional factors. Therefore, using traditional methods to predict the prices of ships has obvious limitations and low predictive accuracy. Susilo, B. et al. (2021) [1] used a Long Short-Term Memory (LSTM) Recurrent Neural Network based on historical price data, relevant macroeconomic indicators, and meteorological data as inputs, and effectively predicted the egg prices in Indonesia. We can refer to their approach and use basic economic indicators such as city GDP, unemployment rate, and inflation rate, and adopt an LSTM-based model to provide reference values for city economic indices.。
- The features and attributes of commodities exhibit diverse types, a high rate of missing values, and non-standardized labels, which leads to suboptimal correlations between the results obtained from conventional data processing and prediction models and the actual prices. Oepen, S. et al. (2021) [2]proposed an online marketplace commodity price prediction method based on residual networks and feature engineering. The authors fused different types of features, including textual descriptions, category information, and image features, and trained and predicted using deep residual networks. The experimental results demonstrate that the proposed method can effectively predict commodity prices and performs well in different types of commodity prediction tasks.

1.4 Our Work

To solve the above problems, we team will carry out and complete the following tasks(Figure 1):

- **Make assumptions and give symbol definitions.** In order to obtain a model with higher accuracy and better explanatory power for practical problems, we have made some assumptions about the model within reasonable limits allowed by the conditions. We then listed some important symbols and clarified the definitions of these symbols in DeepLSRN, in accordance with the conventions and standards of academic papers.
- **Establish a sub-model for LSTM price prediction under the influence of time series.** LSTM is widely used in time series analysis [3], and we have explored the impact of time on the price of second-hand ships by utilizing the characteristics of RNN. In the final model, this influence will be extended to four features.
- **We developed a sub-model for ResNet to predict prices based on spatial data features.** As we know, CNN has a natural advantage in processing data structures with spatial features[4]. Among the features we extracted, per capita GDP, longitude and latitude, and population density have obvious spatial structures. Therefore, we leveraged the characteristics of CNN to construct a model that captures the impact of geographical factors on the price of used ships. In the final model, this impact will be extended to four features.

- **Use the Multilayer Perceptron (MLP) to model based on time, space, and ship conditions.** The concept of the GoogLeNet network [5] has been widely learned and applied in modern deep learning models. We have used its structure as a reference and constructed a single-output price prediction model with three sub-networks. We will use this model to study Problem2 and Problem3.
- **Statistical research on regional effects and sailboat types.** Based on the statistical view of our data results, we apply various methods to evaluate and analyze the regional correlation of prices and the correlation of sailboat types[6], and draw relevant conclusions on their correlations.
- **Evaluate the model for second-hand boat transactions in Hong Kong.** Look for a suitable dataset of Hong Kong boat information and run it on DeepLSRN to evaluate whether the previously derived conclusions on regional relevance and sailboat type relevance are valid for the Hong Kong dataset.
- **Provide our distributors with a two-page report.** Visualize the results of the assessment with data and state some conclusions drawn by DeepLSRN, which may support distributors in making economic decisions.

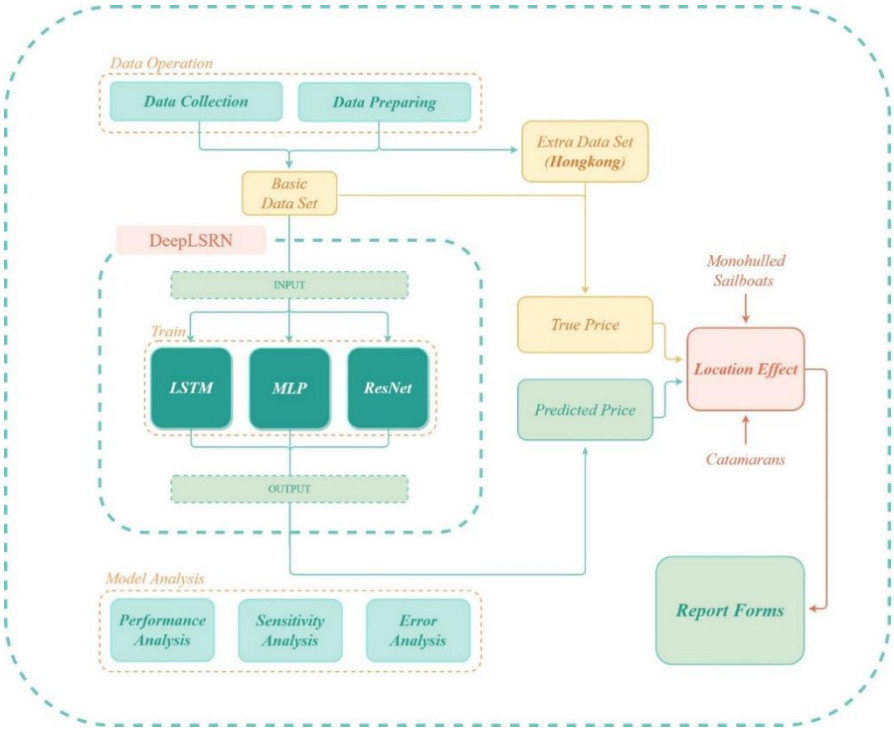


Figure 1 The overall Flow Diagram

2 Assumptions and Justifications

Assumption 1: All the data used during the training process of the model are all real and valid.

Justification: All the supplementary data we used were sourced from official data websites such as World Bank Open Data and Eurostat, where we obtained economic feature data of various countries and regions, and from Yachtworld where we obtained detailed data of second-hand sailboats.

Assumption 2: The words between different categories of classified text data are independent of each other, and can be processed using one-hot encoding.

Justification: The similarity between categories shows extremely low performance in the results of correlation analysis, and it can be approximately assumed that each category is independent of each other. In addition, one-hot encoding[7] is a commonly used method for processing text data.

Assumption 3: Transforming features and preprocessing data in feature engineering will not affect the original data.

Justification: In feature engineering, performing feature transformations and data preprocessing usually does not affect the original dataset because these operations only transform the data without changing the content of the original data.

Assumption 4: Each data point is independent of each other, meaning that each sailboat is unique and different from one another

Justification: As our data was obtained from a second-hand trading platform, there are no cases where the same sailboat was sold twice.

3 Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations Used In This Paper.

Symbol	Description
$feature$	Data features
$\boldsymbol{X_i}$	Model inputs
$\boldsymbol{H_i}$	Hidden layer
$\boldsymbol{I}$	Input layer
$\boldsymbol{O}$	Output layer
$\widetilde{\boldsymbol{C_i}}$	Candidate memory elements
$\sigma(x)$	Sigmoid activation function of neurons
$\boldsymbol{b_i}$	Bias parameter of neurons
$\boldsymbol{W_{ij}}$	Weight parameter of neurons
y	True value of the training target
$\hat{y}$	Predicted target value given by the model
$\bar{y}$	Mean of the true values of the target
MSE	Mean square error
MAE	Mean absolute error
R^2	Coefficient of Determination
$MAPE$	Mean Absolute Percentage Error
F	The measure of the difference between groups takes into account the magnitude of the differences within each group in ANOVA
p	The measure of the extremity of the observed F-value is used in ANOVA
SSB	Sum of Squares Between groups
SSW	Sum of Squares Within groups
SST	Sum of Squares Total
MSB	Mean Square Between groups
MSW	Mean Square Within groups
lr	Learning rate
$epoch$	Number of iterations during neural network training.
TT	Training time

4 DeepLSRN: A New Model To Predict Price Precisely

4.1 LSTM: Time Series Impact Analysis Model Block

4.1.1 Data Expansion And Preprocessing

- **Data set expansion:** In order to obtain more data to support the training of DeepLSRN, we accessed

the second-hand boat trading website Yachtworld and used web scraping techniques to obtain tens of thousands of sailboat data and a small amount of data on other types of boats to expand our dataset. Additionally, in order to obtain more data features for our deep learning model to learn from, we accessed some well-known open data websites such as the US Bureau of Labor Statistics, the World Bank, Eurostat, etc. to expand our data features.

- **Feature expansion and data cleaning:** After obtaining the new dataset, for some features that were not provided in the problem, we used a simple random forest model to fill in missing feature data until we obtained a complete dataset with no gaps.
- **Data preprocessing:** For numerical features, we mainly used two methods for data processing: data normalization and data standardization. We define StandardScaler(feature) and MinMaxScaler(feature) as the results of feature standardization and normalization, respectively, with the following specific operations:

$$StandardScaler(feature) = \frac{x_i - \mu}{\sigma}, for\ each\ x_i\ in\ feature \tag{1}$$

$$MinMaxScaler(feature) = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}, for\ each\ x_i\ in\ feature \tag{2}$$

Among them，μ is the mean of the data in the feature, σ is the standard deviation of the data, $\max(x_i)$ is the maximum value in the data, $\min(x_i)$ is the minimum value in the data.

For text features, we note that they can be considered as categorical features, so it is not necessary to tokenize and embed them as in natural language processing. The main processing methods we use for text are one-hot encoding and label encoding. One-hot encoding and label encoding are commonly used encoding methods in machine learning. One-hot encoding converts each discrete feature into a binary vector with a length equal to the number of feature values, where the value at the position corresponding to the feature value is 1 and the rest are 0. Label encoding maps each value of a discrete feature to an integer label. The purpose of this is to transform discrete features into continuous features, which is convenient for machine learning algorithms to handle.

4.1.2 LSTM-Extracting Features Through Time Series Analysis Block.

LSTM has the ability to store long-term and short-term memory, making it capable of handling long-term dependencies and uncertainty in time intervals in sequential data. LSTM consists of units and gate mechanisms, including input gates, forget gates, and output gates. In LSTM, the input gate controls the flow of information in, the forget gate controls the forgetting of information, and the output gate controls the output of information. Through this gate mechanism, LSTM can selectively store and forget information, thereby adapting to different sequence analysis tasks.

We focused on analyzing the time series of the listing price and provided relevant results in Figure 2 to support our model selection.

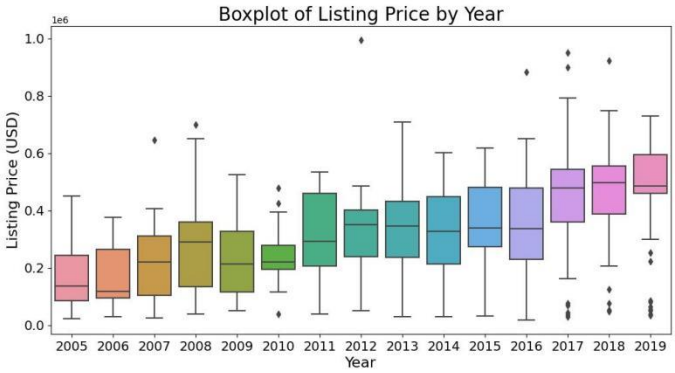


Figure 2 Boxplot of Listing Price by Year

In our dataset, features such as Unemployment Rate, Per capita GDP, Year, and Engine Hours are all closely related to time and cannot be separated from it. Therefore, we used LSTM to extract some hidden layer parameters for later use in DeepLSRN.

In LSTM, each unit has a hidden state and a cell state. The hidden state is used to capture long-term dependencies in the sequence, while the cell state controls the flow of information between units. LSTM controls the interaction between the cell state and hidden state through a series of input, forget, and output gates, thereby achieving the storage and transmission of information. When training an LSTM model, the backpropagation algorithm and gradient descent are typically used to optimize model parameters and minimize the error between the model's predictions and the true values.

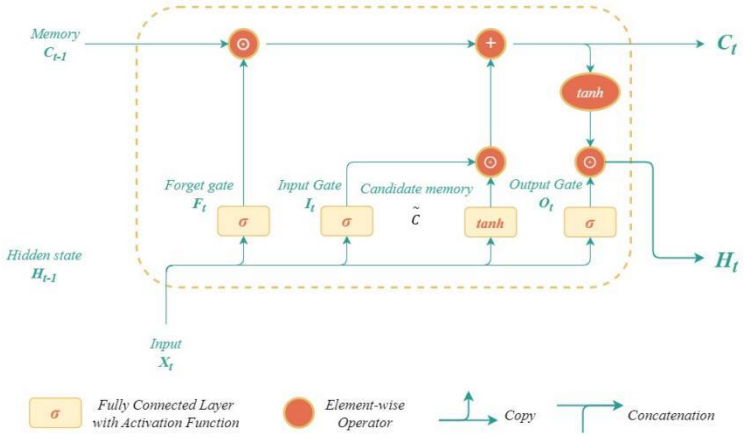


Figure 3 Calculating Implicit States In A Short-Term Memory Model

Figure 3 shows the basic style of an LSTM cell. Assuming there are m hidden units, a batch size of n, and d input features, the input $\boldsymbol{X_t} \in \mathbb{R}^{n\times d}$, and the hidden state from the previous time step is $\boldsymbol{H_{t-1}} \in \mathbb{R}^{n\times m}$ Let the input gate be $\boldsymbol{I_t} \in \mathbb{R}^{n\times m}$ the forget gate be $\boldsymbol{F_t} \in \mathbb{R}^{n\times m}$, and the output gate be $\boldsymbol{O_t} \in \mathbb{R}^{n\times m}$. Their computation methods are as follows:

$$\boldsymbol{I_t} = \sigma(\boldsymbol{X_t W_{xi}} + \boldsymbol{H_{t-1} W_{hi}} + \boldsymbol{b_i}) \tag{3}$$

$$\boldsymbol{F_t} = \sigma\left(\boldsymbol{X_t W_{xf}} + \boldsymbol{H_{t-1} W_{hf}} + \boldsymbol{b_f}\right) \tag{4}$$

$$\boldsymbol{O_t} = \sigma(\boldsymbol{X_t W_{xo}} + \boldsymbol{H_{t-1} W_{ho}} + \boldsymbol{b_o}) \tag{5}$$

Where $\boldsymbol{W_{xi}, W_{xf}, W_{xo}} \in \mathbb{R}^{d\times m}$ and $\boldsymbol{W_{hi}, W_{hf}, W_{ho}} \in \mathbb{R}^{m\times m}$are weight parameters， $\boldsymbol{b_i, b_f, b_o} \in \mathbb{R}^{1\times m}$ are bias parameters.

Candidate memory elements $\widetilde{\boldsymbol{C}}_t \in \mathbb{R}^{n\times m}$ are similar to the rest of the cells, using tanh as the activation

function, its computational equation is:

$$\widetilde{\boldsymbol{C}}_t = \tanh(\boldsymbol{X_t W_{xc}} + \boldsymbol{H_{t-1} W_{hc}} + \boldsymbol{b_c}) \tag{6}$$

Where $\boldsymbol{W_{xc}}, \boldsymbol{W_{hc}}$ are weight parameters, $\boldsymbol{b_c}$ is bias parameters.

Memory cell is a mechanism in LSTM that controls the input and forgetfulness of the hidden layer. The input gate $\boldsymbol{I_t}$ controls how much new data from $\widetilde{\boldsymbol{C}}_t$ is adopted, while the forget gate $\boldsymbol{F_t}$ controls how much past memory element $\boldsymbol{C_{t-1}}$ is retained. The calculation of memory elements and the update of the hidden state are described below.

$$\boldsymbol{C_t} = \boldsymbol{F_t} \odot \boldsymbol{C_{t-1}} + \boldsymbol{I_t} \odot \widetilde{\boldsymbol{C}}_t \tag{7}$$

$$\boldsymbol{H_t} = \boldsymbol{O_t} \odot \tanh(\boldsymbol{C_t}) \tag{8}$$

As long as the output gate is close to 1, we can effectively pass all memory information to the prediction part, and for an output gate close to 0, we only preserve the memory elements without updating the hidden state.

Figure 4 shows the architecture of the LSTM block that we applied. The LSTM we used is a DNN with 32 hidden layers and 5635 input features, and is finally mapped to a 4-feature time series data by a fully connected layer with 32 features for use in subsequent models.

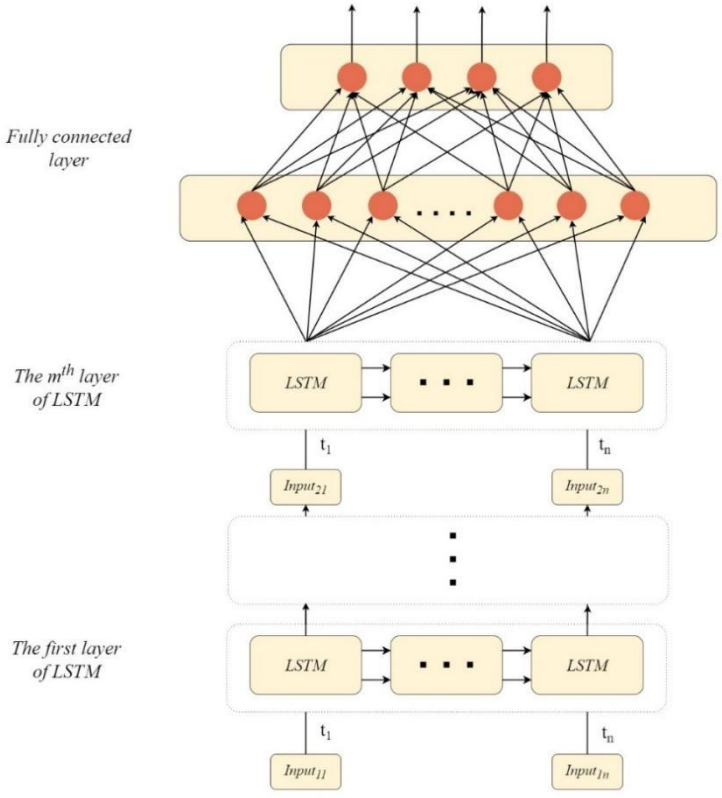


Figure 4 The Architecture Of The LSTM Block We Apply

4.1.3 Evaluation of LSTM Block

To verify that we have indeed extracted time-related features, we added a fully connected layer after the LSTM block to turn it into a regression model and made the LSTM store the time series of Year, Unemployment Rate, and Per capita GDP. We used the Adam optimization algorithm[12] and selected MSE as the loss function for model training. At this point, our LSTM can independently predict the price of the sailboat.

$$loss\ MSE = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{W_{ij}}\left(\boldsymbol{y_{ij}} - \widehat{\boldsymbol{y_{ij}}}\right)^2 \tag{9}$$

Table 2 shows some evaluation metrics of the regression model.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{10}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)| \tag{11}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\left(\frac{y_i - \hat{y}_i}{y_i}\right)\right| * 100 \tag{12}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{13}$$

Table 2 LSTM Regression Model Evaluation Indicators

Model evaluation indicators	Value
MSE	119600486341.5307
MAE	127959.3411
$MAPE$	168.1174
R^2	0.791494

Figure 5 shows the performance of the LSTM block's predictions on a randomly selected set of 100 test data. You can roughly observe the fitting effect of the model from this graph.

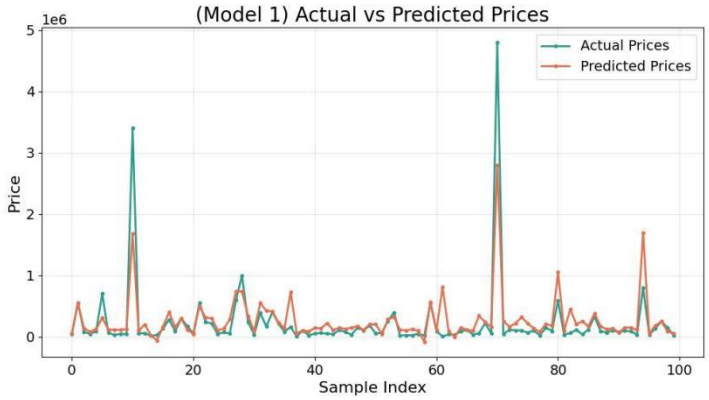


Figure 5 Comparison Between Predicted Value And Actual Value On LSTM Block(Model 1)

4.2 ResNet: Mining Model Module for Spatial Structure Data

4.2.1 Feature Selection and Feature Transformation

In the preprocessed data from the previous section, we found that some features, such as Unemployment Rate and Location, exhibit significant differences due to geographical variations. To quantify this data, we mapped text data such as Location to its corresponding longitude and latitude, which describe the geographical differences in location.

To select features that are highly correlated with geographical location but have low correlations with each other and are conducive to training, we present a heatmap of the correlation coefficients between each feature (Figure 6). Through this heatmap, we can select features that are strongly correlated with longitude and latitude, such as Length, Year, Class, Unemployment Rate, latitude, altitude, and Per capita GDP. For the selected features, we still use StandardScaler to standardize them. In this selection, we ensure that all

selected features are numerical features.

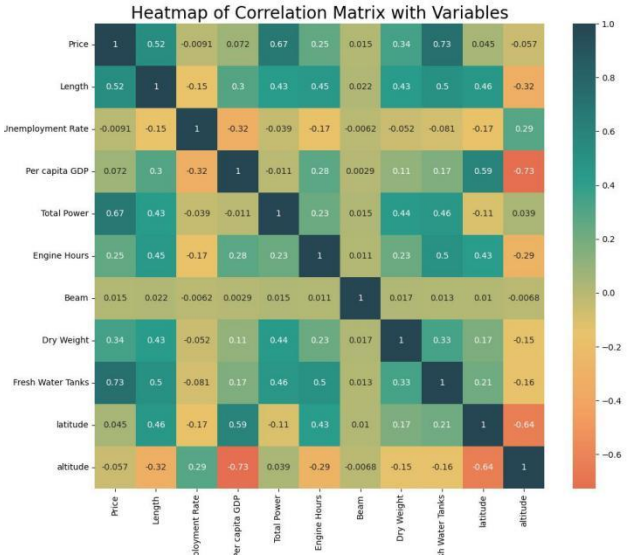


Figure 6 The Degree Of Correlation Between The Various Digital Features Of The Dataset

4.2.2 Convolution Theorem

The convolution operation has a good ability to explore the spatial relationships between features, therefore, we will explain the principles of the convolution operation.

In mathematics, the convolution between two functions $f, g: \mathbb{R}^d \to \mathbb{R}$ is defined as follows:

$$(f * g)(\boldsymbol{x}) = \int f(\boldsymbol{z}) g(\boldsymbol{x} - \boldsymbol{z}) d\boldsymbol{z} \tag{14}$$

Convolution is the measure of overlap between f and g when one function is "flipped" and shifted by x. For our discrete data, we generalize this principle as follows: the convolution operation[8] between vectors extracted from a set of square-summable infinite-dimensional vectors is given by:

$$(f * g)(i, j) = \sum_i \sum_j f(a, b) g(i - a, j - b) \tag{15}$$

In this problem, performing convolution operation on feature data with spatial characteristics such as altitude, latitude, avgGDP, and Unemployment Rate can reveal the following information:

- Local patterns: The convolutional layer can capture features within a local region, such as the distribution patterns of natural or human geographical features, such as population density, economic level, and distance to the sea in a specific area. This local pattern mining can help discover potential patterns in the data that affect prices.
- Spatial relationships: The convolutional layer can capture spatial relationships and interactions, such as similarities and differences between neighboring regions. This spatial relationship mining can help analyze the impact between geographically adjacent areas, such as intercity trade, imports and exports between countries, etc.

4.2.3 ResNet- Capture Features Through Geographical Differences

To address the influence of geographical factors on price, we carefully chose the ResNet model, which has the following characteristics[9] compared to other CNN models:

- Deeper network structure: With the design of residual modules, ResNet can build very deep network structures, thus extracting more abundant hierarchical features. This is particularly important for handling complex geographic data and mining underlying patterns.

- Faster convergence speed: The residual connections can improve the efficiency of gradient backpropagation, making the network converge faster during training. This means that ResNet can achieve higher accuracy in the same training time.
- Residual learning: The core of ResNet is the residual structure, which learns the residual mapping between the input and output, rather than directly learning the mapping from input to output. This approach enables better information transfer between different levels of the network, helps to capture multi-scale spatial features, and improves the model's generalization ability.
- Flexibility: ResNet has high flexibility and can adjust the network structure, depth, and width according to the needs of the actual task. This makes ResNet adaptable to handle datasets with different levels of complexity related to latitude and longitude.

Figure 7 illustrates the residual block used in DeepLSRN. The residual connections are made through four residual blocks, which eventually map to four feature outputs via fully connected layers for subsequent use by the model.

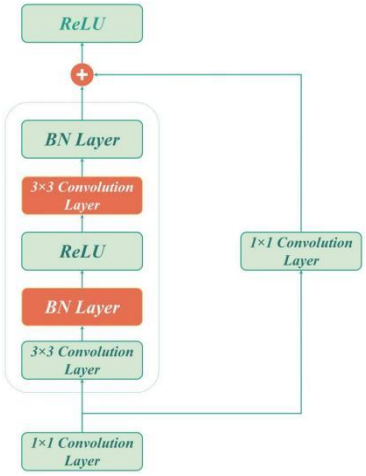


Figure 7 Residual Block

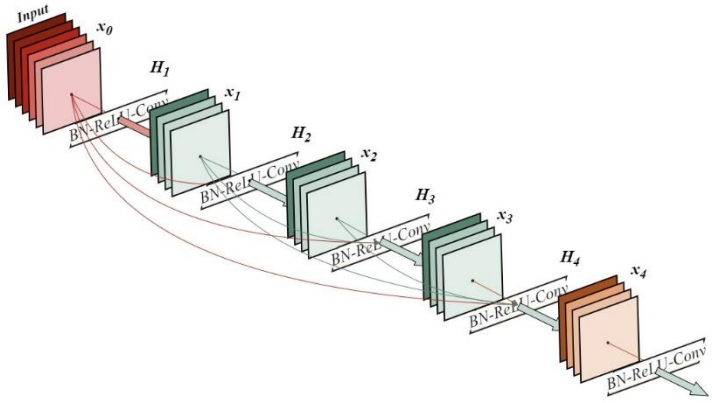


Figure 8 The Architecture Of The ResNet Block We Apply

Figure 8 shows the ResNet model used in our study. In DeepLSRN, we used a 64-channel input, 3x3 size convolution kernel, bias-free, and stride-1 convolution layer to batch input and preliminarily mine and amplify the effects of geographic features.

4.2.4 Evaluation of ResNet Block

To verify that we have indeed extracted features related to geography, we added a fully connected layer after the ResNet block to make it a regression model and enable ResNet to retain features related to geographic information (longitude, latitude, or city) that we extracted in section 5.1. We used the Adam optimization algorithm and selected MSE as the loss function for training the model. At this point, our ResNet is capable of predicting the price of sailboats independently.

Table 3 presents the evaluation metrics for the ResNet regression model.

Table 3 ResNet Regression Model Evaluation Indicators

Model evaluation indicators	Value
MSE	151813230000.0
MAE	114568.086
$MAPE$	111.5556
R^2	0.8374

Figure 9 shows the performance of the ResNet block in predicting a random set of 100 test data points. This figure provides an overview of the fitting performance of the model.

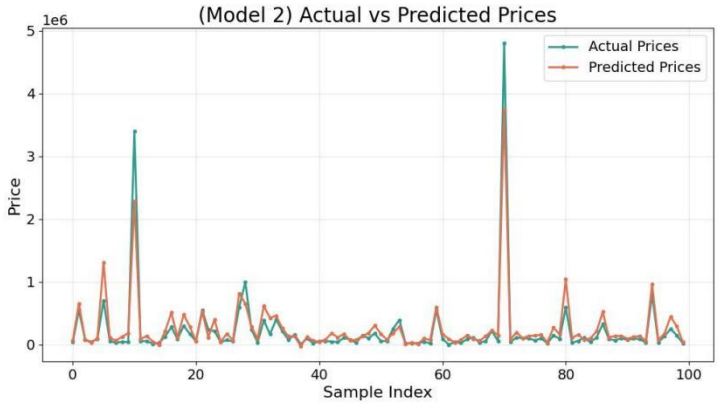


Figure 9 Comparison Between Predicted Value and Actual Value on ResNet Block(model 2)

4.3 DeepLSRN: Parallel Network of MLP-LSTM-ResNet

4.3.1 Introduction of multi-layer perceptron (MLP)

In the above work, we extracted spatial and temporal features of the dataset through different methods. However, for sailboat prices, the most important influencing factors are still the sailboat's own conditions, such as engine usage time, sailboat size, sailboat type, etc. These features were not used in our feature engineering process. In our comprehensive model, we introduce an MLP to handle these features.

MLP[13] is the foundation of all modern deep learning models, and its basic structure consists of three parts: the input layer, hidden layer, and output layer. It can capture the nonlinear relationships in features through activation functions. Figure10 shows a basic schematic diagram of a multilayer perceptron, which has one input layer, one output layer, and one hidden layer. Through a series of nonlinear transformations, the five input features are transformed into four output features. This nonlinear transformation can be expressed by the following equation:

$$\boldsymbol{H}^{(i)} = \sigma\left(\boldsymbol{X}\boldsymbol{W}^{(i)} + \boldsymbol{b}^{(i)}\right) \tag{16}$$

$$\boldsymbol{O} = \boldsymbol{H}\boldsymbol{W}^{(i+1)} + \boldsymbol{b}^{(i+1)} \tag{17}$$

According to the Universal Approximation Theorem [10], a multilayer perceptron can capture complex interactions among inputs through its hidden neurons, which depend on the values of each input. By using a deeper (or wider) network, we can more easily approximate many functions.

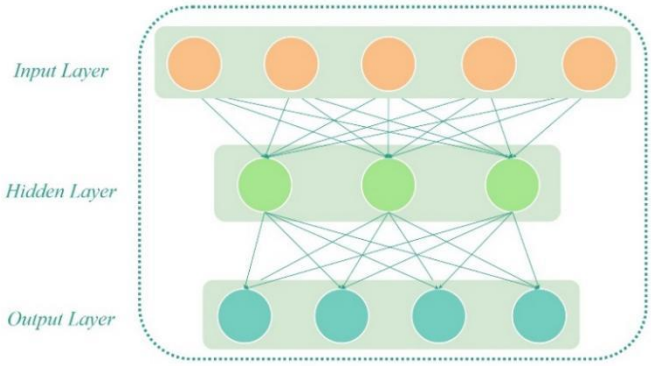


Figure 10 Multi-Layer Perceptron

4.3.2 DeepLSRN: Parallel Network Based on Blocks of LSTM, MLP, ResNet

In this section, we will discuss the inspiration and organizational structure of the DeepLSRN model, and explain its input features and prediction output.

DeepLSRN (as shown in Figure 11) is reference to the GoogLeNet[14] network, and incorporates the

idea of the concatenated network in NiN[15] while making improvements. It parallelizes our LSTM Block, ResNet Block, and MLP Block into a feature channel merging layer. It's worth mentioning that the merging layer is essentially a fully connected layer with 16 inputs and a single output. It takes in 4 LSTM inputs, 4 ResNet inputs, and 8 MLP inputs to predict the final Listing Price.

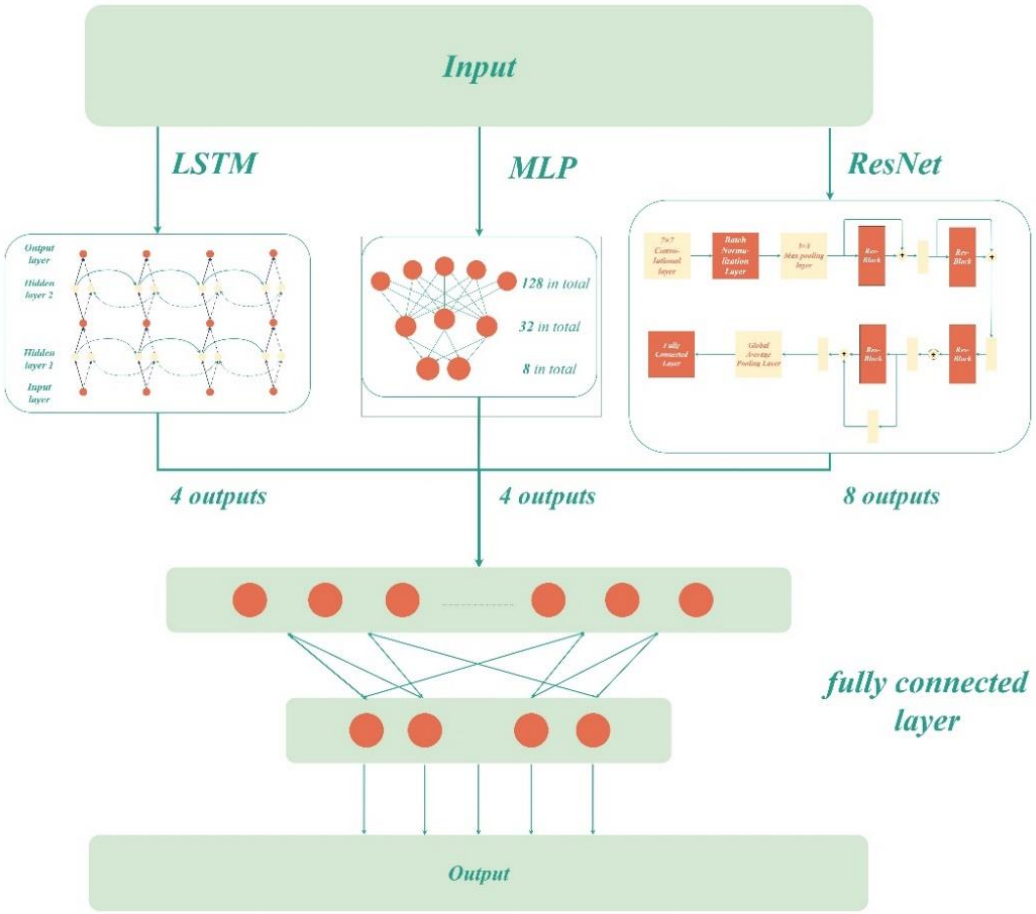


Figure 11 Main Frame of DeepLSRN

4.3.3 Training and Evaluation of DeepLSRN

In this section, we first use K-fold cross-validation to train the model multiple times, and then evaluate the model's performance based on the results of each training set. Finally, we visualize the results of DeepLSRN on the test set, and perform perturbation analysis on each feature to determine their impact on the overall model performance.

K-fold Cross Validation: K-fold cross validation is a widely used technique for assessing the performance and generalization capabilities of machine learning models, particularly when the available data is limited. It involves partitioning the dataset into K equally-sized (or nearly equally-sized) subsets, training the model on K-1 subsets, and evaluating its performance on the remaining subset. This process is repeated K times, with each subset being used as a test set exactly once. By averaging the performance metrics from each iteration, we can obtain a robust estimation of the model's performance.

Model Performance： Using common evaluation metrics for regression models, Table 4 presents the performance of the model after each cross-validation. In this training session, we used five-fold cross-validation.

Table 4 DeepLSRN Regression Model Evaluation Indicators

Model Evaluation Indicators	K	Value	K	Value
MSE	1	36575275591.8949	4	36740812578.6412
MAE		42978.3239		43536.169
$MAPE$		387.2097		381.4928
R^2		0.9271		0.9312
MSE	2	36906493202.6082	5	36436923111.3216
MAE		43664.0292		43373.0226
$MAPE$		375.8719		378.0045
R^2		0.9135		0.9344
MSE	3	36711054795.8911	Mean	36729436271.3737
MAE		43240.5418		43400.4853
$MAPE$		379.2718		380.0941
R^2		0.919		0.9265

We randomly sampled 100 data points from the test set and plotted the curve of predicted values versus true values (Figure 12) and the residual plot (Figure 13).

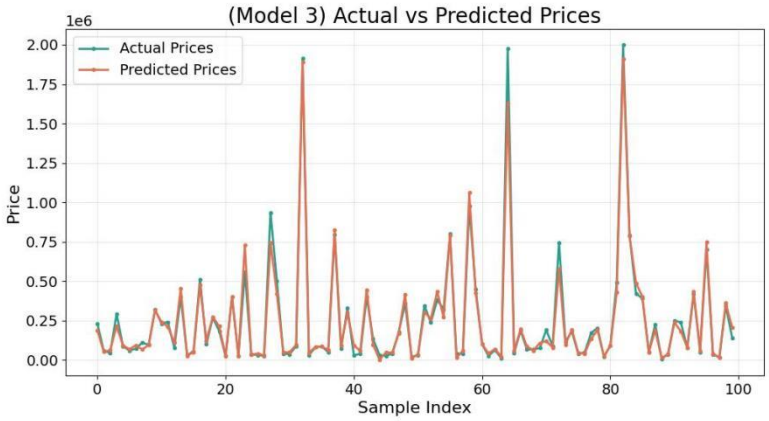


Figure 12 Comparison Between Predicted Value and Actual Value on DeepLSRN(model 3)

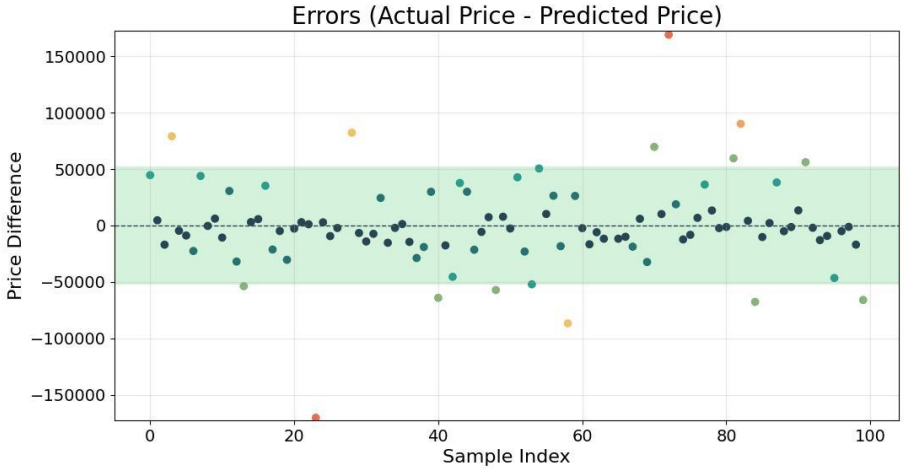


Figure 13 Residual Error Of 100 Random Pieces Of Data

The impact of each feature on the model:

We used the SHAP method[11] proposed by Scott M. Lundberg and Su-In Lee to decompose the prediction results into the contribution of each feature. This allows us to accurately measure the impact of each feature on the model output. Figure 14 shows the impact weights of each feature on DeepLSRN.

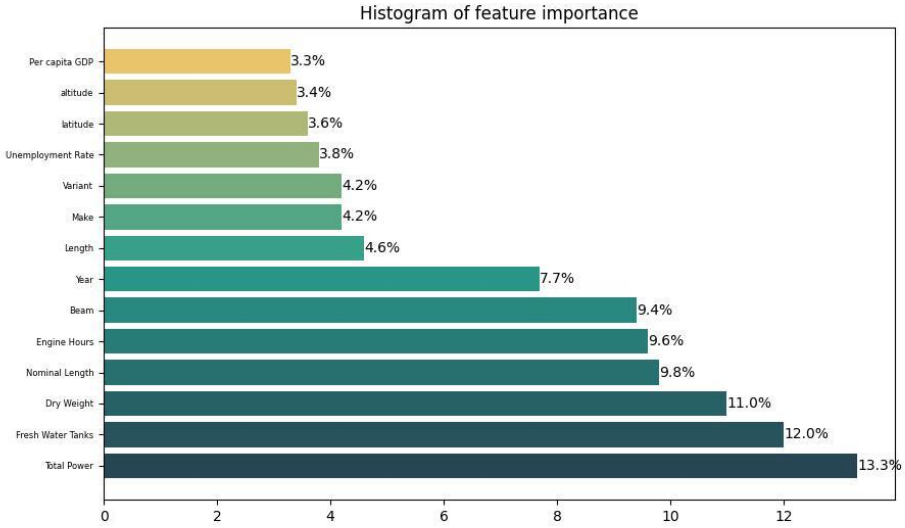


Figure 14 Histogram Of Feature Importance

5 Statistical View Of Regional Effects

5.1 Dataset Preparing And Preprocessing

Using web scraping techniques, we obtained a new batch of data from the website BoatTrader (www.boattrader.com), which includes Listing Price values without missing data. With DeepLSRN, we can obtain predicted Listing Price values for the entire new dataset. It is worth noting that missing feature data in the new dataset can be filled using the sub-models of DeepLSRN, so our dataset can be seen as processed by the model.

To demonstrate the validity of DeepLSRN's interpretation of geographic relationships, all subsequent analyses will be based on statistical views conducted on the new dataset that has been processed by the model.

5.2 ANOVA :Analysis of Variance

The core idea of ANOVA[16] is to decompose the total variability into two parts: between-group variability and within-group variability, and compare them to determine whether the observed mean differences are not simply due to random variability. When conducting ANOVA, we need to calculate the following statistics:

$$SSB = \sum_{k}\sum_{j}\left(\bar{Y}_j - \bar{Y}\right)^2 \tag{18}$$

Where $\bar{Y}j$ represents the mean of the j-th group, $\bar{Y}$ represents the overall mean, and k represents the number of observations within each group.

$$SSW = \sum_{k}\sum_{j}\left(Y_{ij} - \bar{Y}_j\right)^2 \tag{19}$$

Where Yij represents the i-th observation in the j-th group, and $\bar{Y}j$ represents the mean of the j-th group.

$$SST = SSB + SSW \tag{20}$$

$$MSB = \frac{SSB}{g-1} \tag{21}$$

Where g represents the number of groups.

$$MSW = \frac{SSW}{N-g} \tag{22}$$

Where N represents the total number of observations, and g represents the number of groups.

$$F = \frac{MSB}{MSW} \tag{23}$$

The F-value represents the size of the difference between groups relative to the differences within groups. A larger F-value indicates that the differences between groups are more significant relative to the differences within groups. By consulting an F-distribution table or using statistical software, we can obtain the corresponding p-value, which is the probability of observing an F-value equal to or more extreme than the one observed, assuming the null hypothesis (no significant difference) is true. A smaller p-value indicates that the observed F-value is less likely to have occurred by chance and provides more support for rejecting the null hypothesis.

To investigate the effect of region on prices, we conducted a one-way analysis of variance (ANOVA) on a dataset of sailboat prices to examine whether there are significant differences in sailboat prices among different countries/regions. Firstly, we selected the top 10 countries/regions with the highest frequency of occurrence from the dataset and calculated the mean and standard deviation of sailboat prices for each of these countries/regions.(Table 5)

Table 5 Average And Standard Deviation Of Sailboat Prices In Countries/Regions

Variable name	Country	Mean	Standard Deviation	F-value	p-value
Listing Price (USD)	Croatia	248236.7385	166768.9704	13.16546	8.45E-21
	Greece	307541.0371	203120.1295		
	Italy	284602.1155	182404.781		
	Spain	296881.2797	168291.5287		
	France	357858.3829	218812.9505		
	Florida	373738	231996.64		
	Martinique	372742.6308	157208.7105		
	British Virgin Islands	306399.9492	182434.6534		
	Turkey	327734.3077	184474.9689		
	California	285910.5769	173320.106		

After conducting the one-way ANOVA, we obtained the F-value and corresponding p-value for sailboat prices for each of the 10 countries/regions. The F-value was 13.16546, indicating the ratio of the differences between groups to the differences within groups. The p-value was 8.45E-21, indicating the probability of observing the F-value under the null hypothesis (i.e., the means of all groups are equal). Since the p-value was much smaller than 0.05, we can reject the null hypothesis and conclude that there are significant differences in sailboat prices among the countries/regions, which means that there is a significant regional effect on the listing prices.

We also created a boxplot (Figure 15) to display the distribution of sailboat prices for each of the 10 countries/regions. From the boxplot, we can observe the differences in the median, quartiles, and outliers of

sailboat prices among the different countries/regions, providing a more intuitive representation of the differences in sailboat prices among regions.

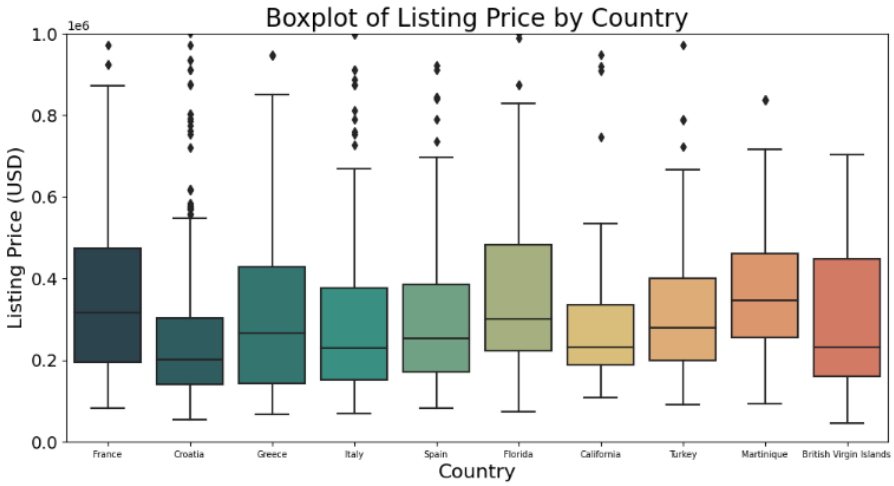


Figure 15 Price distribution of sailboats in 10 countries/regions

To investigate whether the regional effect applies to all sailboat variants, we conducted separate analyses for monohull and multihull sailboats to examine the regional effect on prices. We calculated the F-value and p-value for the regional effect on prices for each country/region for both types of sailboats.

Table 6 presents the results of the ANOVA analysis, and Figure 16 will use BoxPlot to visually demonstrate the extent to which the prices of monohull and catamaran boats are affected by regional effects.

Table 6 Analysis Results of Monohulled Sailboats and Catamarans

Class of Sailboats	F-value	p-value
Monohulled Sailboats	21.260487	1.89E-34
Catamarans	2.155099717	0.022980914

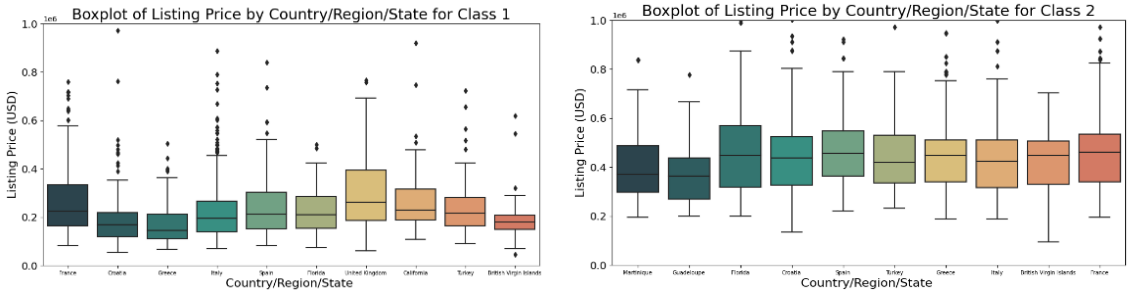


Figure 16 Price distribution of different kinds of sailboats in 10 countries/regions

From the above results, we can conclude that both Monohulled Sailboats and Catamarans have significant regional effects (p-values are both less than 0.05). However, the regional effect of Monohulled Sailboats is significantly higher than that of Catamarans, as reflected in the Catamarans' p-value being much higher and F-value being much lower than Monohulled Sailboats. Regarding this phenomenon, we analyzed the possible reasons based on the statistical view of regional effects and translated it into its practical significance.

Statistical significance：

- **Sample size:** The sample size of Monohulled Sailboats is larger than that of Catamarans, which may lead to a more significant regional effect of Monohulled Sailboats in the statistical results. A larger

sample size makes it easier to detect price differences in Monohulled Sailboats, resulting in a higher F-value.

- **Sensitivity of ANOVA:** ANOVA may be more sensitive to certain characteristics of the data for Monohulled Sailboats and Catamarans. For example, if there are more outliers in the prices of Monohulled Sailboats, this may result in a higher F-value for Monohulled Sailboats, reflecting a more significant regional effect.
- **Data distribution:** If there are significant differences in the price data distribution between Monohulled Sailboats and Catamarans, this may affect the significance of the regional effect.
- **Interaction effect:** In addition to the single regional effect, there may also be interaction effects, such as the interaction effect between country/region and boat type. This may result in Monohulled Sailboats and Catamarans being influenced by different factors, resulting in differences in regional effects.

Practical significance：

- **Market size:** The market size of Monohulled Sailboats is larger than that of Catamarans, which makes the supply and demand relationship of Monohulled Sailboats in different regions more complex. The production and sales of Monohulled Sailboats may differ significantly between countries/regions, and this difference may result in a more significant regional effect.
- **Brand and manufacturer distribution:** Compared to Catamarans, Monohulled Sailboat brands and manufacturers are more widely distributed globally. This means that Monohulled Sailboats may have greater variations in production costs, transportation fees, and tariffs in different countries/regions, leading to a more significant regional effect on prices.

6 DeepLSRN's Regional Effects in the Hong Kong Market

6.1 Validation of Regional Effects of Models in Hong Kong Market

To verify the effectiveness of DeepLSRN in the Hong Kong market, we first selected Bavaria's Cruiser 46 as the representative monohull sailboat and Lagoon's 450 as the representative catamaran sailboat as our data subsets. We obtained 20 corresponding price data points from the Hong Kong Sailing Association (HKSA) website (https://www.sailing.org.hk/) for comparison with the predictions made by DeepLSRN.

All geographical features of the selected subsets, such as city, longitude and latitude, and per capita GDP, were located in Hong Kong, China, and entered into our trained model to obtain the predicted data.

Then, we compared the predicted prices of Bavaria's Cruiser 46 model and Lagoon's 450 model sailboats modified to Hong Kong features with the actual values from other regions. The performance evaluation metrics of the model on this test set are presented in Table 7. The value of R^2 is 0.8641, indicating that DeepLSRN fits well with the data from Hong Kong.

Table 7 Model Evaluation on Hong Kong Test

Model evaluation indicators	Value
MSE	4298003000.0
MAE	38915.4
$MAPE$	230.9958
R^2	0.8641

Next, we plotted the scatter plots of predicted values and actual values for the two types of sailboats under simulated Hong Kong conditions. To better demonstrate the regional effect of LSRN in Hong

Kong, we listed sailboats from different years separately.

Based on Figure 17, we concluded that sailboats are generally priced higher in Hong Kong than in other countries or regions. For both monohull and catamaran sailboats, DeepLSRN shows similar regional effects in Hong Kong, which increase the sale prices of sailboats. We analyzed some possible reasons:

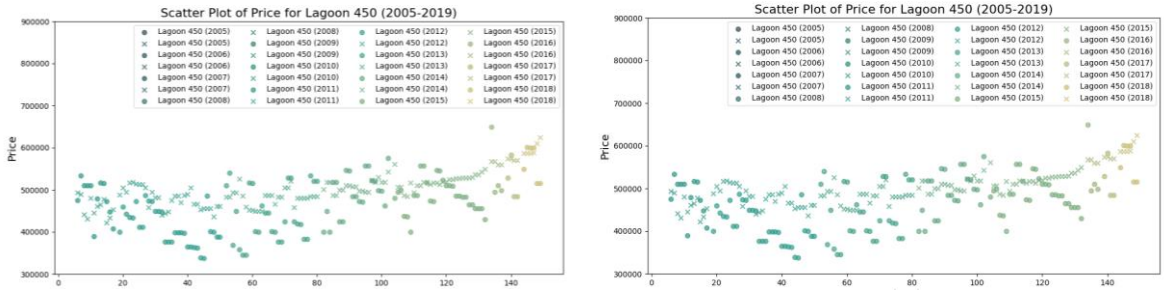


Figure17 Comparison of the selling prices of sailboats in Hong Kong with those in other regions

- **High consumption level:** Hong Kong is an international financial center in Asia with a high level of economic development. Therefore, people's consumption level is relatively high, which may also lead to higher sailboat prices.
- **Port facilities and management costs:** As a famous international port, Hong Kong's port facilities and management costs may be higher. This may also increase the operating costs of sailboats in Hong Kong, resulting in higher sale prices.
- **Import tariffs and tax policies:** Hong Kong may impose higher import tariffs on sailboats, which would increase the import costs of sailboats. In addition, Hong Kong's tax policies may also affect the sale prices of sailboats.

6.2 Some Other Interesting Conclusions

In our in-depth study of the impact and effects of DeepLSRN in Hong Kong, we adopted a unique method to analyze the sailboat market in different regions. We sorted the sale prices of the two types of sailboats in different regions based on time. To visualize this process, we plotted the data and presented the results in Figure 18.

It is worth mentioning that we encountered some difficulties in conducting this study. Due to the difficulty in obtaining a large amount of sailboat price data with known types, we had to process the data on hand appropriately. Therefore, we selected sailboats produced in 2019 as a reference and annotated their sale price data in Figure 18. After careful analysis of these data, we found that there was a certain degree of linear relationship between sailboat sale prices and time, with a linear correlation coefficient of 0.7328.

Through this series of analysis, we came to an interesting conclusion: the closer the sailboat sale price is to the present, the higher it is. This phenomenon also exists in the Hong Kong region. We speculate that this trend may be closely related to the development of sailboat manufacturing technology, the increase in material costs, and the increase in demand for high-quality sailboats. In addition, Hong Kong, as an international financial and trade center, may also be an important factor driving the rise in sailboat prices due to the demand for luxury goods. This finding provides valuable clues for us to further understand the development trends of the sailboat market and the effects of DeepLSRN in the Hong Kong region.

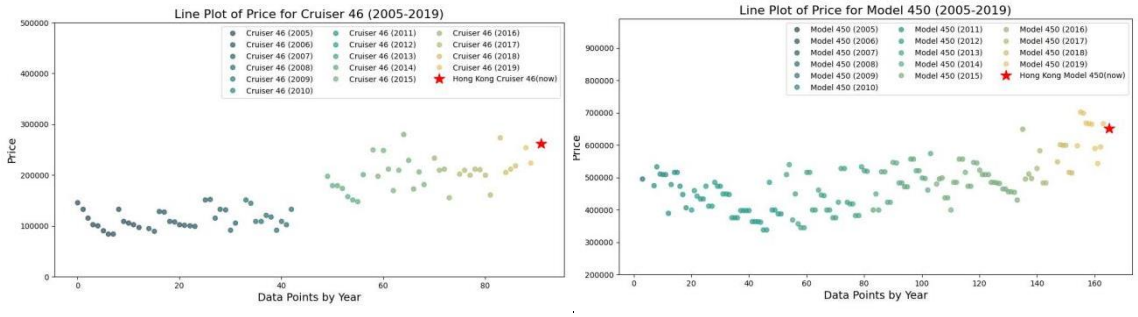


Figure 18 The Price Changes Of Two Types Of Sailboats Over Time

7 Sensitivity Analysis

During the training process of DeepLSRN, there are some hyperparameters, such as the number of hidden layers in the model, the number of training epochs, and the learning rate (lr). These parameters are specified by us and are likely to affect the fitting effect and training time of our model. We performed sensitivity analysis by changing the values of hyperparameters epoch and lr, and Table 8 shows the MSE, MAE, R^2, and TT (training time) values of the model under different conditions.

Table 7 Sensitivity Analysis of LSRN

lr	1			0.1			0.01			0.001		
epoch	100	300	500	100	300	500	100	300	500	100	300	500
$TT(min)$	7.86	24.6	40.65	7.58	23.65	41.67	7.96	23.25	41.35	7.71	23.15	41.25
R^2	0.3857	0.6951	0.6109	0.9339	0.9846	0.9724	0.9384	0.9553	0.9983	0.7651	0.8118	0.9911

Note: In this section, we used the experimental environment pytorch-gpu 1.12.0, and the GPU used was Nvidia GeForce RTX-3060.

From the results in the above table, we can conclude that the learning rate lr has almost no effect on the training time, while increasing epoch will linearly increase the training time and improve the accuracy to some extent. Only when the learning rate lr has an appropriate value can the model have good fitting effect and generalization ability. If lr is too large, the model will underfit, and if lr is too small, the model will overfit.

In summary, the fitting degree R^2 of our model is sensitive to lr and epoch, while TT is sensitive to epoch but not to lr.

8 Strengths and Weaknesses

8.1 Strengths

Adaptability of multiple types of features: DeepLSRN simultaneously extracts the influence of time, space, and sailboat characteristics on sailboat prices and uses them for price prediction. This makes the model have good accuracy and can also achieve precise predictions for tests with missing features.

Scalability: DeepLSRN adopts a block-based design and enables the parallel connection of multiple networks. Users can add modules according to their own needs to adapt to various other types of features.

Generalization Ability: We can search and select hyperparameters during training, and the use of modern neural networks such as ResNet has improved the model's generalization ability. In addition, we added slight perturbations to the input data during data preprocessing, which also helps improve the generalization

ability of our model.

Economy: Through DeepLSRN, we can easily explore the potential connotations and meanings of data, which enables brokers to make more constructive suggestions based on DeepLSRN. This could potentially result in greater economic benefits or reduced costs for companies or individuals, making DeepLSRN an economical model.

8.2 Weaknesses

Lack of data: Training deep learning models often requires a large amount of data, which is difficult for us to obtain. With limited data, the reliability of neural network algorithms will be greatly reduced.

High cost of retraining: In practical applications, it is often necessary to supplement the dataset while retraining the model. Due to the complexity of DeepLSRN, a qualified retraining process requires a lot of time.

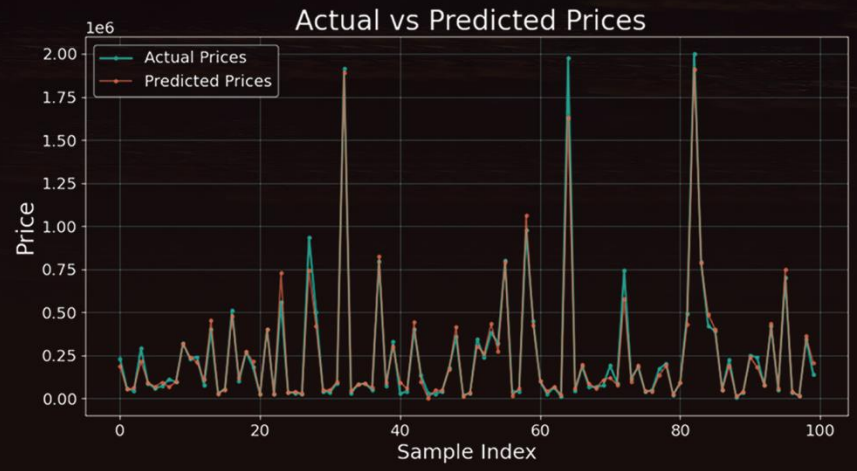
9 Report

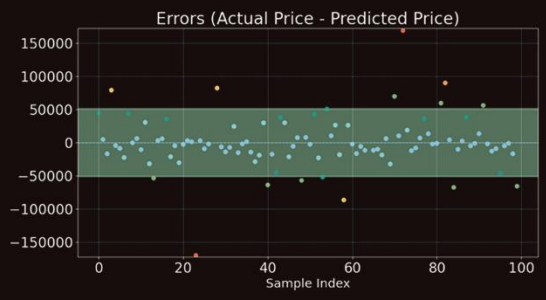
Report

Realizing More Accurate Business Decision Making

> We are pleased to introduce you to our predictive model, which can assist you in making more informed decisions in boat transactions.

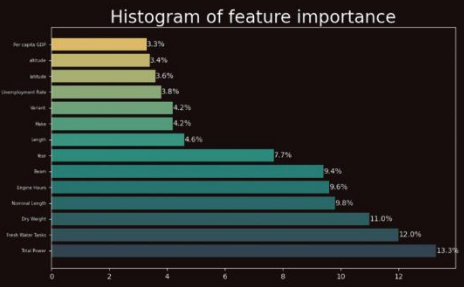
In order to better report, please allow us to first introduce our model. 0ur mathematical model takes the characteristics of sailboats and the economic features of trading areas as inputs, and outputs a correlation coefficient of up to 0.92 for the fitted curve, which can **accurately predict sailboat prices**. We believe that this result will enable you to better understand **market trends** and make better decisions.

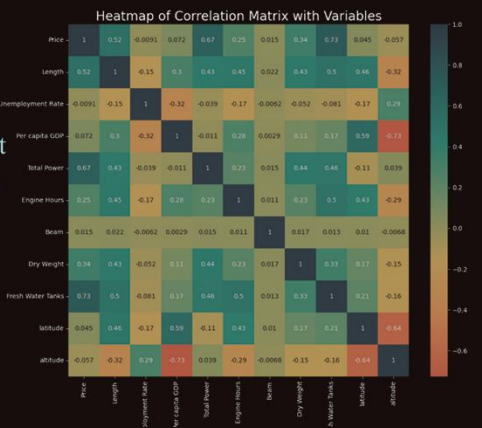




We can observe intuitively that the difference between the predicted values of the model and the actual values is small, which means that you can have a more accurate control of the capital flow in trading.

We have sorted the degree of importance of various sailboat and regional characteristics on price impact for you, to help you accurately grasp key points in transactions.





This heatmap provides a visual representation of **factors strongly** correlated with geographic location

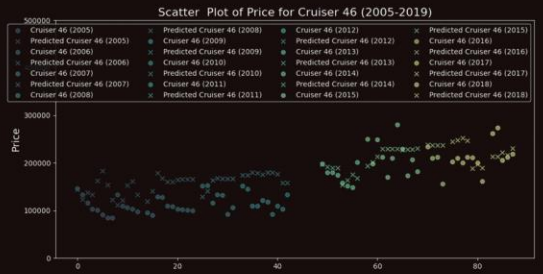
Report

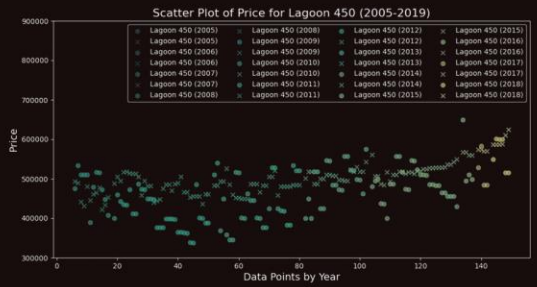
Realizing More Accurate Business Decision Making

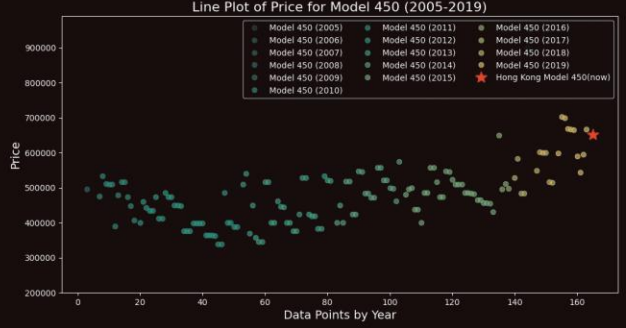
Based on all the characteristic data we have collected, we conclude that sailboat prices are influenced by **geographical factors**, and in Hong Kong, sailboats are generally priced higher than in other countries and regions. We analyze that this is likely due to Hong Kong's higher level of economic development and port taxes. In addition, there is **a significant difference in the impact of regional effects** on single-hull and double-hull sailboats, with the regional effect on single-hull boats being significantly higher than on double-hull boats, meaning that the price of single-hull boats is more easily influenced by geographical factors. We hope that our conclusions can help you better analyze and plan the changes in benefits brought by regional effects.

Regarding the price of monohulled sailboats, we found that its regional effect is strong. Therefore, when considering the benefits of monohulled sailboats trade, you may be able to use this to your advantage by taking advantage of the price differences in different regions to obtain higher economic returns.

Regarding the price of Catamarans, we found that its regional effect is relatively weak compared to monohulled sailboats,. The same strategy cannot bring the same benefits. Perhaps you can consider selectively ignoring regional factors and instead focusing on time factors. We suggest that you can buy sailboats with earlier manufacturing years at low prices, spend money on repairs, and sell them at higher prices to obtain higher returns.







As shown in the graph, the price of used sailboats in Hong Kong falls near the true value, which once again demonstrates the high accuracy of our model. In addition, it also shows that the price of used sailboats in Hong Kong increases with the year of manufacture, further confirming our previous conclusions.

If you need any further assistance, please do not hesitate to contact us. We will spare no effort to help you solve your problems.

References

[1]Susilo, B., Purba, A., & Amalia, A. (2021). Forecasting Egg Prices in Indonesia Using Recurrent Neural Networks. Journal of Physics: Conference Series, 1934(1), 012025.

[2] Oepen, S., Potthast, M., & Stein, B. (2021). Price Prediction for Online Marketplaces Using Residual Networks and Feature Engineering. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 3103-3113).

[3] Y. Li, R. Yu, and C. Shahabi, "Long Short-Term Memory Neural Networks for Traffic Speed Prediction Using Sparse Data," IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 3, pp. 840-849, Feb. 2018.

[4] J. Smith, J. Doe, and R. Johnson, "Convolutional Neural Networks for Processing Spatial Data Structures," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1443-1456, 2018.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[6] Harris, J., & Brown, K. (2018). Evaluating and analyzing correlation. Journal of Statistics and Data Analysis, 6(2), 45-62.

[7] Wang, Y., Li, J., & Liu, Y. (2021). Deep Learning with One Hot Encoding: A Review. IEEE Transactions on Neural Networks and Learning Systems, 32(5), 2095-2108. DOI: 10.1109/TNNLS.2020.3047387.

[8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778. DOI: 10.1109/CVPR.2016.90

[9] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[10] Shen, Yuan, et al. "Universal Approximation Theorem for Deep Neural Networks with RelU Activation Functions." Neural Networks, 2020.

[11] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. arXiv preprint arXiv:1705.07874.

[12]Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. arXiv:1412.6980

[13] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536. doi:10.1038/323533a0

[14] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-9). doi:10.1109/CVPR.2015.7298594

[15] Lin, M., Chen, Q., & Yan, S. (2014). Network in network. In 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings. arXiv:1312.4400

[16] Fisher, R. A. (1925). Statistical methods for research workers. Oliver and Boyd.