Forecasts for the Ecology and Fisheries Economy of Scottish herring and mackerel

Yibo Li *, Yuyang Lin, Pengtao Zhang, Jingrui Pan

Macau University of Science and Technology, Macau, China

* Corresponding Author Email: liyiboqwer@outlook.com

Abstract. In this project, we aimed to develop an efficient supervised learning system foranalyzing second-hand sailboat prices in Hong Kong. The project was divided into three main steps: data cleaning and denoising, dimensionality reduction, and efficient supervised learning. In the first step, we successfully cleaned and denoised the data by filling the missing values using mean, mode. We also detected and removed 154 outliers and abnormal points through Q-Q diagram. The remaining data passed the normality test after dimensionless standardization, and we confirmed that they all conformed to a normal distribution. In the second step, we reduced the dimensionality of the data by combining Pearson correlation coefficient and the mRMR algorithm. We selected the top 6 features as the inputs for the supervised learning model. In the third step, we established a supervised learning system with the second-hand sailboat price as the top layer, quality of the sailboat, year of production, region, and volume as the middle layer, and the remaining small index features as the bottom layer. We used an SVM model with penalty conditions for the S-N layer and a random forest model with parameter adjustment for the S-P layer. The S-N model achieved an accuracy of over 85%, AUC of 0.93, while the R2 of the S-P model was greater than 0.87 and the RMSE was less than 0.7, indicating the model was well optimized. Overall, our project successfully established an efficient supervised learning system to analyze second-hand sailboat prices in Hong Kong, providing insights into the regional effect and enabling better decision-making in the sailboat market.

Keywords: Random Forest; SVM Vector Machine; mRMR Algorithm; Pearson Correlation Coefficient.

1. Introduction

1.1. Problem Background

The background of the global second-hand sailboat market is shaped by a variety of factors, including the popularity of recreational sailing, advancements in boat building technology, and the growth of the boating industry. Additionally, the emergence of online marketplaces for buying and selling boats has made it easier for buyers and sellers to connect across geographic regions, contributing to the global nature of the market. With the rise of the sharing economy, boat rental services have also become more popular, further expanding the market for sailboats. However, economic factors such as recessions or changes in consumer spending patterns can also impact the demand for sailboats and therefore affect the overall health of the market.

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

Can you develop a mathematical model that predicts the listing price of each sailboat in the provided spreadsheet? Please also describe all data sources used and discuss the precision of your estimates for each sailboat variant's price.

Problem 2

How does region affect the listing prices of sailboats? Is this effect consistent across all sailboat variants? Please discuss the practical and statistical significance of any regional effects observed.

Problem 3

How can your modeling of the given geographic regions be useful in the Hong Kong (SAR) market? Please choose a subset of informative sailboats, including both monohulls and catamarans, from the provided spreadsheet. Then, find comparable listing price data for this subset from the Hong Kong (SAR) market. Please model the regional effect, if any, of Hong Kong (SAR) on each sailboat's price for the subset. Is the effect the same for both catamarans and monohulls?

Problem 4

What other interesting and informative inferences or conclusions can be drawn from the data on sailboats provided?

Problem 5

Please prepare a one- to two-page report for the sailboat broker in Hong Kong (SAR). The report should include a few well-chosen graphics to help the broker understand your conclusions.

1.3. Literature Review

Scholars studying sailing prices are currently making progress using new data analysis techniques and research methods. Among them, technologies such as machine learning and artificial intelligence have shown great potential in analyzing large amounts of data and predicting price trends. At the same time, scholars will also focus on studying the supply and demand relationship and pricing mechanism of the market, as well as the price trend of specific brands and models. In addition, they are also studying how to deal with factors such as economic changes, policy changes and climate change on the price of sailboats. In general, scholars who study sailing prices are constantly exploring new methods and techniques to better understand the market, predict price trends, and provide useful information to those in the industry.

2. Assumptions and Justification

Hypothesis: Sailboat listing prices are influenced by a combination of factors such as beam, draft, displacement, rigging, sail area, hull materials, engine hours, sleeping capacity, headroom, electronics, and economic data by year and region.

Rationality: Sailboat prices are likely to be determined by a complex combination of factors, including the features of the boat itself and broader economic trends. By incorporating these various factors into a mathematical model, we can gain a more precise understanding of how sailboat prices are determined.

Hypothesis: Region has a significant effect on sailboat listing prices, but this effect may not be consistent across all sailboat variants.

Rationality: Economic and cultural factors can vary widely across different regions, and these factors are likely to influence the demand and pricing of sailboats. However, the relative importance of these factors may differ for different types of sailboats, leading to differences in the impact of region on pricing.

3. Definitions and Notations

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

Table 1. Notations used in this paper

Symbol	Description
P	The level of satisfaction
S	Hierarchy where the remaining three variables are located
N	Hierarchy of feature variables
f	Implicit functions between the S-N hierarchy
$R_{x,y}$	Correlation coefficient
G_i	Gini coefficient
V_i	Gini score for layer i
w_i	Weight of the ith feature variable

Designed Waterline Length (L_{WL}): Designed waterline length refers to the length of the designed waterline on a ship. The design waterline refers to the intersection line between the hull-shaped surface and the water surface when the ship floats freely and upright on still water in the expected design state. That is, the waterline corresponding to the design displacement. For civil ships, the design waterline is usually the waterline when the ship is fully loaded.

Beam: Briefly referred to as the ship's width. The maximum breadth from the outer edge of one side frame to the top of the other side frame at or below the deepest subdivision load line.

Draft(ft): Draft refers to the deepest length of the submerged part of the ship in the water, and different ships have different drafts. The same ship also has different drafts according to different loads and the salinity of the waters where it is located. Larger ships will not be able to enter shallow bays, ports, or canals because their draft is too deep. The waterline scale is engraved on the outside of the hull to show the draft.

4. Used Sailing Yacht Price Prediction Model

4.1. Description and analysis of raw data

The price of the second hand-boat is separate statistical analysis, with a total chart of 2,768 data points. Of these, 10 properties, a total of 2,788 data, all known data, lost values 6, total outward-oriented numbers are 52 values, the actual meaning of these outcomes is transmission errors, human causes, and writing errors that create data distortions that seriously distort people's evaluation.

4.2. Data noise reduction and Kolmogorov-Smirnov (K-S) normality test

4.2.1 Kolmogorov-Smirnov (K-S) normality test

The Kolmogorov-Smirnov(K-S) test is a test method based on the empirical distribution function, if the number of population distribution plots F(x) is unknown, but there are sample observations, then the n observations in the sample are arranged in order from smallest to largest into, and the empirical distribution function can be obtained:$x_1, x_2, \cdots, x_n$ $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$

$$F_n(x) = \begin{cases} 0, x \leqslant x_{(1)} \\ \frac{i}{n}, x_{(i)} \leqslant x \leqslant x_{(n)}, i = 1, 2, \cdots, n-1 \\ i, x > x_{(n)} \end{cases} \tag{1}$$

According to the Glivenko-CantelII theorem, the empirical distribution function F_n (x) derived from sample observations is a good approximation of the population distribution function F(x) when n is large. In 1933 Kolmogorov proposed to test the statistical indicator D_n, and in 1948 Smimov gave a table of fits for estimating the empirical distribution of Dn. Statistical indicator D_n, the specific form is:

$$D_n = \max_{1 \leqslant i \leqslant n} \left\{ |F(x_i) - \frac{i-1}{n}|, |\frac{i}{n} - F_n(x_i)| \right\} \tag{2}$$

The Kolmogorov-Smirnov test is a test method that is easier to successfully accept the normality hypothesis under large samples, and the Kolmogorov-Smirnov test results are subject to the sample sum of , and the Kolmogorov-Smirnov test results are subject to the statistical software SAS when the sample is .$N \geqslant 5000$ $N \geqslant 2000$

In view of the fact that the statistical software SPSS is widely used by the majority of medical workers, it is particularly pointed out here that there are two ways to implement the Kolmogorov-Smirnov test in SPSS, one is the ExplorePlot option in the analysis module, and the second is the single-sample Kolmogorov-Smirnov test in the nonparametric test module; The test results of Method 1 often have an an a-note number in the upper right corner, indicating the corrected Kolmogorov-Smirnov test results, which are suitable for general normality tests. The nonparametric test results of Method 2 are not corrected, and this test method can only infer whether the data follow the standard normal distribution. In practice, if the conclusions of the two test methods are inconsistent, the results of Method 2 generally prevail.

When using the statistical software R for the Kolmogorov-Smirnov normality test, it is specifically required that the same value cannot appear in the test sample, because the R software requires the test sample to be continuous when performing the normality test, and the probability of the same value in the continuous distribution is 0, and if the same value occurs, a sufficiently small value can be added to it.

In addition, the Kolmogorov-Smirnov test can test the normality of the population distribution represented by the sample, and can also test whether the population represented by the sample obeys other distributions, which is 'universal'. Compared with other methods, the normality test efficiency of this method is low, and it is most easily affected by factors such as sample content and outliers.

4.2.2 Data noise reduction

Here, the method of $Q-Q$ test is used to visualize and express the results of the normality test for each feature value.

Assuming that the evaluation feature value $X_1, X_2, ..., X_n$ comes from people's overall call satisfaction $N(\mu, \sigma^2)$, the sample observations are arranged in order from largest to smallest to obtain the order statistic $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$ and the empirical distribution function

$$F_n(x) = \begin{cases} 0 & x < x_{(1)} \\ \dfrac{k}{n} & x_{(k)} \leqslant x \leqslant x_{(k+1)}, k = 1, 2 ... n \\ 1 & x \geqslant x_{(n)} \end{cases} \tag{3}$$

F Since when the sample size is sufficiently large,$\forall \varepsilon > 0$has $P(|F_n(x) - F(x)| > \varepsilon) = 0$.

$F(x) = P(X \leqslant x) = \varphi\left(\frac{x-\mu}{\sigma}\right)$, when the sample is sufficiently large, $F_n(x) = \varphi\left(\frac{x-\mu}{\sigma}\right), a.s$, that is, $\frac{x-\mu}{\sigma} = \varphi^{(-1)}(F_n(x)), a.s.$

So $\forall x \in R$. Order $\varphi^{-1}(F_n(x)) \equiv \omega$, has $x \equiv \sigma\omega + \mu, a.s$.

The above equation means that when the population is $N(\mu, \sigma^2)$ and the sample size n is sufficiently large, the point (ω , x) consisting of the sample observation x and the inverse standard normal distribution ω is almost in a straight line.

The resulting correlation coefficient is

$$r=\frac{\sum\left(\frac{x_{(i)}-\overline{x}}{x}\right)\left(\frac{u_{(i)}-u}{u}\right)}{\sqrt{\sum\left(\frac{x_{(i)}-\overline{x}}{x}\right)^2}\sqrt{\sum\left(\frac{u_i-u}{u}\right)^2}} \tag{4}$$

And test its normality (if point $(u_i,x_{(i)})$ is approximately on a straight line, the correlation coefficient r=...), so the sample can be considered to be from a normal population.)

4.3. Data cleansing

The missing values do not conform to the laws of mathematical models and big data statistics, and are more affected by external factors. Some missing values will lose valid information during data mining modeling, resulting in more significant uncertainty in the model, and the laws contained in the model are more difficult to control, which may also cause confusion in the modeling process and form unreliable output. The supplement of data missing points will make the overall change range of the model more accurate, reduce the error, so that it is more perfect, more optimized, and the results are more correct, and these supplementary data missing points will not have much impact on the overall change trend.

Using the Hermite interpolation method to avoid unnecessary Longer phenomenon, the team chose the segmented cubic Hermitian interpolation method, and took two known points at each end of any missing point and calculated by CR/CI test, and it can be known that they are all satisfied:

$$\begin{aligned}&\text{i.}G(x)\textit{The polynomial degree on each cellinterval is}3\\&\text{ii.}G(x)\in C_1(\alpha,\mathrm{b})\\&\text{iii.}G(x_1)=f(x_1),G'_{(x_1)}=f'_{(x_1)},\left(\dot{I}=0,1,\cdots,\ n\right)\end{aligned} \tag{5}$$

After the conditions are satisfied, we express the formula, which is a cubic Hermitian interpolation polynomial:

$$\begin{aligned}&G(x)=h_k y_k(x)+h_{(k+1)}y_{(k+1)}(x)+H_k(x)y'_k+H_{(k+1)}(x)\ y'_{(k+1)}=\\&\left(1+2\frac{x-x_k}{x_{k+1}-x_k}\right)\left(\frac{x-x_{k+1}}{x_k-x_{k+1}}\right)^2y_k+\left(1+2\frac{x-x_{k+1}}{x_k-x_{k+1}}\right)\left(\frac{x-x_k}{x_{k+1}-x_k}\right)^2y'_{k+1}\\&+(\chi-x_k)\left(\frac{x-x_{k+1}}{x_k-x_{k+1}}\right)^2y'_k+(x-x_{k+1})\left(\frac{x-x_k}{x_{k+1}-x_k}\right)^2y'_{k+1}\end{aligned} \tag{6}$$

So we can completely supplement most of the missing values, and finally get the global annual average temperature change trend as shown in the Figure 1:

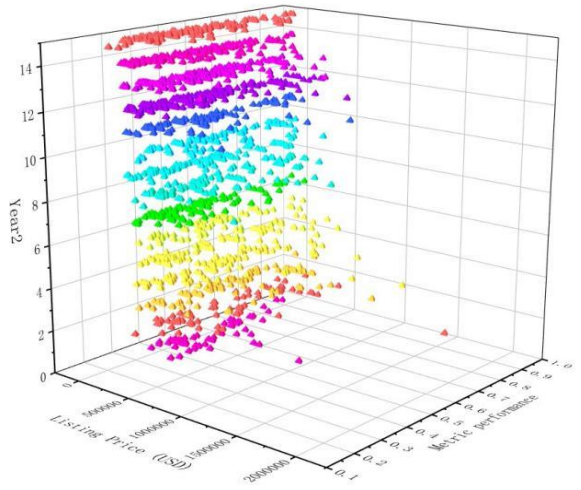

Fig 1. The global annual average temperature change

5. Selection and analysis of indicators

5.1. Construction of the quality system of sailboats

5.1.1 Selection of sailing indicators

Among the many performance indicators of sailboats, we have selected four indicators that best represent the value and performance of the sailboat, which are used to indicate the value for money of used sailboats:

Displacement/sailing area

The weight of the ship includes the hull and keel, without affecting the strength of the structure, the lighter the hull of the same size, the better the performance of the vessel, we know that the FRP hull is the cumulative result of FRP lamination (simply understood as fiberglass cloth and resin paste), in this process, if more composite materials or processes such as vacuum technology are used, both strength and weight reduction can be achieved.

The smaller the ratio of sail displacement to area, the better the boat's performance.

Length * width

The width of the hull, the shape and weight of the keel determine the stability of the vessel when it rolls.

The keel of the T-shaped structure generally adopts the form of a stabilizer plate and a lead heavy hammer, which can not only optimize the profile of the stabilizer plate through design, so that the resistance generated by the water flow is smaller, but also increase the overturning moment of the hull because the overall center of gravity is shifted down.

LWL

Looking at the size of a boat is to look at the comprehensive length, and what really affects the performance of the boat is actually the length of the LWL waterline. There are advantages to a long waterline, first of all, the top speed of the boat may be higher, and secondly, the wave resistance of the boat will be better, these are the core elements of a hull performance.

Draft

It determines whether the water depth of the water suitable for this boat is sufficient, whether the bridge is passable, etc., and the above also considers the influence of tides.

5.1.2 Topsis analysis

Scores were calculated using TOPSIS

$$Q=\begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n1} & q_{n2} & \cdots & q_{nn} \end{bmatrix} \tag{7}$$

After forward and standardized processing, the scoring matrix Q obtained is all extremely large data, from which the ideal optimal solution and the worst solution can be extracted.

The optimal solution is:

$$q_{\max}=[\max(q_{12}\cdots q_{m1}),\max(q_{1n}...q_{mn})] \tag{8}$$

The worst solution is:

$$q_{\min}=[\min(q_{12}\cdots q_{m1}),\min(q_{1n}...q_{mn})] \tag{9}$$

The distance between the i-th component and the optimal solution can then be calculated:

$$d_i^{+}=\sqrt{\sum_{j=1}^{m}[w_j(q_{ij}^{+}-q_{ij}\ )]^2} \tag{10}$$

The distance between the ith part and the worst solution can then be calculated:

$$d_i^- = \sqrt{\sum_{j=1}^{m} [w_j (q_{ij}^- - q_{ij})]^2} \tag{11}$$

In turn, the score of each part of the i can be calculated: $S_i = \frac{d_-}{d^+{}_i + d_i^-}$

$0 \leqslant S_i \leqslant 1$, When d_i^+ is smaller, the smaller the distance from the optimal solution, and the higher the score; d_i^- with hours, saying that the smaller the distance between the i name and the worst solution, the lower the score. In this way, we can choose a second-hand sailboat according to the rating level.

5.2. K-means clustering of countries

In order to find the center point of the sample, we decided to process the data using the k-means method to achieve the effect of minimizing the average error of the tie of the results.

K-means belongs to unsupervised clustering algorithm. Its basic idea is to divide the sample set into K clusters (C_1, C_2 ... C_k) based on the distance between samples for a given dataset, so that the points within the cluster are as closely connected as possible, while the distance between clusters is as large as possible. I

Our learning goal is to minimize the average error, which is

where is the mean vector of cluster, also known as the centroid, with the expression is μ_i C_i

$$\mu_i = \frac{1}{|C_i|} \sum_{x \epsilon C_i} x \tag{12}$$

K-means algorithm flow:

Input: Sample set, the cluster type is numeric; $D = \{x_1, x_2, \cdots, x_m\}$ Output: Cluster division; $C = \{C_1, C_2, \cdots, C_K\}$

(1) Randomly select several points from the data as the initialized center point.

(2) calculate the distance from each remaining sample to the center point: which will belong to the center cluster with the smallest distance; $\mu_i (j = 1, 2, 3, ...) d_{ij} = ||x_i - \mu_j||_2{}^2$

(3) Recalculate cluster center $\mu_j = \frac{1}{c_i} \sum_{x \epsilon c} x$

(4) If the cluster center changes, return to step 2, and if the cluster center remains unchanged, the algorithm ends.

After the above series of processing, we get the result shown in the Figure 2:

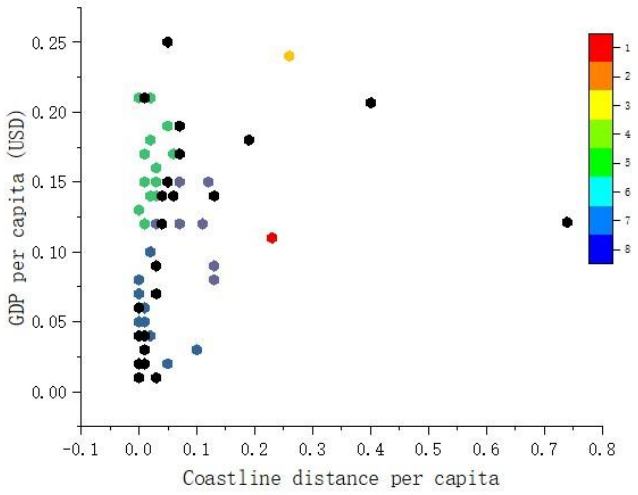

Fig 2. The relationship between coastline distance per capita and gdp per capita

6. Establish a supervised learning prediction model for used sailing prices

6.1. Establish a supervised learning system at the S-P level

By juxtaposing the categories of second-hand sailboats, we perform ordinary algebraic classification learning of several models on the data, and obtain a model with a higher supervised learning effect of the model, and perform grid search parameter tuning and Bayesian optimization respectively, and the learning results are as follows:

In order to evaluate the prediction error and accuracy of the model, the root mean square error (RMSE), mean squared error (MSE) and (R^2) goodness-of-fit are used to evaluate the prediction results of the model, and the three calculations are defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)}$$

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

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(\overline{y}_i - \hat{y}_i)^2}$$

y_i——*The data test set is the true value*;

$\hat{y}_i$——*Model predictions*;

$\overline{y}_i$——*The mean of the data test set sample*。

After preliminary learning and selection, we find that SVM vector machine and integrated decision tree have better accuracy, lower RMSE and higher R^2 value, so that the error mean square deviation is controlled within 0.9, and the accuracy is higher than 85%, which has high applicability.

The team adjusted the following parameters of the SVM vector machine by grid parameter adjustment method to achieve the goal of optimizing the RMSE score when the highest is achieved.

The iteration parameters and generation process are shown in the Figure 3:

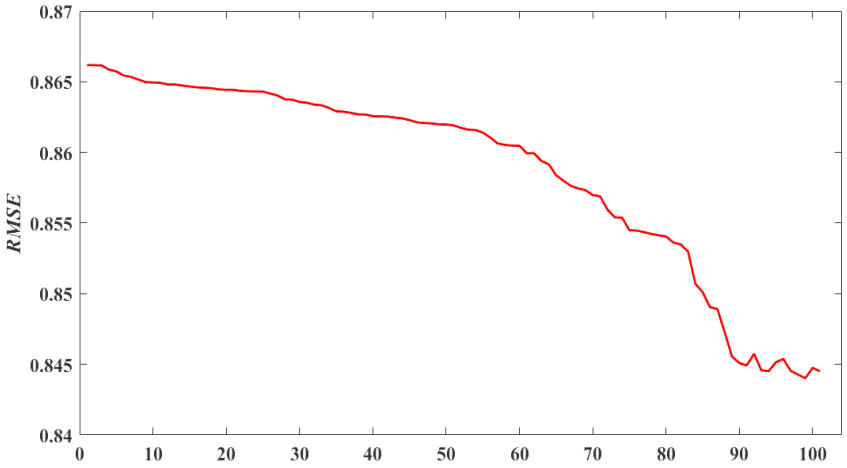

Fig 3. The iteration parameters and generation process

The obtained RMSE and heat map (Figure 4), R^2 as shown in the figure below, and the importance of our team, we prefer to pursue the smallest value for *RMSE*, but R^2 there are still some reference values, so we think the smaller the RMSE, the better, and R^2 the larger the better, so we forward the two values and perform weighted analysis

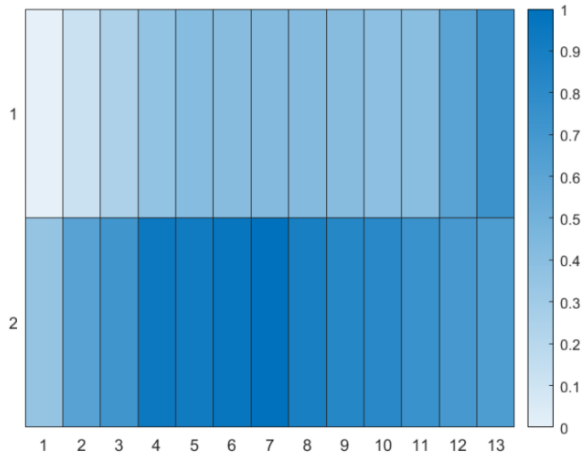

Fig 4. Heat map

Table 2. The team found that when the parameters are

Kernel functions	Box constraint level	Nuclear-scale model	Specifies the kernel scale
gauss	1	Manual	12

According to the parameters in the Table2, we finally use the self-parametric Gaussian kernel function-SVM vector machine for supervised learning, and its mid-kernel scale is KernelScale=12, close to the medium Gaussian SVM level, high accuracy, and the sample size is as high as 4000+, so there will be no overfitting phenomenon, and the box constraint level remains unchanged

Below we will use a self-parametric Gaussian medium SVM vector machine for supervised learning, and the main principles of Gaussian SVM are explained below

Assuming a hyperplane, the support vectors $w^T x+b=0$ of dichotomous *SVMs* can be represented. The $\boldsymbol{m}$ input-output space of a dimension under the two-classification condition is $K=(x_1,y_1),(x_2,y_2)...(x_m,y_m),y_i\in\{1,-1\}$。

The main idea of SVM is to find an optimal hyperplane that separates the two types of data, and to maximize the distance between the points on the dataset and the hyperplane. To do this, we can simplify the problem to a nonlinear programming problem. The objective function is:

$$\underset{w,b}{Max}\left\{\frac{2}{|w|}\right\} \tag{14}$$

The restrictions are:

$$s.t.\begin{cases} w^T x_i+b>1, y_i=1 \\ w^T x_i+b<1, y_i=-1 \end{cases} \tag{15}$$

Continue to simplify the optimization of this planning function, and you can get it

$$\begin{gathered} \underset{w,b}{Min}\left\{\frac{1}{2}|w|^2\right\} \\ s.t.\ y_i(w^T x_i+b)>1, i=(1,2...m) \end{gathered} \tag{16}$$

In order to improve the generalization ability of SVM, a relaxation variable is introduced, ξ_i and the formula is:

$$\begin{gathered} \underset{w,b}{Min}\left\{\frac{1}{2}|w|^2\right\}+C\sum_{i=1}^{m}\xi_i \\ s.t.\ y_i(w^T x_i+b)>1-\xi_i, \xi_i>0, (i=1,2...m) \end{gathered} \tag{17}$$

This dual problem is simplified by using the Lagrangian solver to obtain the equation

$$L(w,b,a,\xi,\mu)=\frac{1}{2}|w|^{2}+C\sum_{i=1}^{m}\xi_i-\sum_{i=1}^{m}\mu_i\xi_i+\sum_{i=1}^{m}a_i(1-\xi_i-y_i(w^{T}x_i+b))$$
$$s.t.\ a_i>0,\mu_i>0 \tag{18}$$

At this point, the KKT condition is obtained.

$$\begin{cases}a_i\geq 0\ \mu_i\geq 0\\ 1-y_i(w^{T}x_i+b)-\xi_i\leqslant 0\\ \xi_i\geq 0\ \xi_i\mu_i=0\\ a_i\left(1-yi^{(w^{T}x_i+b)-\xi_i}\right)=0\end{cases} \tag{19}$$

In order to save computational costs and improve efficiency, SMOs usually introduce the constant k to the above equation (4.6) transformation, that is, (4.7).

$$\begin{cases}a_i(y_i f(x_i)-1+\xi_j)=0\\ \xi_i(c-a_i)=0\end{cases}\quad i=1,2,\cdots,m$$
$$y_1a_1+y_2a_2=k \tag{20}$$

According to the above equation, we can solve a_i for the sum$b$, and then we can solve the relevant result.

When the original dataset space has a finite number of dimensions and is indivisible, it can be mapped to a high-dimensional space. Therefore, choosing the right kernel function to map to a high-dimensional space is the key to solving the classification problem. In this paper, the cubic polynomial SVM model is used for the first time to divide all samples into 0-10 points and invalid classification, and its kernel function adopts the Gaussian-based kernel function, that is, the formula combines the most accurate and accurate classification points (integer points) The best classification recognition effect was obtained, ***P=11***

$$k(x_i,x_j)=(x_i^{T}x_j)^{d}$$
$$1<(x_i,x_j)=e^{-\frac{|x_i-x_j|^{2}}{2(4\sqrt{p})^{2}}},p=11 \tag{21}$$

The program model is established by the calculation formula, and the parameter tuning optimization is performed by using Matlab's templateTree function, and the final supervision system is obtained, and then the final prediction confidence histogram (Fig.5) is obtained by the cross-validation method of K=10 generation

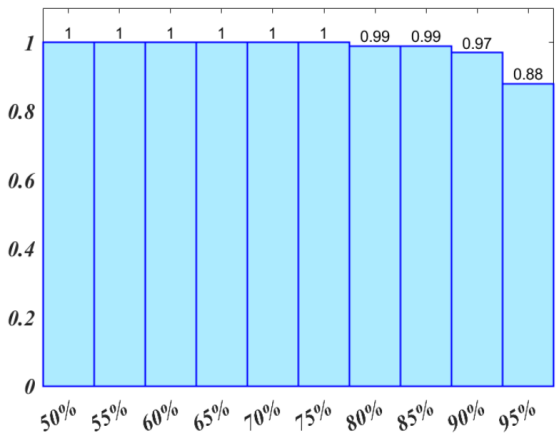

Fig 5. The final prediction confidence histogram

Note: The X-axis represents the confidence interval for each raw data point, and the Y-axis represents the proportion of data points under that confidence interval

The team found that when the confidence interval is 90%~95%, the proportion of data is as high as about 90%, which means that most of the predicted data is within 1 point of the original data, and 81% of the data is within 0.5 points.

By substituting the data to be predicted in the attachment into the model, the following prediction data will be obtained (only show the prediction of the voice satisfaction part of the data in Annex 1), first we show the difference between the predicted data distribution graph and the accurate data (Figure 6):

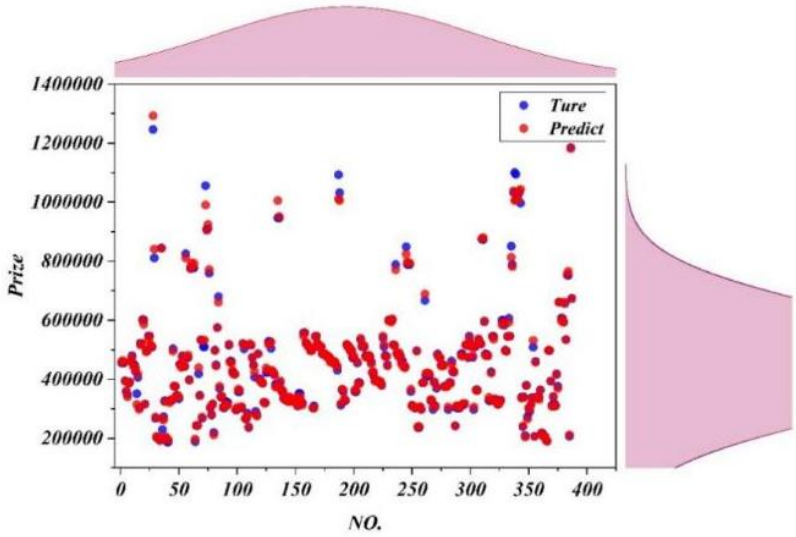

Fig 6. The difference between the predicted data distribution graph and the accurate data

7. The name of model 3

We will use the characteristic importance analysis of random forests to analyze the correctness of our conclusions

An important feature of random forest is that the importance of feature relocation can be evaluated, and the basic idea is that after adding noise values to each feature variable, the prediction accuracy of *RF* is significantly reduced, indicating that the feature change is more important, and we will use this method to perform the final screening of the remaining more important feature values

(1) First, the OOB data is used to test the performance of the generated random forest to obtain an *OOB* accuracy

(2) Then, the noise value is artificially added to the feature change *v* in the OOB data, and then the performance of the random forest is tested with the OOB data after adding noise, and a new *OOB* is obtained Accuracy

(3) The difference between the original OOB accuracy (Gini value) and the *OOB* accuracy after adding noise is used as an important measure of the corresponding feature variable *v.*

Using the characteristic of random forest, the importance of characteristic variables can be ranked, and banks should pay attention to more important information in the process of credit business and try to ensure its authenticity and integrity.

We use the variable importance score and the $V_i Gini$ index as the representation, assuming that there areG_i J features.

$X_1, X_2, X_{3,\cdots}, X_J$ A decision tree, C category, now has to calculate theX_j *Gini* index score for each feature, i.e.$V_j^{G_i}$ *the jth* feature in *RF* The average amount of change in node split impurity across all decision trees.

The *Gini* index of the i-th tree node *q* is calculated as:

$$CI_q^{\ i} = \sum_{C=1}^{|c|} \sum_{C' \neq 1} P_{qc}^{(i)} P_{qC'}^{(i)} = 1 - \sum_{c=1}^{|c|} \left(p_{qc}^{(i)}\right)^2 \tag{22}$$

Intuitively speaking, it is the probability that two samples are randomly selected from node Q, and their class labels are inconsistent. The importance of the feature at node q of x_j*the ith* tree, that is, the *amount of Gini* index change before and after the branching of node *q* is:

$$V_{ijq}^{(G_i)} = CI_q^{(i)} - CI_l^{(i)} - CI_r^{(i)} \tag{23}$$

where, respectively represents the Gini index of the two new nodes after the branch, and if the node where the feature appears in the decision tree i is a set Q, then it is in i The importance of the tree is:$X_j X_j$

$$V_j^{(G_i)} = \sum_{q \in Q} V_j^{(G_i)(i)} \tag{24}$$

Assuming that there are I trees in *RF*, then

$$V_j^{(G_i)} = \sum_{i=1}^{I} V_j^{(G_i)(i)} \tag{25}$$

Finally, all the obtained importance scores are normalized

$$V_j^{(G_i)} = \frac{V_j(G_i)}{\sum_{j'=1}^{J} V_{j'}^{(Gi)}} \tag{26}$$

The three methods of verifying importance are weighted, that is, the final importance indicator chart (Figure 7) is as follows:

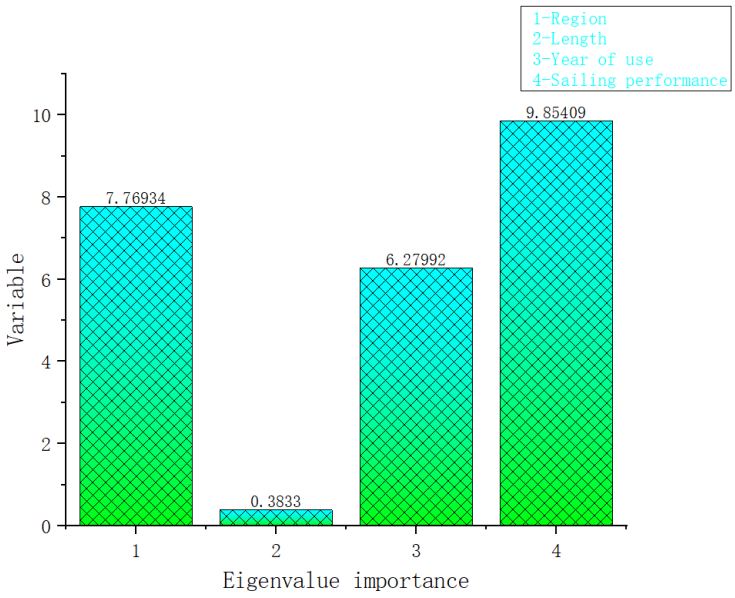

Fig 7. Feature importance analysis diagram

The price of a second-hand sailboat can be affected by geographical factors.

Location: The location of the sailboat can have a significant impact on its price. For example, a sailboat located in a popular sailing area or a region with high demand for sailboats will likely be more expensive than one located in a less desirable location.

Climate: The climate in which the sailboat is located can also affect its price. Sailboats located in areas with harsher weather conditions or shorter sailing seasons may be less expensive than those in areas with mild weather and longer sailing seasons.

Availability: The availability of second-hand sailboats in a particular area can also affect prices. If there are few second-hand sailboats available in a particular location, prices may be higher due to increased demand.

Maintenance: The level of maintenance and upkeep that has been performed on the sailboat can also affect its price. A sailboat that has been well-maintained and is in good condition will generally be more expensive than one that has been neglected.

Market demand: Finally, market demand for sailboats in a particular area can have a significant impact on prices. If there is high demand for sailboats in a particular area, prices may be higher due to increased competition among buyers.

8. Sensitivity Analysis

Sensitivity analysis is an important part of evaluating the performance and robustness of a predictive model. In the context of a supervised regression model for price prediction, sensitivity analysis involves assessing the impact of changes to input variables on the model's output.

To perform sensitivity analysis on a price prediction model, we can vary one or more input variables and observe the corresponding changes in the predicted price. For example, we can perturb the values of the input features by a fixed amount or percentage and record the resulting changes in the predicted price.

One commonly used technique for sensitivity analysis is to calculate the partial derivatives of the predicted price with respect to each input variable. These partial derivatives represent the sensitivity of the predicted price to changes in each input variable, and can be used to identify the most important input variables for the model.

Another approach is to use Monte Carlo simulations to generate random perturbations to the input variables and observe the resulting changes in the predicted price. By repeating this process many times, we can obtain a distribution of predicted prices and assess the sensitivity of the model to changes in the input variables.

Overall, sensitivity analysis can help us identify which input variables have the greatest impact on the predicted price, and how much the predicted price changes in response to changes in these variables. This information can be useful for understanding the limitations of the model and for making decisions based on its predictions.

9. Model Evaluation and Further Discussion

9.1. Strengths

a) Accurate Predictions: Supervised regression learning algorithms can produce highly accurate predictions. By using historical data to train the model, it can learn patterns and relationships in the data, which can then be used to make accurate predictions on new, unseen data.

b) Generalization: Supervised regression learning algorithms can generalize well to new data, meaning that they can make accurate predictions on data that they have not seen before. This is because the algorithms learn the underlying patterns in the data, rather than just memorizing the training data.

c) Flexibility: Supervised regression learning can be applied to a wide range of problems, including predicting stock prices, weather patterns, and disease outcomes. The flexibility of these algorithms makes them useful in many different industries and fields.

d) Interpretability: Many supervised regression learning algorithms are easy to interpret, which means that we can understand why the algorithm makes a certain prediction. This interpretability can be useful in decision-making and can help build trust in the algorithm's predictions.

e) Incremental Learning: Supervised regression learning algorithms can learn from new data over time. This means that as new data becomes available, the model can be updated to incorporate the new information, leading to improved predictions over time.

9.2. Weaknesses

Data Requirements: Supervised regression learning requires large amounts of high-quality data to train the model. In some cases, it may be difficult or expensive to obtain the necessary data, which can limit the accuracy and generalization ability of the model.

9.3. Further Discussion

To promote the prediction of used sailboat models, the following strategies can be considered:

Content Marketing: Creating and publishing high-quality content such as blog posts, articles, and videos that provide valuable insights and information about used sailboat models, market trends, and pricing predictions can help build credibility and trust with potential users.

Social Media Marketing: Social media platforms such as Facebook, Instagram, and Twitter can be used to reach a wider audience and engage with potential users. By sharing useful content and interacting with followers, social media can help build a community around the sailboat prediction model.

Search Engine Optimization (SEO): Optimizing the sailboat prediction model's website for search engines can help increase its visibility and attract more traffic. This can be done by using relevant keywords, creating high-quality content, and building backlinks to the website.

Influencer Marketing: Collaborating with sailing influencers and bloggers who have a large following can help reach a wider audience and build credibility. By having influencers review and promote the sailboat prediction model, it can help increase awareness and interest.

Paid Advertising: Paid advertising on platforms such as Google Ads or Facebook Ads can help target specific audiences and increase visibility. This strategy can be particularly effective for reaching users who are actively searching for used sailboat models or related topics.

Overall, a combination of these strategies can help promote the sailboat prediction model and attract potential users. By providing value and building trust with the sailing community, the model can become a valuable tool for predicting the value of used sailboat models.

References

[1] Fan Xia. A brief discussion on the construction of social stability index system [J]. Journal of Hubei College of Economics, 2005, 3(2):4.

[2] JC Zhu, YM Yang, JF Huang. Research on the Competitiveness Differences and Coordinated Development of Provincial Central Cities in Central Region [J]. Geographical Research and Development, 2010, 29(3): 6

[3] Wenyuan Niu. Social physics and early warning system of social stability in China [J]. Proceedings of the Chinese Academy of Sciences, 2001(01): 15-20.DOI: 10.16418/j.issn.1000-3045.2001.01.004.

[4] ZN Li, WQ Pan. Econometrics. 5th edition [M]. Higher Education Press, 2020.