Predicting Second-hand Sailboat Market with Computer Modeling (2024)

Anonymous, Unknown Department, Unknown Research Institute

Abstract—**This article presents an in-depth analysis of the used sailboat market, using data analysis techniques to provide insights into pricing and influential factors. The study utilized Python and Excel to process and analyze data, employing different algorithms and statistical methods to build models and make predictions. We prepared a heartfelt and accessible report based on the conclusions and the results we derived from models and experiments for the Hong Kong broker.**

Index Terms—**Ordinary Least Squares, Regression Analysis, ANOVA, Multicollinearity, MLP.**

I. INTRODUCTION

1.1 Problem Background

Despite fluctuating demand and rapidly changing national conditions, the global second-hand sailboat market is showing a steady expanding trend with continuous development of technology, presenting a great business opportunity.

However, as a high-priced luxury goods transaction, asset pricing and market positioning analysis for second-hand sailboats are fraught with difficulties due to the numerous parameters and models of sailboats, significant randomness involved, and relatively obscure specialized data. Therefore, it is necessary to collect valid data in relatively niche fields and conduct the most detailed statistical analysis and forecasting on the numerous and unexpected attributes of the data.

1.2 Restatements and Clarifications

In this study, we obtained pricing data for different types of used sailboats from various manufacturers and regions up to 2019. The data includes information on boat length, place of origin, and year of production. Based on this data, we aim to analyze and address the following issues for an intermediary based in Hong Kong:

1. Develop mathematical models to explain and predict the pricing of sailboats.
2. Use the model to explain regional impact on pricing and validate cross-regional consistency.
3. Discuss the practicality of the established model for the Hong Kong market and retrieve data todetermine if there are different regional effects in the Hong Kong market.
4. Discover interesting information and conclusions.
5. Write a concise, two-page report for the interested Hong Kong sailboat broker.

1.3 Our Approach

Noticing that using the data provided can not yield satisfactory results in the Hong Kong used sailboat market, we collect specific parameters of yachts and macroeconomic data of the regions based on the problem background. After processing the data using Python and Excel, we accomplish the tasks for each question in the following way:

1.To construct an explanatory model, we initially try OLS. However, this approach is susceptible to a severe issue known as multicollinearity. In order to mitigate this problem, we explore two alternative methods. The first involves removing correlated predictors in a heuristic manner to obtain a simplified OLS model, called mini-OLS. The second employs regularization techniques to obtain a stable but biased ENR-OLS model. Additionally, we introduce a neural network model solely for predictive purposes and disregard its potential for explanation.

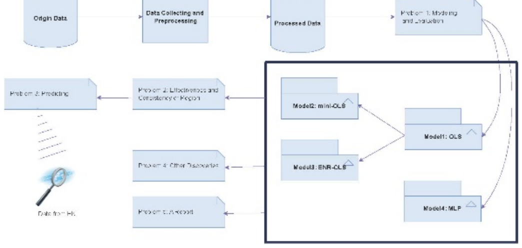
2.We use statistical methods, including ANOVA and t-tests of coefficients, to analyze the effect of region on pricing, and test its consistency.

3.We utilized macroeconomic data from Hong Kong to model its impact on regional macroeconomic factors. We selected a subset of thirty records and analyzed the predicted results in comparison to actual yacht prices in Hong Kong. Our analysis demonstrated the effectiveness of using geographic regions as a predictor, and we tested this approach on both catamarans and monohull sailboats.

4.We have made two noteworthy discoveries. One is that when simply using OLS analysis on original data, the error increases as the length does, indicating potential missing factors. The revised model included the square and cube of length as estimated specifications of sailboats, getting better outcomes. The other discovery is that, although the OLS model exhibits multicollinearity, we can utilize its data to make persuasive conclusions by employing a Bayesian-like technique, namely partial regression analysis.

5.Based on the work we did before, we write a report to help the broker understand Hong Kong used sailboat market.

Fig. 1. Flow Chart of Our Work


Origin Data
Data Collecting and Preprocessing
Processed Data
Problem 1: Modeling and Evaluation
Problem 3: Predicting
Problem 2: Effectiveness and Consistency of Region
Model2: mini-OLS
Model1: OLS
Model3: ENR-OLS
Problem 4: Other Discoveries
Model4: MLP
Data from HK
Problem 5: A Report

This paper is submitted on March 25th, 2024 for review.

1.4 Notation

Symbol	Explanation
LOA	Length over all in meters
S.A.	Total surface area of sails when fully raised
GDP	Gross Domestic Product of country
GDPPC	Gross Domestic Product Per Capita of country
marine	Marine protected area of country
OLS	Ordinary Least Squares
MLP	Multilayer Perceptron
VIF	Variance Inflation Factor

II. IMPLEMENTION

2.1 Data Preprocessing

2.1.1 Data Collecting

Surprisingly any model built directly using the data provided in "2023_MCM_Problem_Y_Boats.xlsx" does not exhibit a good fit. We think that this phenomenon is because the data is insufficient for the model to analyze the market well, and therefore caused failed practical analysis and prediction.

Based on this, we decide to collect more adequate data. After collecting dozens of yacht specifications, we choose the yacht's LOA, model, beam, draft, displacement, sail area, fuel tank, and water tank as additional parameters, along with the derived S.A./disp. and disp./len. To make the model more precise, we also collected the macroeconomic parameters of different countries in different years. Then chose GDPPC, import, export, and marine protected area as additional parameters to make 'country' a quantitative parameter rather than simply qualitative data. We will provide the sources of data used in the references. [1][2][4][6][7][10][11][12][13][14][15]

Fig. 2. Cloud Chart of Indicators

2.1.2 Dealing with Missing Data

As a relatively niche luxury market, the sailboat market does not provide a comprehensive and convenient database for boat specifications. Therefore, it is challenging to find rounded specifications of all given yachts. We choose to discard the corresponding records for the missing categorical data that does not have an apparent logical relationship. We use a semi-supervised learning approach to obtain estimated values as fill-ins for the missing quantitative data that have certain regularity.

2.2 Modeling and Evaluation

2.2.1 OLS: Ordinary Least Square

Considering the goals of given problems, selecting an interpretable model is crucial. Linear regression appears to be a suitable option, and therefore Ordinary Least Squares (OLS) was initially employed[8].

The formula for multiple linear regression can be expressed as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

where:

y is the listing price (the one we want to predict)

$x_1, x_2, \ldots, x_p$ are the predictors

β_0 is the intercept

$\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients

ϵ is the error term. The sample size is large enough, so it can be assumed that the ϵ follows a normal distribution $N(0, \sigma^2)$.

After running the model, we get some evaluation values:

Table 1. Evaluation of OLS

R^2	$\overline{R^2}$	F-statistic	σ	d
0.933	0.919	68.18	56434.4	1.65

The R^2 value of 0.933 represents 93.3% of variation in the listing price variable can be accounted for by the predictors included in the linear regression model.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y_i})^2}{\sum_i (y_i - \bar{y})^2}$$

where y_i denotes the observed value, $\hat{y_i}$ denotes the predicted value, $\bar{y}$ denotes the mean of the observed values.

The adjusted R^2 value is similar but considers the number of independent variables in the model, indicating that 91.9% of the variability in the listing price can be explained by the predictors after adjusting for overfitting.

$$\overline{R^2} = \frac{1-K}{N-K} + \frac{R^2(N-1)}{N-K}$$

The F-statistic is used to test the null hypothesis that all coefficients in the regression model are zero, suggesting that none of the predictors have a significant impact on the listing price. A high F-statistic value of 68.18 suggests that we can reject the null hypothesis, indicating that the model adequately fits the data and the predictors have a significant joint effect on predicting the listing price.

The Durbin-Watson d statistic serves as a means for mexamining the presence of autocorrelation within residual or error terms. Typically, a d value between 1.5 and 2.5 is deemed satisfactory, whereas a d value outside of this range suggests probable serious autocorrelation existing among error terms.

2.2.2 mini-OLS:Drop Correlated Predictors

Multicollinearity in regression occurs when two or more predictor variables are highly correlated with each other, making it difficult to determine the individual effect of each predictor. This can lead to unreliable and unstable estimates of the regression coefficients and can make it difficult to interpret the results of the regression analysis. It is important to detect and address multicollinearity before interpreting the results of a regression analysis.

The data we selected exhibited a multicollinearity issue, as indicated by the output of the code presented below:

```
Notes:
[1] ......
[2] The smallest eigenvalue is 2.68e-29. This might indicate
that there are strong multicollinearity problems or that the
design matrix is singular.
```

To avoid such problem, several factors need to be considered. Dummy variable trap is an usual way to cause multicollinearity. To avoid it we need to perform one-hot encoding on $(n-1)$ dummy variables when the category size is n.

One common way to solve multicollinearity is to eliminate predictors that are highly correlated with each other[5]. We applied the 'Variance Inflation Factor' technique to identify any interdependence among the predictors we selected, and obtained the results in a file named '$vif.csv$'. The analysis revealed that it would be difficult to simultaneously achieve optimal variable selection and maintain complete leverage of the available data, so we have to drop lots of features to obtain a stable and interpretable model.

After selecting predictors based on the 'Variance Inflation Factor' in a heuristic manner, we were left with the following 7 predictors.

Table 2. Model-2 Statistical Analysis to Predictor Coefficients

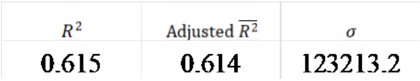
predictor	coef	std err	t	*P*>|*t*|	[0.025	0.975]
const	-3.55E+07	1.04E+06	-34.294	0	-3.76E+07	-3.35E+07
Year	1.73E+04	516.176	33.515	0	1.63E+04	1.83E+04
Length(ft)	1.39E+04	504.605	27.601	0	1.29E+04	1.49E+04
Beam	4.53E+04	2084.809	21.71	0	4.12E+04	4.93E+04
Fuel	142.9576	12.399	11.53	0	118.648	167.267
S.A./Disp.	4.59E+06	1.06E+06	4.335	0	2.51E+06	6.66E+06
gdppc	3.2075	0.214	14.989	0	2.788	3.627
Is_Europe	-2.02E+04	6925.268	-2.914	0.004	-3.38E+04	-6603.782
Is_USA	9.83E+04	8165.696	12.033	0	8.22E+04	1.14E+05

The results demonstrate that our chosen predictors provide consistent coefficients, enabling us to effectively interpret their impact. This is evidenced by the p-values obtained from performing hypothesis $H_0: \beta_j = 0$ for each predictor, which are almost all close to zero. Additionally, their coefficients have confidence intervals that fall within a narrow range of epsilon. These findings suggest that Model 2 is robust and possesses a strong explanatory power.

The observation also aligns with empirical evidence that the listing price exhibits a positive correlation with building materials, their efficacy, GDP per capita, and years they are built. The coefficients of these variables signify the degree to which the listing price could potentially rise with unit increase of the predictor, assuming linearity.

The evaluation of this model is shown below:

Table 3. Evaluation of mini-OLS



R^2	Adjusted $\overline{R^2}$	σ
0.615	**0.614**	**123213.2**

As can be inferred, the adjusted $\overline{R^2}$ value shows a significant decrease to 0.614 while σ doubles in magnitude. This indicates that only 61.4% of the variation in listing price can be accounted for by this model, and the substantial increase in σ suggests that the predictive accuracy is limited.

2.2.3 ENR-OLS: Elastic Net Regularization

An alternative strategy to mitigate multicollinearity is to use regularization[3]. In this study, we specifically opted for Elastic net regularization, which combines the strengths of ridge and lasso regularizations.

$$Loss = \frac{SSR}{2n} + \alpha\left(\frac{1-w}{2}|params|_2^2 + w|params|_1\right)$$

where SSR refers to the Sum of Square residuals. The symbol α represents a hyperparameter used for regularization, while w denotes the weight assigned to the L_1 component.

We experimented with various combinations of hyperparameters and observed that upon applying regularization, the coefficients of predictors exhibited higher stability than those in Model 1. However, the interpretations derived from the model did not align with real-world observations. For instance, contrary to expectations, an increase in the parameters of year, beam, water, and draft led to a decrease in the listing price. This discrepancy appeared consistently across numerous hyperparameter adjustments, indicating that regularization led to a biased estimation of the data.

Table 4. Coefficients of predictors in Model 3

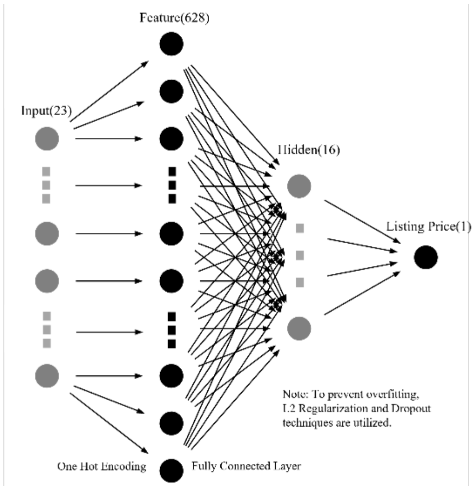
predictor	coefficients
const	3.036208e+05
Length (ft)	1.841988e+02
Year	-3.291069e+00
Monohulled	-8.066942e+04
Catamarans	1.601292e+05
LOA	1.761799e+02
Beam	-5.413574e+01
S.A.	1.081311e+02

The evaluation of this model is: σ = 66411.60

2.2.4 MLP: MLP prediction

To improve the accuracy of our prediction, we have integrated a Multi-layer Perception (MLP) nonlinear model. The MLP approach provides increased adaptability in capturing intricate relationships among variables by incorporating multiple layers of neurons and nonlinear activation function within the network architecture.[9] The purpose of this model is to improve the prediction's performance and accuracy, without prioritizing explanatory power.

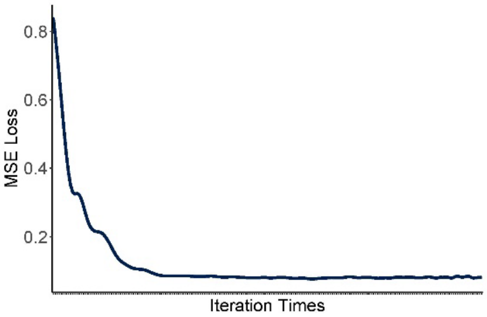
Fig. 3. Multilayer Perceptron (MLP) Architecture


Feature(628)
Input(23)
Hidden(16)
Listing Price(1)
Note: To prevent overfitting, L2 Regularization and Dropout techniques are utilized.
One Hot Encoding
Fully Connected Layer

In order to capture the hidden characteristics of the problem, it may be necessary to incorporate additional neural network layers beyond our linear model. Adding a single hidden layer to the network architecture can facilitate the identification and modeling of these features.

In order to align the complexity of the model with the size of the data set, the hidden layer was configured to have 16 neurons with 0.3 dropout rate. Meanwhile, L_2 regularization was employed.

Fig. 4. MSE loss of MLP with epoch iterations (200 Iters)



We can see from the figure that the MSE value dropped drastically within the initial 50 iterations, post which it remained approximately constant at 0.11. The σ value for this model is 51016.35, which is the smallest among all our models.

2.3 Effect of Region

2.3.1 Coefficient Testing in Model 1

In order to dig out the effect of region on the listing price, we calculated the p-values for the following hypotheses in Model 1:

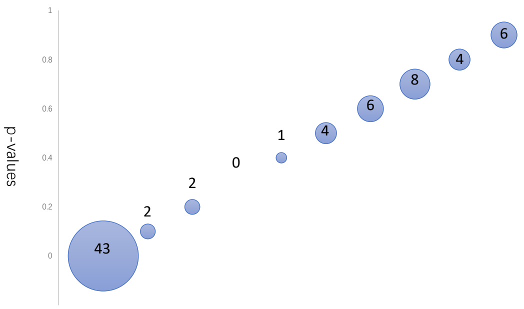
$$H_0: \beta_j = 0$$

where β_j refers to the j-th coefficient of the dummy variables for the predictor variable $Region$.

The p value generated by t-statistic is a measure to show the effectiveness of the predictors. In particular, the p value of 0.1 associated with a predictor indicates that there exists a 10.0% likelihood that this predictor has no significant impact on the outcome variable, listing price.

To show the effectiveness of $Region$, we need to synthesis the p value of all dummy variables generated by its different categories. Here we show the distribution of p values.

Fig. 5. Bubble chart of P-values for Region Dummies' Coefficients



It indicates that a relationship exists between region and listing price, as most of the p-values are concentrated around 0.

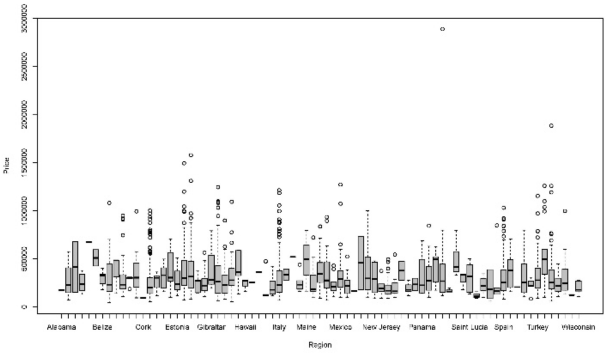
2.3.2 Evidence from Model 2

In model 2, the region related predictors we reserved is Geographic Region. To prevent the dummy variable trap, we selected Caribbean as the base reference. Therefore, the coefficients for Is_USA and Is_Europe demonstrate the contrast of these regions with respect to the Caribbean region as 2.2.2 shows. The data indicates a conspicuous disparity between the listing prices in the United States as opposed to those in the Caribbean and Europe, with the former being considerably higher.

2.4 Consistent Effect of Region

To determine if the regional effect is consistent across all sailboat variants intuitively, we draw the Box-plot grouped by Region.

Fig. 6. Box-plot of listing price grouped by region



ANOVA is a statistical method often used to determine whether there are statistically significant differences between groups in terms of their mean values.

To discuss whether regional effect is consistent across all sailboat variants. We consider each region as a group and use ANOVA to test whether it is consistent. Consequently, we have converted the problem into a Single-factor K-level experiment wherein the "single factor" pertains to regional factors and "k-level" refers to the number of various regions.

$$Y_{ij} = a_i + e_{ij}, j = 1, \cdots, n_i, i = 1, \cdots, k$$

where Y_{ij} refer to the j-th listing price of region i, a_i means the mean value of region i , e_{ij} is random error, we assume:

$$E\left(e_{ij}\right) = 0, 0 < Var\left(e_{ij}\right) = \sigma^2 < \infty$$

and all e_{ij} are $i.i.d.$

To test the consistency, we hold the null hypnosis:

$$H_0: a_1 = a_2 = \cdots = a_k$$

Divide the sum of squares of residuals into within-group SS_e and among-group SS_G components.

$$SS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2, \bar{Y} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}/n$$

$$SS_e = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2, \bar{Y}_i = \left(Y_{i1} + \cdots + Y_{in_i}\right)/n_i, i = 1, \cdots, k$$

$$SS_G = SS - SS_e = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2$$

where n_i represents the quantity of data in a given group denoted by i, SS_e corresponds to the residual variation within groups, which is considered as the same random error for all groups, SS_G refers to the residual variation among all groups, which is considered as their differences.

Assuming that the null hypothesis H_0 is true, and given a sufficient number of samples, we can make the assumption that the error term e_{ij} follows a normal distribution with mean 0 and variance σ^2. This leads us to the following conclusion:

$$MS_G = SS_G/(k-1), MS_e = SS_e/(n-k)$$
$$MS_G/MS_e \sim F_{k-1,n-k}$$

Now we can use F-statistics to test the consistency of the effect of regional predictor. Specifically, we will accept the null hypothesis H_0 if the subsequent inequality is satisfied:

$$\frac{MS_G}{MS_e} \leq F_{k-1,n-k}(\alpha)$$

Table 5. ANOVA of Region

	Sum_Square	DoF	F	PR(>F)
C(Region)	8.95E+12	7.60E+01	3.14E+00	8.06E-18
Residual	1.28E+14	3.41E+03	-	-

The p-value result for C(Region) indicates that there is sufficient evidence to support accepting the null hypothesis, which suggests that the regional effect is consistent across all sailboat variants.

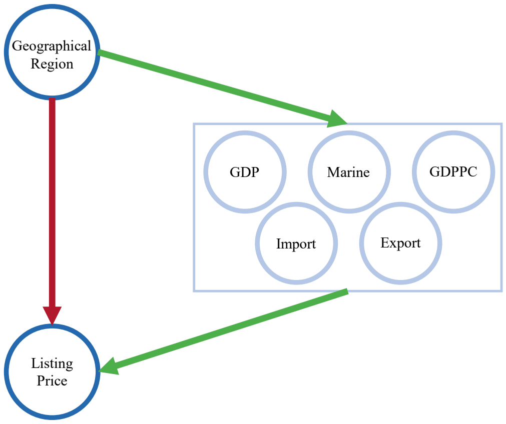
III. PREDICTING AND ANALYTICS

3.1 Boat Subsets and Market Data

We collected the Hong Kong market data from web[15]. Two subsets of 15 records each were chosen, one for monohulled sailboats and the other for catamarans. The selection was made with consideration for the balance of data size between different categories due to the limited availability of sailboat pricing data in Hong Kong.

In order to assess the potential effect of our geographic region modeling approach in the Hong Kong (SAR) market, we assume that the primary determinant of listing price variability across geographic regions is macroeconomic data. Specifically, we assume that the causal relationship indicated by the red pathway in the following diagram may be substituted with that of the green pathway:

Fig. 7. Two pathway assumed equal



Specifically, our assumption is that the macroeconomic data has the ability to capture and represent the distinctions in geographical regions at a relatively elevated level. Therefore, we have gathered macroeconomic data for Hong Kong from 2005 to 2019.

3.2 Final Results

Due to the limited availability of data from Hong Kong, we have opted to utilize the geometric mean absolute residual metric in order to present the predictive performance of various models across different datasets. This approach allows for a stable overall evaluation while minimizing the impact of excessive noise.

$$GMAR = \left(\prod_{i=1}^{n}\left|y_i - \hat{y}_i\right|\right)^{\frac{1}{n}}$$

Fig. 8. HK Monohulled Price Prediction with 4 Models

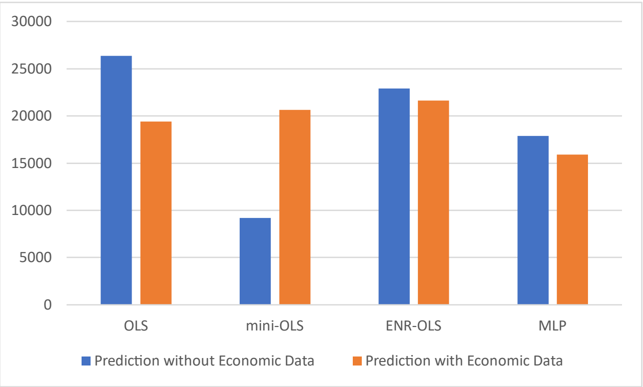
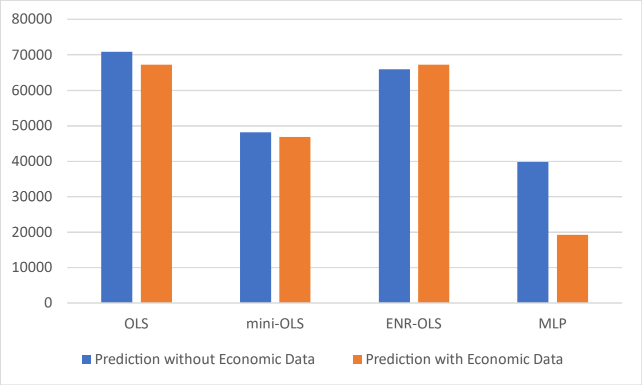


Fig. 9. HK Catamaran Price Prediction with 4 Models



The data indicates that in most cases, precision improves with the availability of macroeconomic data, thereby demonstrating that geographic region information can be useful in the Hong Kong market, so our hypothesis that the impact of a geographic region can be substituted by examining macroeconomic data at a level commensurate with that particular region is supported. The enhancement of precision is evident in both catamarans and monohull sailboats.

It can also be observed that when economic data is included, MLP outperforms other linear models significantly. This indicates that machine learning methods have stronger predictive capabilities compared to linear models. Although MLP performs well even without economic data, mini-OLS significantly improved. This could be due to the small sample size of the selected data in HK market.

Among three linear models, OLS and ENR-OLS perform poorly while mini-OLS performs well. This is because although OLS has the highest R^2, this only indicates that the selected features are good predictors of listing price. However, due to the presence of multicollinearity, OLS is highly unstable. The regularized model can improve coefficient stability but introduces bias that results in similar prediction performance to OLS. By using the mini-OLS model, which eliminates correlated variables, we achieve the best stability, accuracy, and interpretability among the linear models.

IV. CONCLUSION

This research set out to develop a succinct pricing model for sailboats and to gain a comprehensive understanding of the factors that affect the pricing of used boats.

We used several statistical methods including but not limited to VIF, ANOVA and t-tests of coefficients to select independent variables as predictors, to test the effect and to valid the consistency.

The regression analysis revealed how the price is potentially related to the given predictors. We also introduced a neural network model solely for predictive purposes and disregarded its deficiencies for explanation.

We also utilized macroeconomic data from Hong Kong to model its impact on regional macroeconomic factors. We selected a subset of thirty records and analyzed the predicted results compared to actual yacht prices in Hong Kong. Our analysis demonstrated the effectiveness of using geographic regions as a predictor, and we tested this approach on both catamarans and monohulled sailboats and got a consistent result.

However, the study is not flawless. The major limitation of this study is the lack of variety of the data set. The given data has limited dimensions and less than 5 thousand samples. Due to limited time, we chose to expand the width of the data instead of the size because it is relatively enough for linear regression. Since we have got a usable result, a larger set of data is bound to make the model more convincing.

Apart from our main goals, we have also made two noteworthy discoveries. One is that when simply using OLS analysis on original data, the error increases as the length does, indicating potential missing factors. The revised model included the square and cube of length as estimated specifications of sailboats, getting better outcomes. The other discovery is that, although the OLS model exhibits multicollinearity, we can utilize its data to make persuasive conclusions by employing a Bayesian-like technique, specifically, partial regression analysis.

The insights gained from this study may be of assistance to brokers all around the world in deciding the boat price. Considering the economic significance, this would be a fruitful area for further work.

REFERENCES

[1] ayc-yachtbroker. 2023. url: https://www.ayc-yachtbroker.com/ .

[2] b-yachts. 2023. url: https://www.b-yachts.com/ .

[3] G. Casella and R.L. Berger. Statistical Inference. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002. isbn: 9780534243128. url: https://books.google.com.hk/books?id=0x%5C_vAAAAMAAJ.

[4] edwardsyachtsales. 2023. url: https://www.edwardsyachtsales.com/ .

[5] R.V. Hogg, J.W. McKean, and A.T. Craig. Introduction to Mathematical Statistics. What's New in Statistics Series. Pearson, 2019. isbn: 9780134686998. url: https://books.google.com.hk/books?id=V1SzswEACAAJ.

[6] itboat. 2023. url: https://itboat.com/models/.

[7] katamarans. 2023. url: https://www.katamarans.com/.

[8] D.C. Montgomery, E.A. Peck, and G.G. Vining. Introduction to Linear Regression Analysis. Wiley Series in Probability and Statistics. Wiley, 2013. isbn: 9781118627365. url: https://books.google.com.hk/books?id=lSyiRZh09oEC.

[9] J. Quinn et al. Dive Into Deep Learning: Tools for Engagement. SAGE Publications, 2019. isbn:9781544385402. url: https://books.google.com/books?id=eaCgDwAAQBAJ.

[10] vismara-mc. 2023. url: https://www.vismara-mc.com/ .

[11] whitsundayescape. 2023. url: https://www.whitsundayescape.com/ .

[12] worldbank. 2023. url: https://www.worldbank.org/ .

[13] worldwideyachtbroker. 2023. url: https://www.worldwideyachtbroker.com/ .

[14] yachtvillage. 2023. url: https://www.yachtvillage.net/ .

[15] yachtworld. 2023. url: www.yachtworld.com/ .

APPENDICES

mini-OLS code

```python
import pandas as pd
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
from statsmodels.api import OLS
from statsmodels.graphics.regressionplots import plot_regress_exog
```

Listing 1: Import Envrionments

```python
def VIF(X):
    vif = pd.DataFrame()
    vif['variables'] = X.columns
    vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
    vif.tocsv("tmp.csv")
```

Listing 2: Variance Inflation Factor to make OLS ”mini”

```python
Data = pd.read_excel("2023_MCM_Problem_Y_Boats_selected.xlsx", index_col=None, sheet_name="Boats")
Data.dropna()
X = Data.loc[:, ['Year', 'Length (ft)', "Beam", "Fuel", "S.A./Disp.", 'gdppc','Geographic_Region']]
X = pd.get_dummies(data=X, columns=['Geographic_Region'], drop_first=True)
# X = pd.get_dummies(data=X, columns=['Make', 'Variant', 'Geographic Region', 'Region'], drop_first=True)
y = Data.loc[:, ['Price']]
y = y.values.reshape(y.shape[0])
X = sm.add_constant(X)
```

Listing 3: Data Processing

```python
model = OLS(y, X).fit()
print(model.summary())
y_hat = model.predict(X)
residuals = y - y_hat
print(np.sqrt(np.sum(residuals**2)/DoF))
```

Listing 4: OLS Linear Regression

MLP code

```python
import torch
device=torch.device('cuda' if torch.cuda.is_available() else 'cpu')
from torch import nn
from sklearn.utils import shuffle
import pandas as pd
```

Listing 5: Import Envrionments

```
def zscore(df):
    means=[]
    stds=[]
    for key in df.keys():
        if key!=’Make’ and key!=’Variant’ and key!=’Geographic_Region’ and key!=’Region’:
            mean=df[key].mean()
            std=df[key].std()
            means.append(mean)
            stds.append(std)
            df[key]=(df[key]-mean)/std
    return df, means,stds

Data=pd.read_excel("2023_MCM_Problem_Y_Boats.xlsx",index_col=None,sheet_name="Boats")
MonohulledDataCopy=Data
Data.dropna()
Data=shuffle(Data,random_state=2331567)
X=Data.loc[:,[’Make’,’Variant’,’Length (ft)’,’Year’,"Monohulled","Catamarans","S.A.","Draft","Displacement","Fuel","Water","S.A./Disp.","Disp./Len.","export","import","gdp","gdppc","marine"]]
X,Xmeans,Xstds=zscore(X)
X=pd.get_dummies(data=X,columns=[’Make’,’Variant’])
Y=Data.loc[:,[’Price’]]
Y,Ymeans,Ystds=zscore(Y)

x_train=torch.tensor(X[:n_train].values, dtype=torch.float32,device=device)
x_test=torch.tensor(X[n_train:X.shape[0]].values, dtype=torch.float32,device=device)
y_train=torch.tensor(Y[:n_train].values.reshape(-1, 1), dtype=torch.float32,device=device)
y_test=torch.tensor(Y[n_train:X.shape[0]].values.reshape(-1, 1), dtype=torch.float32,device=device)
```

Listing 6: Data Processing

```
def init_weights(m):
    if type(m) == nn.Linear:
        nn.init.normal_(m.weight,std=0.056)

net=nn.Sequential(nn.Flatten(),nn.Linear(548, 16),nn.ReLU(),nn.Dropout(0.3),nn.Linear(16, 1))
net.apply(init_weights);
net=net.to(device=device)
```

Listing 7: Build Nerual Network

```
batch_size, lr, num_epochs=256, 0.01, 200
loss_p=0
loss_fn=nn.MSELoss(reduction="mean")
optimizer=torch.optim.Adam(net.parameters(), lr=lr, weight_decay=0.001,betas=(0.9,0.999),eps=1e-8)
for i in range(num_epochs):
    net.train()
    y_pred=net(x_train)
    loss=loss_fn(y_pred,y_train)
    with torch.no_grad():
        net.eval()
        y_pred=net(x_test)
        loss1=loss_fn(y_pred,y_test)

    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
```

Listing 8: Train Network

```
y_pred=net(x_test)
y_pred=y_pred*Ystds[0]+Ymeans[0]
y_test=y_test*Ystds[0]+Ymeans[0]
for i in range(len(y_test)):
    error += torch.pow((y_test[i].item()-y_pred[i].item()),torch.tensor(2,dtype=torch.float32,device=device))
error=error / len(y_test)
error=torch.sqrt(error)
print(error.item())
```

Listing 9: Evaluate Network

Regression Analysis code

```
def anova(Data):
    model_anova = smf.ols("Price ~ C(Geographic_Region)",data=Data).fit()
    anova = sm.stats.anova_lm(model_anova, test="F", typ=2)
    print(anova)
```

Listing 10: Analysis of Variance

```
def partial_reg(model):
    plot_regress_exog(model, 'Year')
    plt.show()
```

Listing 11: Partial Regression Analysis