Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2332913

Imbalance Price Prediction Model for Used Sailboats

Summary

Sailboats have become a popular luxury in recent years, attracting the attention of the wealthy and enthusiasts alike. As a result, used sailboats are highly sought after in the sailboat market. The price of used sailboats is affected by many factors, including the age, brand, size, configuration, maintenance condition and market demand of the boat. And as a professional sailboat trading broker, a boat broker needs to consider all these factors to determine the market price and trading strategy of used sailboats and provide the best trading plan.

The challenges of the second-hand sailboat pricing model come from 1. the imbalanced price distribution and 2. the ambiguity of the value of second-hand sailboats. Therefore, our job is to use mathematical models to fit the price data of various sailboats in different regions as much as possible, and to be able to have good results in Hong Kong market as well.

To complete our work, we sequentially performed data acquisition, data cleaning, data exploration, and feature transformation. In data exploration, we selected some features and performed **subsampling** to tackle with the **Imbalanced Price Distribution** problems. We have both continuous-valued and discrete-valued data in our features. For continuous-valued data, we performed a normalization operation, and for discrete-valued data, we used one-hot encoding.

Considering that the features of sailboats have different attributes, we encode the features respectively with neural network **Auto-Encoders** and a **Time-Encoder** to extract multiple features, which are integrated into a model similar to the **Transformer** to realize **Feature Fusion**. These components make up our **model I**. We predict a range of sailboat prices, and propose a new result evaluation method, i.e., **in-range score** to overcome **the ambiguity of the value** of second-hand sailboats. We analyze the results of the model, and our model performs well.

After obtaining the economic data for Hong Kong, we established our **model II**, i.e. **Register Model**, which accepts economic data from other regions. With this model, we analyze the Hong Kong (SAR) market. A subset of sailboats with high information content, divided into monohull and catamaran, is selected from the data set and compared with the collected data on the Hong Kong(SAR) market, respectively, and we derive the effect of using the model in Hong Kong, and the effect on monohull and catamaran boats, respectively.

Then, we show some informative conclusions that we got during the data analysis, including consumer behavior, market competition, etc. Finally, we compile the results of the model used in Hong Kong into a report for brokers' reference.

It has been verified that our model has achieved excellent results. But there is still room for improvement, we will open source the entire project for follow-up work.

Keywords: Regression Analysis; Imbalance Learning; Auto-Encoder; Transformer; Feature Embedding; Used Sailboats

Contents

1 Introduction

1.1 Problem Background

Due to its high production cost, high navigation cost and uniqueness, sailboats are gradually becoming popular as luxury goods, and more and more people choose to buy used sailboats. Compared with new boats, used boats are cheaper, and they also have a special style and fun of renovation. There are many factors that affect the price of a used sailboat, including but not limited to market demand, boat condition and brand, etc. The used sailboat market is huge, estimated at 4 billion globally, and will grow further as technology advances and people become more interested in water sports and tourism activities.

In the process of used transactions, ship brokers play an important role, and need to analyze the market and evaluate the ships to reach a reasonable transaction. Therefore, ship brokers need to know more about the market, and make informed decisions based on consideration of various factors affect prices, so as to ensure that the purchased sailing boats can be traded at a rational price.s

1.2 Restatement of the Problem

In this problem, we have a raw data set, which contains the sale price of sailboats in different countries and regions. In addition, we collect other features of various sailboats and economic data of each region, and use all these data to solve the following questions:

(1) Use the data on features of sailboats and regional economic to model a reasonable explanation of sailboat prices, quantitatively and qualitatively analyze the impact of regional economies on prices, and use the model to predict and analyze the Hong Kong (SAR) market with the addition of economic data.

(2) Based on our model, we address the following needs:

- **Optimality:** Our model achieves the best possible performance on a particular task or problem.
- **Robustness:** Our model maintains its performance and makes accurate predictions on new, unseen data even in the presence of noise, outliers, and other types of perturbations or irregularities in the data.
- **Practility:** We visualize the output of our model to discuss how useful it is in the Hong Kong (SAR) market.

1.3 Our Work

In order to maximize the accuracy of our mathematical model on the dataset, we use the following workflow:

First, Section 3 is the data mining and preprocessing part where we mine, clean, preprocess and select highly correlated features of our data as training features. Afterwards, we find that the impact of each data feature on the listing price of sailboats is highly complex ,and due to the complexity of the sailboat market, the listing price of sailboats is not stable.

The above analysis explains the necessity to choose a good model. In Section 4, we stand in the perspective of sailboat brokers and purchasers to get a model. After considering the feature

factors, we get an access to a model including Auto-Encoders, Time-Encoder, Self-attention Feature Fusion and XGBoost, as the main body of our model I, solving Problem 1.

Then, we discuss the effect of regions on sailboat prices in Section 5. Through analyzing all variants, we solved Problem 2.

After collecting the relevant economic data, Section 6 introduces a Register Model used to put the new region information in the model I, and presents the results of our application of the integrated model to the analysis of the Hong Kong (SAR) market. We select an informative subset of sailboats from the dataset, divided into monohulls and catamarans, and analyze them separately in comparison with the collected data of the Hong Kong (SAR) market.

Section 7 presents some interesting conclusions we obtained for Problem 3 about the relationship between prices and other features.

In Section 8, we compile the results in Section 6 into a report that addresses Problem 4.

Last but not least, section 9 show the strengths and weaknesses of our work. Section 10 is the conclusion and our expectations about future work.

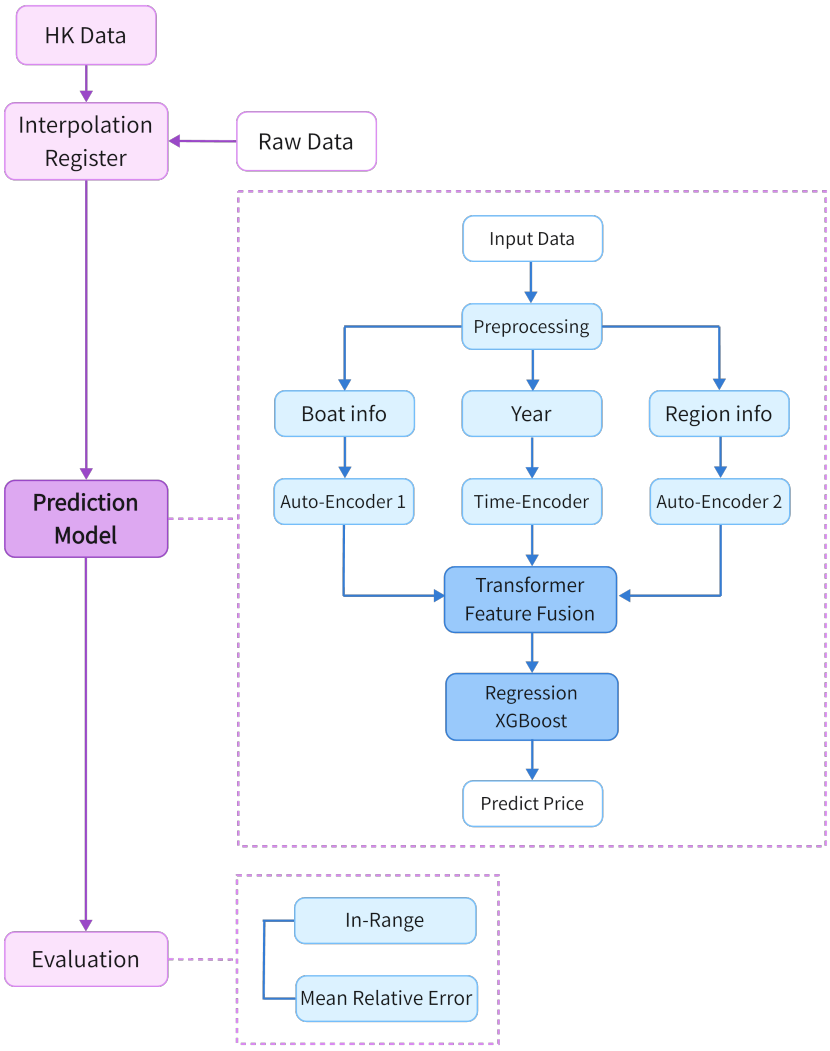

Figure 1: Workflow

1.4 Innovation Points

- We deal with the essential, temporal and spatial factors affecting sailboat prices separately to perform feature fusion for the prediction.
- For the distribution of prices, we propose a regression method to handle data with imbalanced distribution.
- Our model has a self-attention mechanism, which has good results on feature fusion tasks.
- Our model outputs a price distribution interval when forecasting, and in-range score which is a special model evaluation method– is used to test the performance of our model.

2 Assumptions and Notations

2.1 Assumptions

To simplify our model and reduce the design and computational complexity of the work, we make the following assumptions, each of which is properly justified.

- **Stationarity:**We assume that the data is stationary, meaning that the statistical properties of the data (such as the mean and variance) do not change over time. This can simplify the modeling process and make it easier to make predictions.
- **Independence:** We assume that the samples of data are independent of each other, meaning that the features of one sample does not depend on the features of another sample. This can make the modeling process more straightforward and help avoid issues with multicollinearity.
- **Rationality:** We assume that the data used in the analysis is reasonable and it can support the conclusions drawn from the analysis.
- **Sufficiency and Representativeness:** We assume that the amount of data used in the analysis is adequate for the research question or problem being studied. In other words, there is enough data to draw meaningful conclusions and make accurate predictions.

2.2 Notations

Symbols	Description
p	the Ground Truth Price
$\hat{p}$	the Predicted Price
ϕ	Multiple Features
T	the Set of Features
t	the Feature in the T

where we define the main parameters while specific value of those parameters will be given later.

3 Data Mining and Preprocessing

3.1 Data Collection

In addition to the original data, we also collected some information on sailboats in various regions, including displacement, LOA, LWL, draft(max), etc.

Some data sources are listed in Table 3.1.

Table 1: Data Source List

Data	Website
GDP	World Bank(https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2021&start=2018)
GDP per capita	World Bamk(https://data.worldbank.org/indicator/NY.GDP.PCAP.CD)
HongKong Boats Data	Asia Boating Ltd(https://www.asia-boating.com/)
	Asia Yachting(https://asiayachting.net/)
	Hong Kong Yachting(https://www.hongkongyachting.com/)
Addtional Data	Sail Boat Data(https://sailboatdata.com/)
Luxury Goods	Luxury Goods(https://www.statista.com/outlook/cmo/luxury-goods/france#revenue)

3.2 Data Cleaning

The data we get contains missing data or other issues that require some data cleaning prior to analysis.

3.2.1 Missing Value Processing

First of all, since the amount of data is large enough, when we encounter a sample lacking a small amount of feature data, we use an interpolation algorithm to compensate the data, that is, we first use the existing data points to establish a suitable interpolation function, and then leverage the function value $f\left(x_i\right)$ to replace missing values. For samples that lack a large number of feature data, we directly delete them to prevent failures in model fitting.

3.2.2 Outlier Processing

We choose **Isolation Forest** algorithm for outlier detection, instead of 3-Sigma algorithm used by many data analysis problems.

The Isolation Forest algorithm implements outlier detection by constructing a random forest, which is an ensemble learning method based on decision trees, each of which is trained by randomly selecting samples from the dataset. The main idea of the Isolation Forest algorithm is that outliers have shallow depths in the decision trees, as their differences from other data points are more easily discovered. Therefore, in the random forest, if a data point has a shallow depth in most decision trees, it is likely to be an outlier.

The advantages of the Isolation Forest algorithm include efficiency and effectiveness for high-dimensional data, however, it also has some disadvantages, such as performing poorly on dense datasets. Experimental results have shown that the Isolation Forest performs well in this data processing situation.

3.3 Data Exploration

3.3.1 Feature Distribution

We visualize the distribution of each feature in the data. Through visual analysis, we can understand the characteristics and structure of the data, and check the effect of data cleaning. This information can help us better understand the data and provide guidance for subsequent data processing and model selection. Due to the peculiarities of the price distribution, we show it in Figure 2. It can be seen from the Figure 2 that price has a long-tailed distribution, which

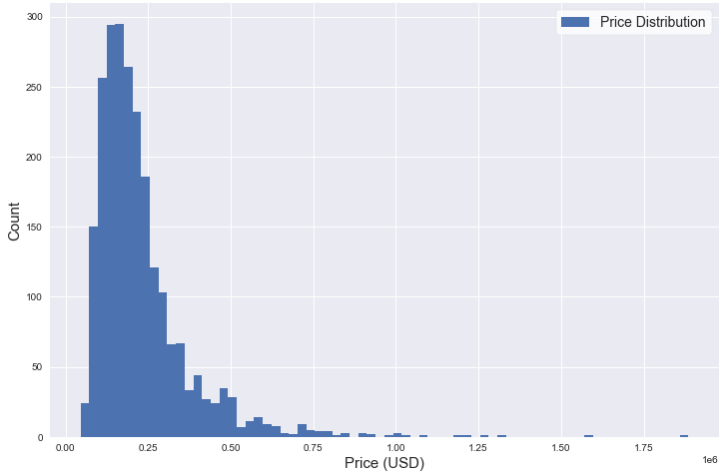

Figure 2: Price Distribution

will reduce the effect of our model. To solve this Imbalance Learning problem, we **subsample** the data of price.

3.3.2 Correlations Between Features

We calculate the correlations between individual feature values and plot them in Figure 3. This information can help us to select many features data that are sufficiently related to the listing price of sailboats, and process model optimization and result analysis.

From the Figure 3, we can get some important features, including length, LOA, etc. These have good correlation with price so they are selected as features in the training data.

3.4 Feature Transformation

In machine learning, continuous features are often standardized, or normalized, to ensure that they have similar importance during model training and do not adversely affect the model due to differences in value range. For discrete features, encoding is required, the purpose of which is to convert discrete features into a form that machine learning algorithms can process. These encoded features can be used as input to the model, thereby improving the performance and accuracy of the model.

The input data features in our regression model contain both continuous and discrete features. We need to normalize and encode them respectively.

Continuous Features Transformation:

We normalize the length, year and beam, etc.

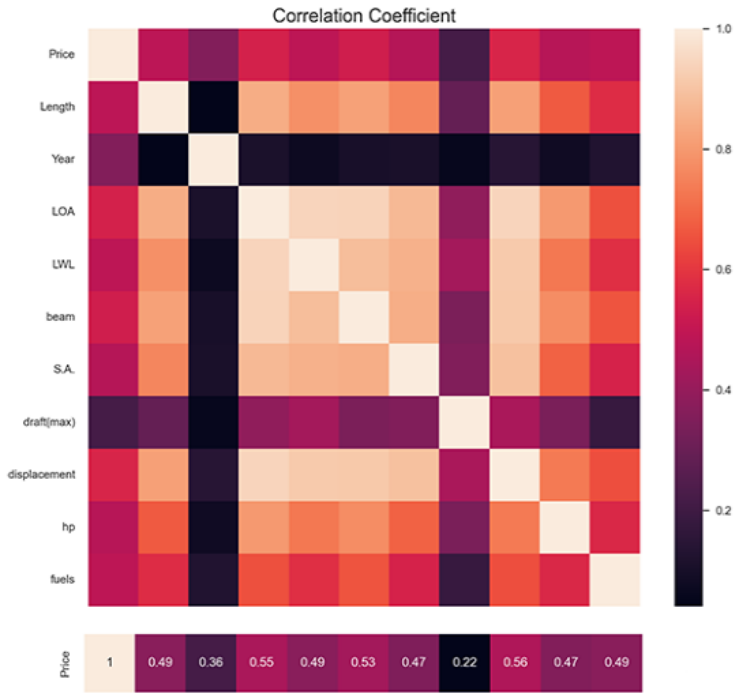

Figure 3: Correlations Between Features

Discrete Features Encoding:

We perform one-hot encoding for discrete features such as make, variant, etc.

4 Model I: Encoders, Feature Fusion, XGBoost Regression

In order to figure out the relationship between the listed prices in the sailboat market and various features of the sailboat, we stand in the perspective of sailboat buyers and brokers, and select a model that can accurately explain the price in the data and has good generalization performance.

Considering that the ship's make, variant, length, etc., correspond to the attributes of the sailboat itself, the year of the sailboat corresponds to the time attribute of the sailboat, and the region corresponds to the space attribute of the sailboat, we use the neural network Auto-Encoder and Time-Encoder we proposed to encode these three types of features respectively. Three kinds of multi-features are integrated in one Transformer model. The Transformer model uses a multi-layer attention mechanism in the encoder, which can encode various feature relationships into a richer representation, thereby improving the generalization ability and performance of the model.

The brief structure of our model is illustrated in Figure xxx

4.1 Auto-Encoder

Auto-Encoder is an unsupervised neural network model that can learn the hidden features of the input data, which is called encoding, and at the same time use the learned new features to reconstruct the original input data, called decoding.

The structure of the Auto-Encoder is shown in Figure 4. The Auto-Encoder model is mainly composed of an Encoder and a Decoder. Building an Auto-Encoder needs to complete the

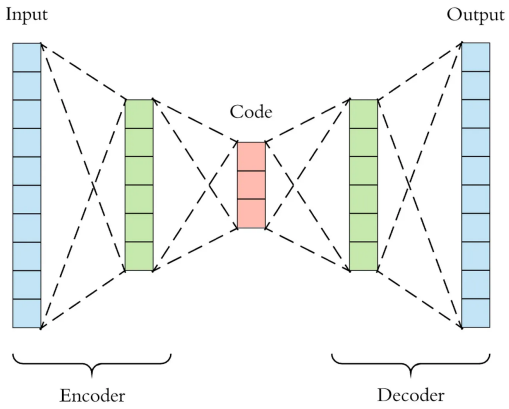

Figure 4: Structure of the Auto-Encoder (Matthew Stewart 2019)

following three tasks: building an Encoder, building a Decoder, and setting a loss function to measure the information lost due to compression.

$$y = h\left(x\right) \tag{1}$$

The Encoder is responsible for receiving the input x and transforming the input into a signal y through the function h.

$$r = f\left(y\right) = f\left(h(x)\right) \tag{2}$$

The Decoder takes the encoded signal y as its input, and obtains the reconstructed signal r through the function f.

Define the error e as the difference between the original input x and the reconstructed signal r, ie. $e = x - r$. The goal of network training is to reduce the mean squared error (MSE), and the error is back-propagated back to the hidden layer.

In our model I, it is used to encode the boat's information and region information.

4.2 Time-Encoder

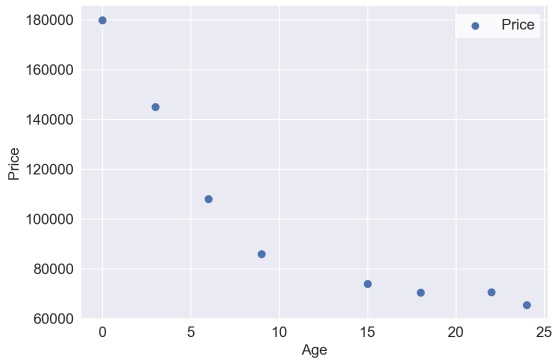

Figure 5: Price Changes over Time

In our model, Time-Encoder is used to encode the time feature information. We observed that different second-hand sailboats showed a consistent decay law in price changes over time, as shown in Figure 5. We explored this decay law and modeled it:

$$P(t) = P(0)(\Delta P e^{-\beta t} + 1 - \Delta P) \tag{3}$$

In order to prove our correctness, we conduct regression analysis on $ln(\frac{P(t)}{P(0)} - 1 + \Delta P)$ and t, and calculate $R^2 = 0.94$. The experimental results are as follows:

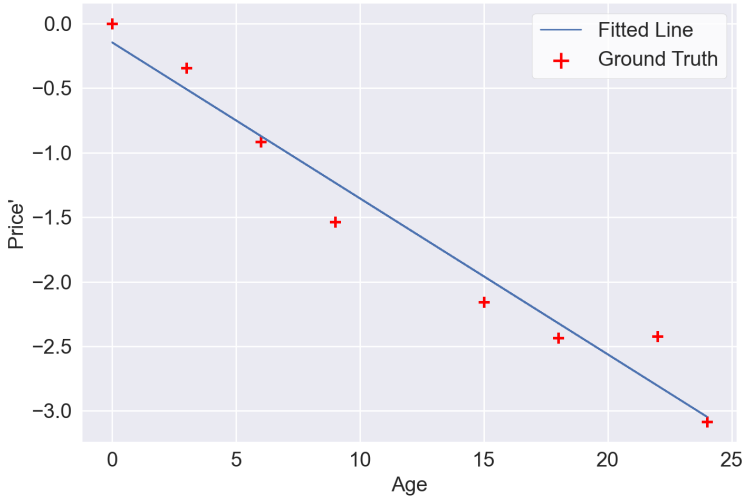

Figure 6: Regression on $ln(\frac{P(t)}{P(0)} - 1 + \Delta P)$ and t

4.3 Self-attention Feature Fusion

We propose a multiple feature fusion method based on self-attention mechanism, its structure is shown in Figure 7.

In practice, given n features, the multiple features are stacked together as ϕ_m, which is then embedded with three different linear layers and self-attention mechanism is applied:

$$attention(\phi_q, \phi_s, \phi_t) = softmax(\frac{\phi_q^T \phi_s}{\sqrt{d_k}})\phi_t \tag{4}$$

where $\phi_q = \phi_m W_q$, $\phi_s = \phi_m W_s$,$\phi_t = \phi_m W_t$ denote the query, source and target feature embedded by learnable weights W, and $\phi_m \in \mathbb{R}^{n \times d_k}$ with d_k as embedding size.

The dot-product result is divided by $\sqrt{d_k}$ to present the gradient vanishing problem. For linear weights W, we use a multi-head attention strategy., i.e. $W_q, W_s, W_t \in \mathbb{R}^{n_{head} \times d_k}$ which encode the multiple features into n_{head} different embedding subspaces, allowing the model to better notice different patterns jointly.

Due to using the gradient descent algorithm to train the self-attention module, it is necessary for the subsequent model to be differentiable. Therefore, we use a multilayer perceptron (MLP) to connect the fused features and the output of the model. It should be noted that the MLP is only used to train the self-attention module, not to predict the price. We will select other regression models to predict the price of second-hand sailboats.

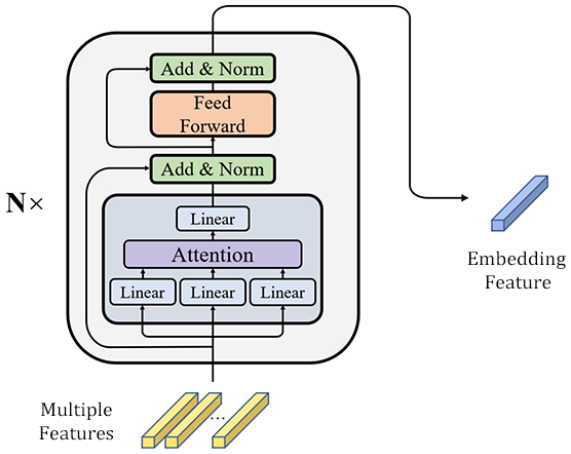

Figure 7: Structure of Self-attention Feature Fusion (Yang Zheng *et.al* 2021)

4.4 Regression XGBoost

At the end of the model, we use XGBoost to regress the fused features to get the final price.

XGBoost is also a kind of gradient boosting tree model. It generates models serially, and takes the sum of all models as the output. XGBoost greedily chooses whether to split the node according to whether the loss function is reduced. At the same time, XGBoost adds regularization, learning rate, column sampling, and approximate optimal segmentation points to prevent overfitting. The hyperparameters are set to the default values of the sci-kit learn package.

Define a loss function $\mathcal{L}(y, \hat{y})$, which is a differentiable arbitrary loss function. For a tree structure q, define

$$g_i = \frac{\partial \mathcal{L}\left(y, \widehat{y_{t-1}}^{(i)}\right)}{\partial \hat{y}_{t-1}^{(i)}}, h_i = \frac{\partial^2 \mathcal{L}\left(y, \widehat{y_{t-1}}^{(i)}\right)}{\partial \hat{y}_{t-1}^{(i)^2}} \tag{5}$$

and

$$G_i = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_j \tag{6}$$

Its loss function is as follows:

$$\begin{aligned} Obj_t &= \sum_{i=1}^{m} \mathcal{L}\left(y^{(i)}, \hat{y}_t^{(i)}\right) + \Omega\left(f_t\right) \\ &= \sum_{i=1}^{m} \left[\mathcal{L}\left(y^{(i)}, \hat{y}_{t-1}^{(i)}\right) + g_i f_t\left(x^{(i)}\right) + \frac{1}{2} h_i f_t\left(x^{(i)}\right)\right)^2\Bigg] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_j + \lambda\right) w_j^2\right] + \gamma T \\ &= \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2}\left(H_j + \lambda\right) w_j^2\right] + \gamma T \end{aligned} \tag{7}$$

where m is the total number of samples, t is the current base learner index, T is the number of leaf nodes of the current base learner, I_j is the sample index set of the j_{th} node, ω_j is the weight of the j_{th} node. If the sample x is on the j_{th} node, then $f(x) = \omega_j$.

XGBoost then updates the learner through a greedy method to find the best tree structure p,so as to find the optimal solution.

4.5 Results of Our Model

4.5.1 Basic Description

We use the original dataset and our model for regression and prediction. After sorting the samples by price, we draw a scatter plot of the ground truth price p and the predicted price $\hat{p}$, as seen in Figure 8. And we evaluated our prediction results using the MRE and in-range score

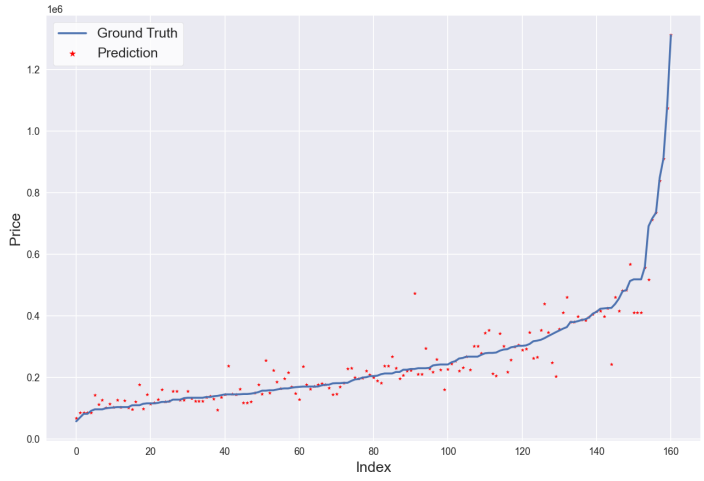

Figure 8: the Ground Truth Price p and the Predicted Price $\hat{p}$

methods we preposed on the Original Dataset and the Additional Dataset, as follows:

Table 2: Best Performance of Model

	MRE	in-range score
Original Dataset	0.070	91%
Addtional Dataset	0.078	88%

4.5.2 Ablation Experiment

In this section, we use the idea of ablation experiments in model evaluation to determine whether our tricks in model training is useful.

In ablation experiments, we directly train the model without using subsampling. The model get retrained.We compare its performance to the original model. We show the results without subsampling in Figure 9

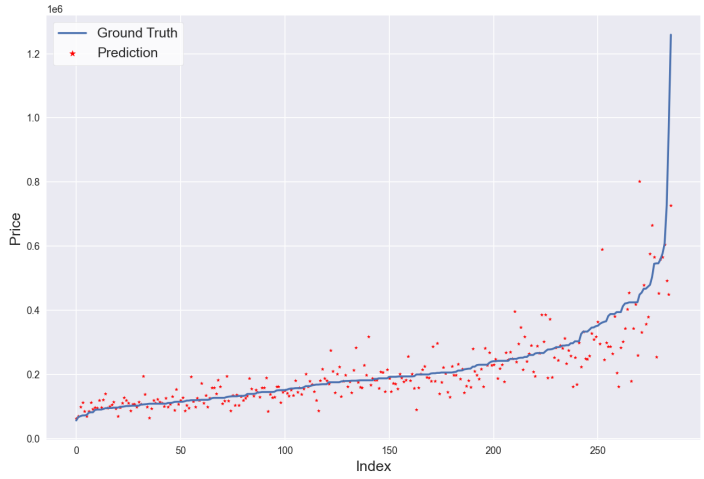

Figure 9: the Ground Truth Price p and the Predicted Price $\hat{p}$ without Subsampling

Comparing Figure 8 and Figure 9, we find that the results without subsampling predict poor performance when the true price is high. This illustrates the importance of downsampling from another perspective.

We also compare our feature fusion methods, and the Table 3 shows that our self-attention based Fusion has the best results.

Table 3: Self-attention based feature fusion

	Raw Dataset		Additonal Dataset	
	MRE	in-range score	MRE	in-range score
Concat-based Fusion	0.102	80%	0.112	81%
Average-based Fusion	0.114	78%	0.120	76%
Self-attention based Fusion	**0.079**	**89%**	**0.084**	**88%**

4.5.3 Analysis of Robustness

We add noise to the original data and put some new data into the model, then we looked at the model performance to confirm whether the model is robust.

Some of our results are illustrated in

5 Analysis of Region Factors

Based on the analysis above, the prediction results of our model is affected by other features, making us aware of the possible existence of a region factor, so we go through the part below to verify it.

5.1 Price Differences in Various Regions

We try to show that there are differences between regions. First, we selected the 8 countries/regions with the highest volume and visualized the distribution of their trading volume, and we could clearly observe a significant difference in the listing prices between different regions as shown in Figure 10, confirming the suspicion that there is a regional factor. Next, we

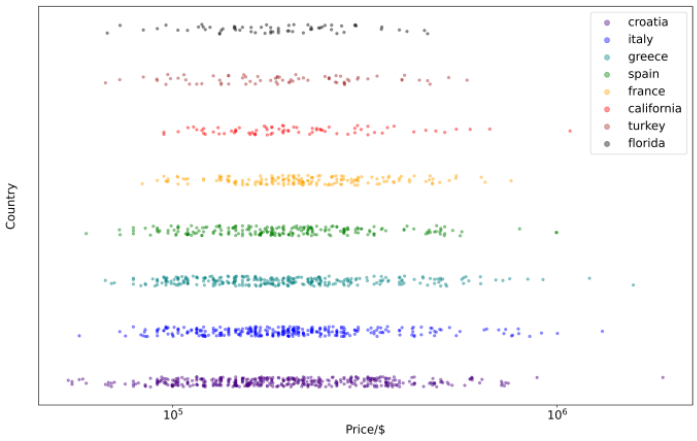

Figure 10: the Listing Prices between Different Regions

explored the existence of regional effects by means of Statistical Tests. We first found through the Kolmogorov–Smirnov Test that the distribution of transaction amounts in each region did not follow a normal distribution and could not be tested by the traditional Student's T Test so we used the following Nonparametric Tests:

1. the Null Hypothesis is that the distribution of sailing transaction prices is consistent across regions, i.e., there is no regional effect, and the alternative hypothesis is that there is a regional effect.

2. a two-parameter Kolmogorov–Smirnov Test was conducted on two groups of data from the United States, Europe, and the Caribbean, respectively.

3. a multi-parameter Kruskal-Wallis Test is conducted on the data from the above three regions simultaneously.

4. test on the raw dataset and the additional dataset expanded to obtain p-values less than 0.01.

5. reject the Null Hypothesis and adopt the Alternative Hypothesis that there is a regional effect.

Table 4: P-values Obtained

	K-S testing P-value(U+E)	K-S testing P-value(U+C)	K-S testing P-value(C+E)	K-W testing P-value(U+E+C)
raw dataset	1.38E-04	2.98E-05	5.61E-03	1.49E-06
additional dataset	1.76E-04	4.09E-04	8.15E-03	1.68E-05

5.2 Discussion about the Relationship between Regions and Sailboat Variants

Similar differences also appear between different Make, as shown in Figure 11.

Inspired by this, we started to think about the relationship between different products and regional effects (we explored whether the regional effects are consistent across sailboat variants)

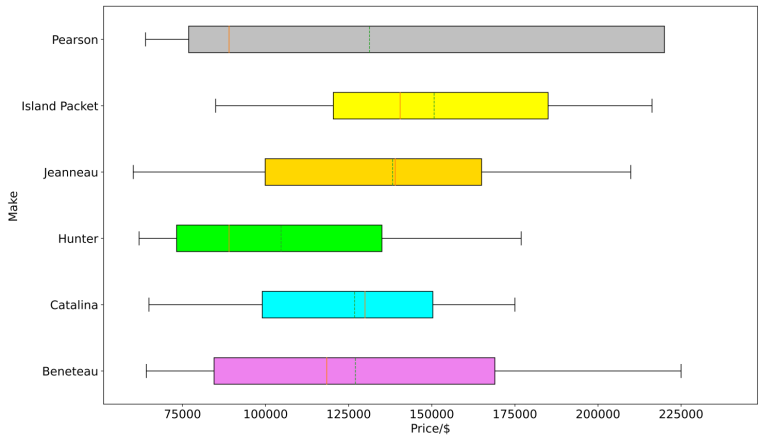

Figure 11: the Listing Prices between Different Make

1. the Null Hypothesis is that the distribution of trade volume is consistent across regions for the same sailboat variant, i.e., the regional effect is consistent across sailboat variants, and the Alternative Hypothesis is that the regional effect is not consistent across sailboat variants.

2. the Kruskal-Wallis Test was performed on the nine most traded sailing vessel variants in the United States, Europe and the Caribbean.

3. the obtained p-values are plotted in the Figure 12 .

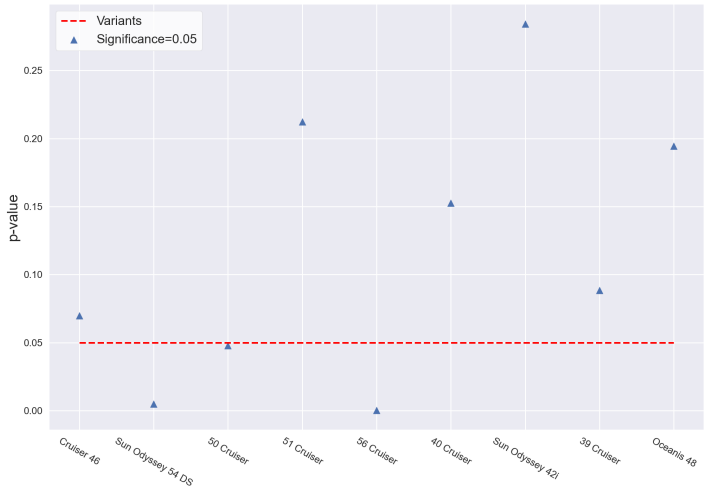

Figure 12: the Obtained P-values

It can be seen that the regional effect is inconsistent for all sailboat variants: for the overall p-value less than 0.01 indicates the presence of a regional effect, as well as for some variants; but for some other variants, the p-value is bigger than 0.05, representing that for these sailboat variants, the regional effect does not exist. This indicates that there are differences in the preferences and consumption ability of consumers in different regions, and their purchasing power differs for different variants.

6 Model II: Register Model and Its Application in HK Market

6.1 Register Model

There are a lot of feature information related to regions. We collect the GDP, GDP per capita, import and export tariffs and coastline length data of relevant regions in various countries as the input of our Register Model.

Specifically, we train a linear regressor and use the above data (note that we use the root of the coastline length divided by the area, i.e. the characteristic radius, as the input instead of the coastline length itself) to regress to get the luxury trade volume per capita of each country and region , after the training is completed, we can get the weight vectors of different countries and regions. We will use the obtained weight vectors to characterize the regional features and incorporate Hong Kong Register into regions in original data.

6.2 HK Market Analysis

We use these vectors to measure the correlation in the average transaction volume of luxury goods between Hong Kong and other countries and regions, and quantitatively display the correlation on a map, as shown in the Figure 13. The darker the color in the figure, the higher the correlation.

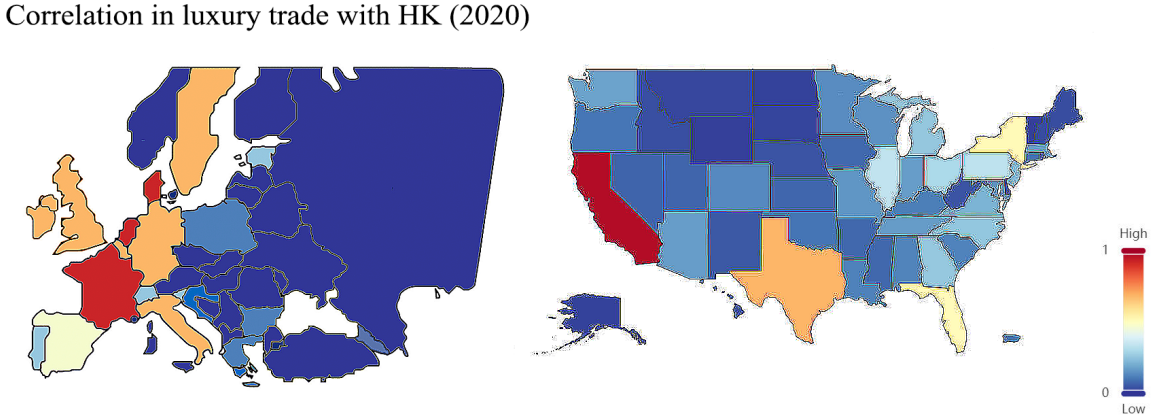

Figure 13: the Correlation between HK and Other Countries and Regions in the Average Transaction Volume of Luxury Goods

We regard the correlation as the Euclidean distance, and determine the feature data of Hong Kong through the size of the Euclidean distance. We propose two methods, nearest neighbor and linear weighting. The nearest neighbor method directly takes the feature data of the region closest to HK as the feature data of Hong Kong, while the linear weighting method selects k special zones closest to HK and linearly weights their features data.

After comparison, we choose the nearest neighbor method to determine the feature data of Hong Kong. So that, Hong Kong obtains feature data of its nearest neighbor France.

After pre-processing, our model divides the input data into three parts, Boat info, Year, and Region info, and then input them into the subsequent processing module through the encoder,

and the model's description of the regional impact is concentrated in the Region info data pre-processing and Auto-Encoder 2. When comparing the regional impact of Hong Kong with other regions, we keep the Boat info and Year info of the other three regions in the original dataset, change the Region info to Hong Kong region info and re-enter it into the model to obtain the prediction data of Hong Kong. We compared this forecast data with the data we found on the web to get the accuracy of our forecast for used sailboat prices in Hong Kong. The prediction results of our model in HK market is shown in Figure 14.

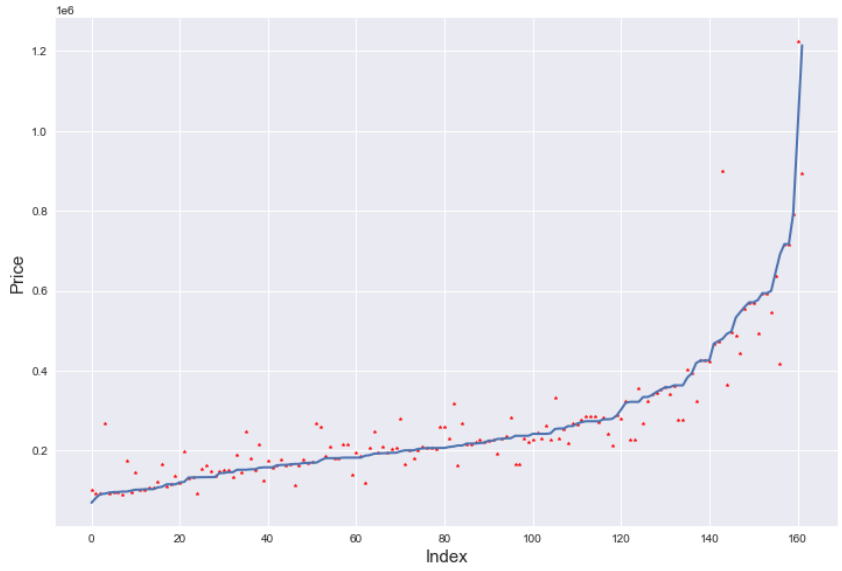

Figure 14: The Prediction Results of Our Model in HK Market

This predicted data was then compared with the original data set for the other three regions, and a statistical test was performed using the KS test, assuming the same regional impact for Hong Kong as for the other regions, and the following p-values were obtained:

Table 5: P-values Obtained

	USA	Europe	Caribbean
P-value	1.4E-5	5.4E-4	7.4E-6

The p-value is less than 0.01, rejecting this hypothesis, and it can be learned that the regional impact of Hong Kong and other regions are not the same.

Then the trained Region info and Auto-Encoder2 of the monohull and catamaran are interchanged, so that the monohull and catamaran use each other's region influence to predict the price distribution. Assuming the same region impact for monohull and catamaran in Hong Kong region, statistical tests are performed using KS test as follows:

Table 6: P-values obtained

	monohulled-input + both-Region-Encoder	catamarans-input + both-Region-Encoder
P-value	2.5E-3	4.7E-3

The p-value is less than 0.01, rejecting this hypothesis, and it can be known that Hong Kong has different regional effects on monohull and catamaran.

7 Some Conclusions about Sailboats Data

1. It is no the secret that the age of a used item has a large effect on its price in this light, at first we used a nonlinear data-fitting to fit the age and price, but found that the fit with the true value was not satisfactory. Afterwards, we went online and found that someone used exponential decay to fit the relationship between price and age. Inspired by which we also look at the relationship between ln(y) and t using the same model plot and found that there was no sharp linearity either. We combine the observed data to reasonably propose an expanded model of the relationship between price and age time encoder The model retains the tendency for the price to decay exponentially with age, but assumes that it does not decay to zero, but to a more stable value which is found to be around 1/3 of the original price after looking at a large amount of data. Recalculating ln(y') versus t after subtracting this value reveals a significant linearity (R^2 = 0.96), and we believe that this stable value can be interpreted as the cost of the vessel, while its profitability gains decay exponentially with time.

2.Since the process of this study involved data on the sale of sailboats in various countries or regions, we naturally thought that there would be some intrinsic connection with the geographic location of the area of purchase. However, after verifying that latitude, longitude, and elevation were largely uncorrelated with trade data, we ventured a guess that it might be related to marine-related factors. So we naturally thought of using the length of a country's coastline L to portray its relationship with its maritime geographic location. However, the area S varies from country to country, and it is unreasonable to use coastline length alone for prediction which was also verified by calculating the correlation coefficient so we introduced a country's area characteristic radius R (R^2=S), and used L/R as a way to portray a country's relationship with the ocean. It was found that this data did correlate with the sailboat sales data, so this value was taken into account in the area coding, resulting in an improvement in model accuracy of about 2

8 Report for the Hong Kong (SAR) Sailboat Broker

We conducted research on the second-hand sailboat sales markets in the USA, Europe, and the Caribbean, delving into the factors that affect the sales prices of second-hand sailboats. Ultimately, we classified these factors into the sailboat's information, age of usage, and regional factors. Based on these insights, we established a machine learning model that predicts the price of a second-hand sailboat by inputting the three types of information mentioned above. Eventually, we achieved a considerable degree of accuracy.

Table 7: Best Performance in USA,Europe,Caribbean

	MRE	in-range score
Original Dataset	0.070	91%
Addtional Dataset	0.078	88%

We explored the impact of regional factors on the price of second-hand sailboats and found that there is a regional effect. Therefore, to generalize our model to Hong Kong, we separately modeled the regional characteristics mentioned above, depicting the region based on factors such as its economy and geography. We also proposed a registration model that can register Hong Kong in our model and obtain reliable prediction results.

In addition, we have found that the second-hand sailboat market exhibits a long-tail price

Table 8: Testing Performance in Hong Kong

	MRE	in-range score
HongKong Dataset	0.093	87%

distribution, meaning that the majority of second-hand sailboats are concentrated in lower price ranges, while a minority are sold at relatively high prices. Please be cautious about this phenomenon, as this imbalanced distribution can result in significant pricing deviations for high-priced second-hand sailboats. For this, we have proposed a regression method to address this imbalance problem, which effectively solves this issue.

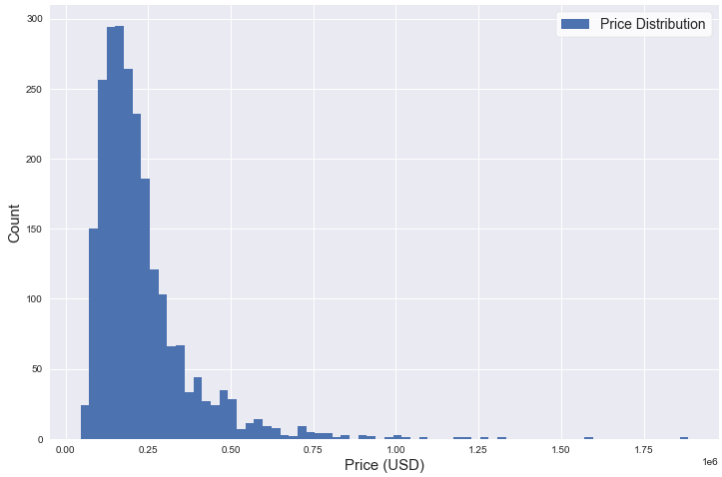

Figure 15: Price Distribution

We have also found that even if three characteristics of a second-hand sailboat are the same, the selling price may still vary. The ambiguity of the price of second-hand sailboats presents great challenges for their pricing. Therefore, our model can predict the minimum and maximum prices of second-hand sailboats, rather than a single point prediction. I believe this provides more information for your pricing strategy.

In summary, we have proposed an imbalanced regression method to address the problem of price imbalance distribution, which solves the issue of inaccurate price prediction for high-priced second-hand sailboats. To eliminate the ambiguity of second-hand sailboat prices, we predict both the highest and lowest prices, providing more information for pricing. Based on these insights, we have established a price prediction model that can effectively forecast the price of second-hand sailboats. To generalize our model on Hong Kong's second-hand sailboat data, we have proposed the Register model, which proves our potential in predicting the prices of second-hand sailboats in Hong Kong.I believe that our model can assist you in gaining a better understanding of the global second-hand sailboat market, including the second-hand sailboat market in Hong Kong, and provide strong support for your price prediction.

9 Strength and Weakness

9.1 Strength

- Improve model performance: We classify various features into different attributes and process them separately and then fuse them, with excellent performance.

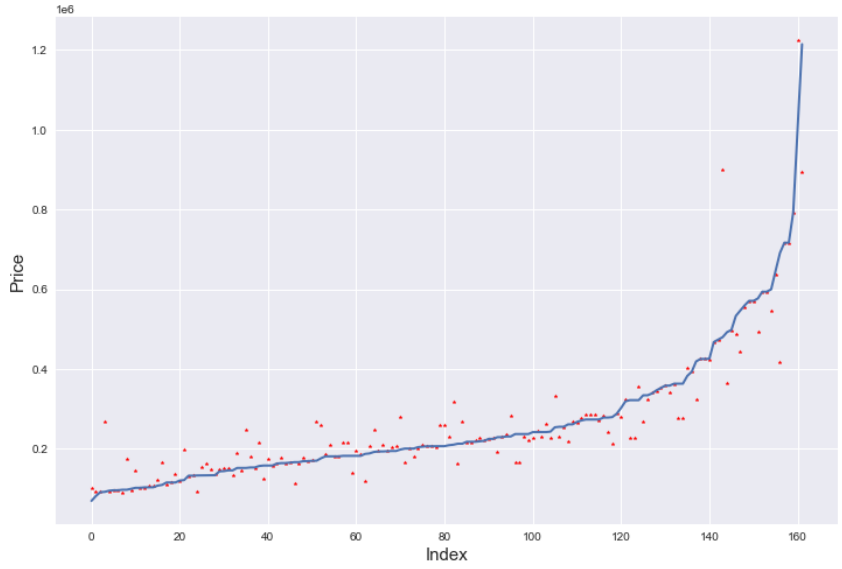

Figure 16: The Prediction Results of Our Model in HK Market

- Strong generalization ability: Transformer's attention mechanism can encode various feature relationships into richer representations, thereby improving the generalization ability of the model.
- Improve model robustness: Our Auto-Encoder is combined with Transformer and XGBoost models, which has strong anti-interference ability.
- Good interpretability: Our feature fusion method can integrate information from different sources, thereby improving the interpretability of the model and making it easier to understand the basis for model decision-making.

9.2 Weakness

- The model structure is complex, and there are design and computational complexities.
- Hard to get better performance on limited data.

10 Conclusion and Future Work

10.1 Conclusion

We collected some sailboat data, and performed a good mining and preprocessing of the data. Using all this data, we first built a model combining Auto-Encoder, Time-Encoder, self-attention based model and XGBoost, which outputs the prediction of sailing price, with excellent results on the dataset. We then analyze the regional impact on sailboat listing prices, discussing the consistency of its impact across all sailboats.

Then we built a small model for adding regional characteristics to the above model, and by regressing the regional characteristic data, we can use its results to measure the correlation between regions. We apply the model to the analysis of the Hong Kong shipping market, and discuss the monohull and catamaran separately. In addition, we also showed some convincing

and interesting conclusions we got during the data analysis process, and finally compiled the analysis results of the Hong Kong ship market into a report for sailing brokers to consider and choose.

10.2 Future Work

Due to the impact of the epidemic and policies on the shipping market in 2020, the data we use for model training in 2020 fluctuates greatly. The current economy is different from the past, but our model still has a good reference value. We believe that with more and better data input and a better model architecture, the prediction of the listing price of second-hand sailboats will become more and more accurate.

References

[1] https://towardsdatascience.com/generating-images-with-autoencoders-77fd3a8dd368

[2] Shihao Gu, Bryan Kelly, Dacheng Xiu, Empirical Asset Pricing via Machine Learning, The Review of Financial Studies, Volume 33, Issue 5, May 2020, Pages 2223–2273,https://doi.org/10.1093/rfs/hhaa009

[3] Fathalla, A., Salah, A., Li, K. et al. Deep end-to-end learning for price prediction of second-hand items. Knowl Inf Syst 62, 4541–4568 (2020). https://doi.org/10.1007/s10115-020-01495-8

[4] Eunsuk Chong, Chulwoo Han, Frank C. Park,Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies,Expert Systems with Applications,Volume 83,2017,Pages 187-205,ISSN 0957-4174,https://doi.org/10.1016/j.eswa.2017.04.030.