Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2332034

Sailboat Listing Price Prediction: XGBoost Based on Multivariate Linear Regression

Summary

In today's increasingly complex market environment, accurate price prediction has become a critical aspect that is difficult to obtain, particularly for luxury goods such as sailboats. At the request of a Hong Kong broker, we have analyzed the provided data and established a mathematical model to estimate the listing price of used sailboats based on our criteria.

Our approach was inspired by the Hedonic Price Method, which posits that goods are valued based on their utility-bearing attributes. We chose to use the methods applied to house price prediction as a reference since the real estate and sailboat markets share strong similarities. As a result, we modeled the used sailboat listing price prediction as a machine learning regression problem and conducted extensive research to understand the algorithms that would be most effective. We found that clustering methods could enrich data features, linear regression could generate explainable results, and XGBoost was a powerful algorithm in this realm.

We decided to use a combination of K-means clustering, multivariate linear regression, and XGBoost algorithm, as the boosting approach suggests that weak models can be integrated into a strong one. We used K-means clustering to process the original data to discover hidden features and included these features in the dataset before sending it to the linear and XGBoost models. These models generated explanations for the listing price and fit the augmented dataset.

To ensure accurate predictions, we gathered sailboat specification information and regional economic traits to generalize the predictions for the Hong Kong market. We developed two crawlers to extend the training samples in both the row and column dimensions from sailboat websites and manually collected data from official websites to obtain authentic values.

After thorough research and experimentation, we have successfully constructed our prediction model using XGBoost regression as the primary algorithm. We conducted multidimensional testing on the model's results and utilized it to observe the data, uncovering intriguing conclusions, which are presented later in this report. Our findings have enabled us to offer a comprehensive and concise report to the Hong Kong broker, with insightful suggestions for his business consideration. Furthermore, we proposed a design for a price prediction system based on our methodology.

Keywords: Price Prediction; Hedonic Price Method; Multivariate Linear Regression; XGBoost Regression; K-means Clustering

Contents

1 Introduction

1.1 Background

Sailboat is a kind of ancient water transportation tool, which has a history of more than 5000 years. Sailboats are powered by sails using the force of the wind. They are also referred to as sailing dinghies, boats, and yachts, depending on their size.[1]Sailing is also one of the hottest water sports, in which athletes drive a sailboat and compete for speed in a defined course. Basing on the hull type(the number of hull), sailboats can be classified into monohull, catamaran, trimaran. Figure1 offers an example of monohull and catamaran. To the best of our knowledge, monohulls and catamarans almost occupy the whole sailboats market, so only two classes are considered in this work.

Considering the purchase prices, sailboats are surely one of the luxury goods, and the used boat trading constitutes a large market proportion. However, the value of a sailboat varies greatly because of internal and external factors, such as its physical specifications and the characteristics of the selling area. A reasonable price can guarantee sales, energize the market and lead to mutual benefits among buyers and sellers. In this paper, after expanding the original data in both sample and feature numbers, we model the prediction problem as machine learning regression problems, provide the results of our models basing on the demands, and finally offer a report for the Hong Kong (SAR) sailboat broker.

(a) Monohull Sailboat (b) Catamarans

Figure 1: Examples of sailboat

1.2 Problem Analysis

At the request of the broker, we have 4 missions to conquer. Problem 1 requires us to develop a mathematical model to explain the prices of each sailboat. Problem 2 requires us to explain the region's effect on the listing prices and we will discuss the practical and statistical significance of it. Problem 3 requires us to forecast the prices in used sailboat market of Hong Kong, then we need to compare the regional effect of Hong Kong on

each of the sailboat prices for the sailboats in our selected subset. Problem 4 requires us to summarize our work and provide a formal report.

Our idea came from the hedonic hypothesis that goods are valued for their utility-bearing attributes or characteristics.[2]Hedonic pricing illustrates a principle that the willingness and perception of a customer add or detract the value of an asset. The real estate market can be a typical example, where the prices are determined by geographical properties(crime rate, the level of water and air pollution, etc.) and the structural designs(size, appearance, etc.). From where we stand, there exits strong similarity between sailboats and houses, so it's natural to consider using the methods in house price prediction in sailboats scenario. The provided excel data can be treated as tied packages of characteristics, when treating listing prices as label y, other properties are $x_1, x_2, \cdots x_m$, the modeling process is to find a function f that makes the result of $|f(x_1, x_2, \cdots x_m) - y|$ as small as possible.

We notice that the broker shares strong interest in the relation of price and region property, yet the provided excel file has so few data that not a single ship in Hong Kong is included. Intuitively, the data is not adequate to complete the mission. So, we need to extend our data in both samples and features, which will be detailed in section2. After enriching data size, we will separate the original data into subsets, each of them inherit some of the original features. With cluster algorithm, we can generate a corresponding new features respectively. We name this process as stage 1, figure2 is aforementioned idea visualized.

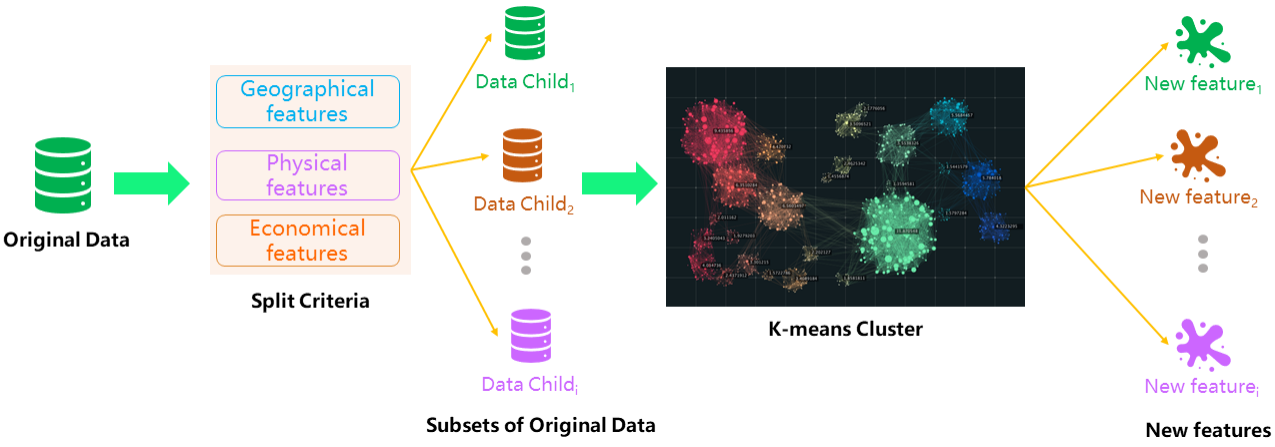


Figure 2: Data Augmentation by Cluster Algorithm

After data augmentation, we will select from new features and original features to forge the final dataset and feed them to our selected algorithms. Specifically speaking, multivariate linear regression[3] hold good reputation for its simplicity and effectiveness, we are confident it will perform well with our extended data. In the meantime, XGBoost[4] enjoys great fame of powerful prominence as well as adequate explanation, which suits our needs. Follow the concept of boosting, we train them both to evaluate their performance and try to find a combination to form our price predictor. Figure3 will be helpful to understand our method.

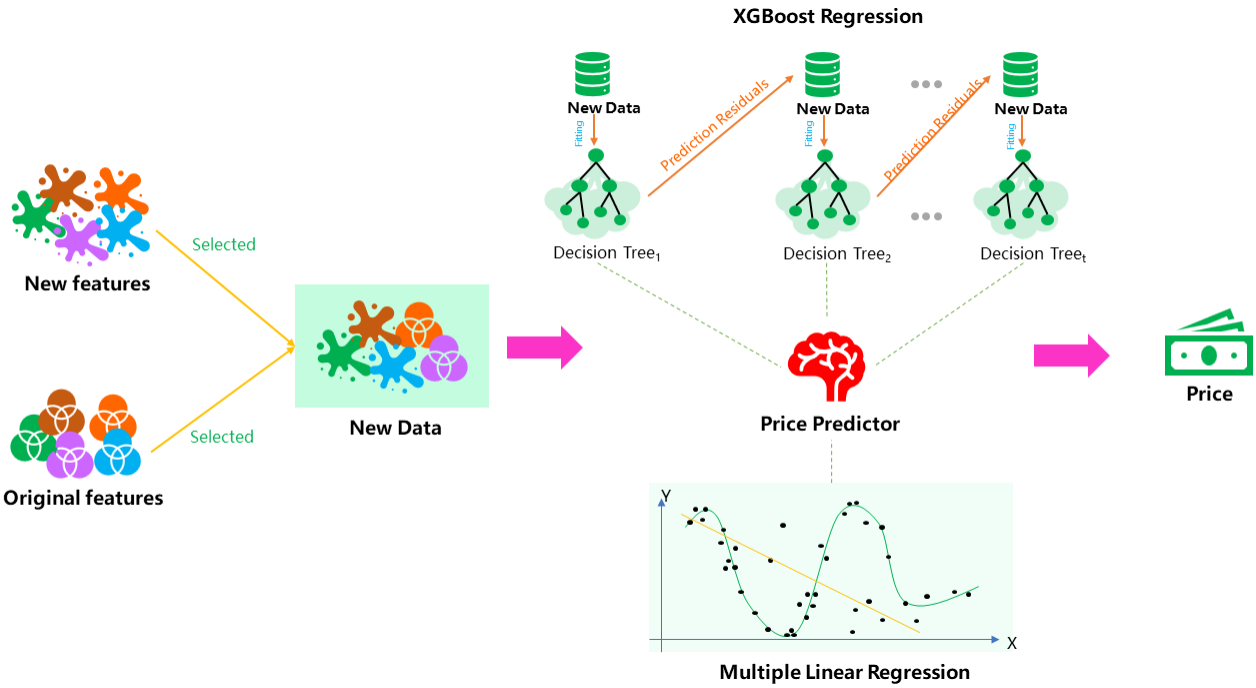


Figure 3: Train price predictor with newly-forged data and regression algorithms

2 Data Processing

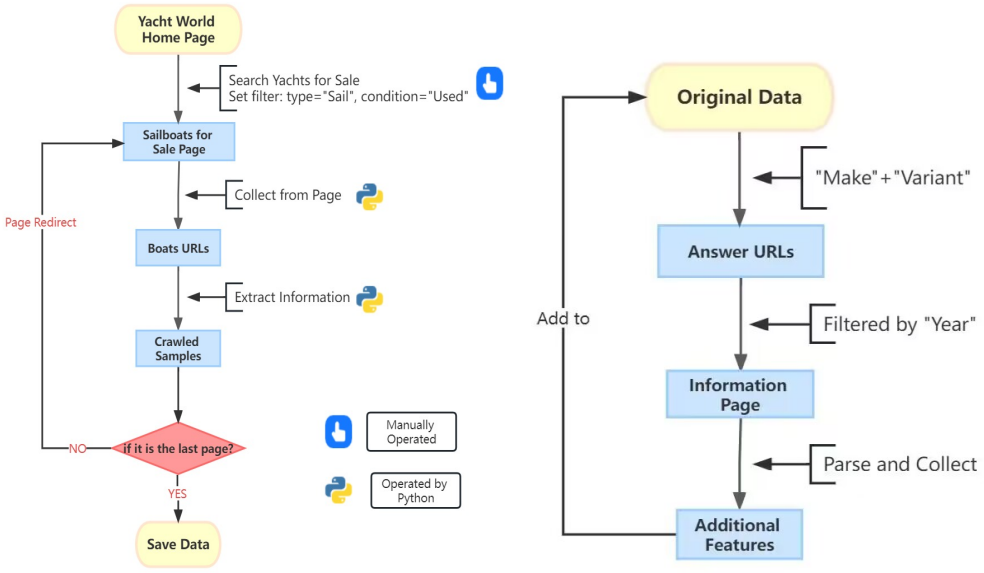
The realm of statistics and machine learning demands data of quality. That's because no matter how subtle the algorithm is, it can't be proven practical without being tested on real data. Lack of qualified data will unavoidably lead to the potential of a design not being fully demonstrated, or result in generating problematic conclusions. Thus, in order to offer a more reliable report, we feel it necessary to extend our dataset in both sample size and data features.

2.1 Data Crawling

One of the best things of Internet era is that you can almost find any information you want from it. We found YachtWorld[5] was extremely helpful in our mission. It is a boat online marketing platform with the largest database of brokerage boats, where we can absolutely obtain data about sailboats. However, since the web page returned us over 10,000 sailboats sales forms, it would be time-consuming for us to collect them all during the context period. That's why we resorted to the web crawler technique to do the favor. A web crawler is an Internet bot that can systematically browse the websites, typically for the purpose of Web indexing[6]. We introduce it to assist in data collection. We built our web crawler in Python, using package Requests to request the website and package BeautifulSoup to do the information parsing and extraction. Our implementation follows the logic in Figure4 (a).

However, the YachtWorld's abundance has no structural patterns, apart form the basic information, extra contents varies greatly, making it impossible to generate a structural table. To compensate for this loss, we resorted to SAILBOATDATA.COM, a database that

contains information on over 8800 sailboats going back as far as 1900. We developed a new crawler that can extract from original excel to generate a search key word, which can be used to request SAILBOATDATA.COM information about the particular ship's physical characteristics like beam, draft, displacement, hull materials, etc. The implementation logic is shown in Figure4(b).



(a) Crawl YatchWorld.com (b) Crawl SailBoatData.com

Figure 4: Implementation logic of Web Crawler

Eventually, we have fetched over 8,000 samples from YachtWorld.com and over 1,300 ship properties from BOAT.com. However, these two data are too complicated and disordered, compared with the given excel. We keep them as alternative if time is available, since it's too time consuming to process. Apart from the physical characteristics, we gathered information about GDP, unemployment rate, inflation rate, etc. to depict the features of given country/state/region. This is because, our methods need train predictors, which means the models learn everything from our data. Unfortunately, we find there is no sample related to Hong Kong, the crawled results have less than 10 to offer. Therefore, without traits describing the regions, our model can hardly generalize on the Hong Kong market.

2.2 Data Cleaning

Before sending data to a machine learning algorithm, it is important to ensure that the data is cleaned and preprocessed to improve the accuracy and effectiveness of the model. In the case of the given excel with columns labeled Make, Variant, Length (in feet), Geographic Region, Country/Region/State, Listing Price, Year, etc. There are 3 missing values for country/region/state in the data, but since they are very few in number and are unlikely to have a significant impact on the overall model, we have chosen to delete them. The data passed the 3-sigma test and there were no outliers. Scaling and normalizing are

also included to ensure all variables are on a comparable scale. After data is prepared, we deem that dividing features according to Figure5 could be inspiring and realistic.

(a) Sailboats Physical Traits (b) Position Traits

Figure 5: Classify the features into two classes

3 Model Construction

3.1 K-Means Clustering

3.1.1 Clustering and K-Means

Clustering algorithm is a renowned type of self-supervised learning. Unlike classification, clustering is the process of dividing samples into classes by the intrinsic relationships within data given no labels in advance. The final aim is to have the classes maintaining strong similarity within, and sharing little in common extra-categorically. And K-means[7] is the most representative clustering method.

3.1.2 Methodology and Process

Simply speaking, K-Means seeks to a partitioning scheme of K clusters, so that the loss function corresponding to the result is minimized. Normally, the sum of the mean squared errors of the distance of each sample is used as loss function:

$$J(c,\mu) = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \mu_{c_i}\|^2 \tag{1}$$

where x_i is the i-th sample containing features in the form of an array, c_i is the cluster x_i belongs to, $\mu_{c_i}$is the center point of clusterc_i,N is the size of dataset. We implement this method on all the subset divided by columns from the original one.

The K is a manually-set hyper-parameter, indicating how many clusters we want the algorithm to find. And the center point for each cluster is required. The standard process goes by:

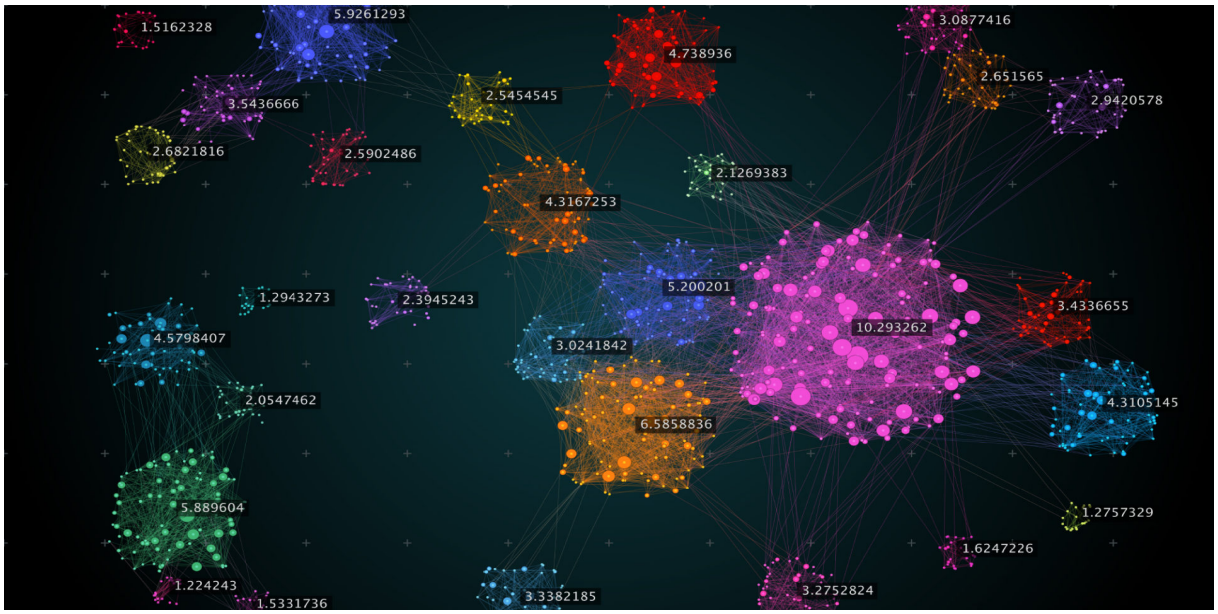


Figure 6: Example of Clustering

1) Data pre-processing. Mainly operate standardization and Anomaly Filtering.

2) Randomly pick K samples as the initialized center points. Let's note them as $\mu_1, \mu_2, \cdots, \mu_K$.

3) Define the loss function. Here, we follow the classical mean squared loss $J(c, \mu) = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \mu_{c_i}\|^2$

4) Set iteration times $t = 0, 1, 2, \cdots$. Repeat the following steps until the function converges or reach a set number:

   (a) For each x_i, it is compared with all center points $\mu_1, \mu_2, \cdots, \mu_K$, and will be arranged the same class with the nearest center:

   $$c_i = \arg\min_K \|x_i - \mu_k^t\|^2$$

   where c_i is the class of samplex_i.

   (b) At the end of each iteration, the center points of K classes will be updated according to:

   $$\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_j$$

   where N_i is the number of samples belong to cluster$i$.

The key operation here is to adjust the class of each sample to reduce the calculation of the loss function J and update the center points by iteration. The alternation of these 2 steps simultaneously minimize the J as well as stabilize the classification.

3.2 Multivariate Linear Regression

3.2.1 Linear Regression

Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables[8]. Given a dataset with N samples:$D = (x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$, where $x_i = (x_{i1}, x_{i2}, \cdots, x_{id})$ has d features, label $y_i \in \mathbb{R}$. Linear regression serves as a model that can take input x_i to forecast y_i as accurate as possible. Assuming that a sample has only one feature, the subscript of it can be ignored, giving us $D = \{x_i, y_i\}_{i=1}^N$, $x_i \in \mathbb{R}$. The linear regression aims to find a function f that:

$$\begin{aligned} f(x_i) &= wx_i + b \\ \text{s.t.} f(x_i) &\simeq y_i. \end{aligned} \tag{2}$$

So the key is about evaluating how close $f(x_i)$ is to the y_i, a popular criteria is the Mean Squared Error:

$$MSE = \frac{1}{N}\sum_{j=1}^{d}(f(x_j) - y_j)^2 \tag{3}$$

where j denotes the j-th feature. When in this single-feature case, the learning process is actually finding a w^* and b^* that:

$$\begin{aligned} (w^*, b^*) &= \arg\min_{(w,b)} \sum_{i=1}^{N}(f(x_i) - y_i)^2 \\ &= \arg\min_{(w,b)} \sum_{i=1}^{N}(wx_i + b - y_i)^2 \end{aligned} \tag{4}$$

3.2.2 Multivariate Linear Regression

This[3] corresponds to the common situation where each sample has d different features. Thus $x_i, w \in \mathbb{R}^d$, but the optimization form and loss function are the same as former section. Computationally speaking, the multi-dimension features can be used to express dataset D as a $N \times (d+1)$ matrix X, each line represents a sample, the first d columns correspond to d characteristics, the last column is all set by 1:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & a_{1d} & 1 \\ x_{21} & x_{22} & \cdots & a_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & a_{Nd} & 1 \end{pmatrix} = \begin{bmatrix} \boldsymbol{x_1^T} & 1 \\ \boldsymbol{x_2^T} & 1 \\ \vdots & \vdots \\ \boldsymbol{x_N^T} & 1 \end{bmatrix}$$

where $\boldsymbol{x_i} = (x_{i1}; x_{i2}; \cdots; x_{id}$. Simultaneously, we can write $\boldsymbol{y} = (y_1; y_2; \cdots; y_N)$, then we have equation similarly to equation(4):

$$\hat{\boldsymbol{w}}^* = \arg\min_{\hat{\boldsymbol{w}}} (X\hat{\boldsymbol{w}} - \boldsymbol{y})^T (X\hat{\boldsymbol{w}} - \boldsymbol{y}) \tag{5}$$

3.3 XGBoost Regression

3.3.1 What's Boosting?

When given a task or problem, whether it can be accomplished or how well it can be handled depending on the difficulty of the problem and the person to do the job. In reality, we can always find a more complex issue, but hardly can we have the suited candidate. However, with combined efforts and abilities, we can always conquered obstacles and even things called impossible. The phenomenon stands true for other spices in the nature7.

(a) Human working together (b) Ant working together

Figure 7: Working together to accomplish difficult tasks

This is the core idea of ensemble learning: uniting weak individual learners to form a powerful model. Basing on the connection between the individuals, those with weak relationships are called Bagging, and those with strong relationships are named as Boosting. And XGBoost is a boosted tree model with tree models(CART regression) as its individuals.

3.3.2 CART Tree

CART[9] is the abbreviation of "Classification and Regression Tree", a popular decision tree algorithm. It starts with setting all samples in the root node, and basing on certain standards to classify the samples by feature. As you perceive, the criteria used for attribute division hold place of significance. As the division goes, what we want to see is that each branch node containing samples of better purity (as similar as possible).

3.3.3 CART for Classification

CART uses Gini index[9] as the classification dividing criteria: Assuming we have K classes, D is our training data, C_k denotes the samples of class k, p_k is the proportion of

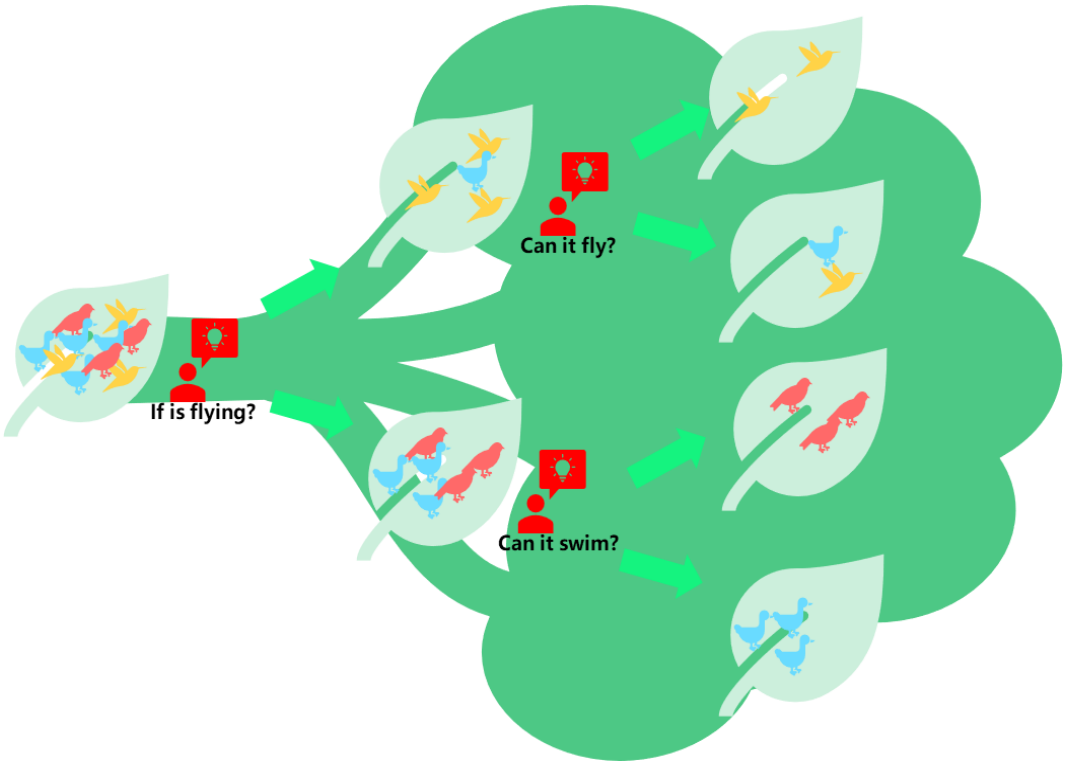


Figure 8: Example of Decision Tree

the k-th class. Then Gini index can be calculated by:

$$\begin{aligned} Gini(p) &= \sum_{i=k}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2 \\ Gini(D) &= 1 - \sum_{k=1}^{K} (\frac{|C_k|}{|D|})^2 \end{aligned} \tag{6}$$

where, $|T|$ returns the total number of set T. For each division, we need to calculate the difference between the Gini coefficient after segmentation according to a feature and the Gini coefficient before segmentation. For two-class classification, according to feature a, the Gini point after division is:

$$\begin{aligned} Gini(D,a) &= p_1 Gini(D_i) + p_2 Gini(D_2) \\ &= \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D2|}{|D|} Gini(D_2) \end{aligned} \tag{7}$$

Then information gain of classifying by feature a is $Gini(D,a) - Gini(D)$

3.3.4 CART for Regression

As for the regression[10], the only problem is that our label becomes continue, compared with discrete in classification. In this case, we can't tell whether the division is good or

not by Gini point. Again, we introduce Mean Squared Error(MSE) as the replacement: we here calculate the MSE loss for each criteria feature:

$$\min_{j,s}[\min_{c_1}\sum_{x_i\in\mathbb{R}_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in\mathbb{R}_2(j,s)}(y_i-c_2)^2] \tag{8}$$

Where j is the different feature, s is the selective dividing point. Since a continuous feature has many values, it's tricky to pick the optimal one. The contents in middle bracket indicating using mean square error to find s for characteristic j.

Moving on, we use the selected (j,s) partition the sample area and decide the corresponding output:

$$\begin{aligned} R_1(j,s)=x|x^{(j)}<s \quad R_2(j,s)=x|x^{(j)}\geq s \\ c_m=\frac{1}{N_m}\sum_{x_i\in R_m(j,s)}y_i \quad x\in R_m, m=1,2 \end{aligned} \tag{9}$$

Equation(9) divides the dataset into two nodes and find the sum of the mean square error of each node. The regression process iterates according to the formula until certain conditions are met, and eventually split the feature space into M regions $R_1, R_2, \cdots, R_M$, the decision tree can be described as:

$$f(x)=\sum_{i=1}^{M}c_m I(x\in R_m)$$

where function I decides which region the x belongs to.

3.3.5 How XGBoost Works

XGBoost[4] keeps adding trees as well as feature splitting to plant a tree. Each time a tree is added means a new function is generated to fit the residuals of the former prediction. XGBoost will return K trees after training, when inputting a sample $x_i = [x_{i1}, x_{i2}, \cdots, x_{id}]$. In each tree x_i will fall to a corresponding leaf node basing on its features $x_{it}, t = 1, 2, \cdots, d$. Each leaf node corresponds to a score, the final prediction is the addition of all scores: $\hat{y} = XGBoost(x_i) = \sum_{k=1}^{K} f_k(x_i)$, where $f_k(x)$ is one of the decision tree within. For simplification, $XGBoost(x_i)$ will be replaced by $\hat{y}_i$.

The object function of XGBoost is defined as:

$$\begin{aligned} Obj=\sum_{i=1}^{n}L(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \\ \text{where} \quad \Omega(f)=\delta T+\frac{1}{2}\lambda\|w\|^2 \end{aligned} \tag{10}$$

In formula(10), $\sum_{i=1}^{n} l(y_i, XGBoost(x_i))$ is used to evaluate the prediction and the ground truth, L is the MSE loss. $\sum_{k=1}^{K}\Omega(f_k)$ is the regularization term to prevent overfitting problems. T denotes the number of leaf nodes, w is the point of the leaves. δ can be used

to control the leaf number and λ can control the points. As mentioned ahead, the newly generated tree is to fit the residuals of the last prediction, i.e., when t trees are generated, the prediction scores can be written as

$$\hat{y_i}^{(t)} = \hat{y_i}^{(t-1)} + f_t(x_i)$$

then we can rewrite object function as:

$$Obj^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)) + \sum_{k=1}^{K} \Omega(f_k) \tag{11}$$

which makes the problem finding a f_t each turn to minimize the Obj. This can be approximated by Taylor second-order expansion at $f_t = 0$:

$$Obj^{(t)} \approx \sum_{i=1}^{n} [L(y_i, \hat{y_i}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \sum_{k=1}^{K} \Omega(f_k) \tag{12}$$

where g_i is the first-order derivative $\partial_{\hat{y}^{(t-1)}} L(y_i, \hat{y}^{(t-1)})$ and h_i is the second-order derivative $\partial^2_{\hat{y}^{(t-1)}} L(y_i, \hat{y}^{(t-1)})$. Because the prediction of the first $t-1$ trees has nothing to do with current tree's object function, it can be ignored:

$$Obj^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \sum_{k=1}^{K} \Omega(f_k) \tag{13}$$

The above equation adds up the loss function values of each sample, since know that each sample will eventually fall into a leaf node, so we can regroup those from the same one:

$$\begin{aligned} Obj^{(t)} &\simeq \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \sum_{k=1}^{K} \Omega(f_k) \\ &= \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 + \delta T \\ &= \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \delta T \end{aligned} \tag{14}$$

Then the object function turns into a quadratic function related to leaf point w, which can be easily solved.

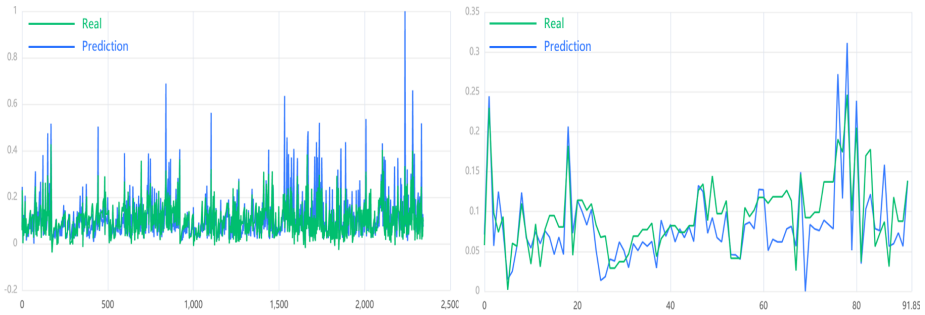
4 Results and Discussion

Due to the social nature of transactions between two individuals, the factors influencing them may exist in various aspects, including but not limited to emotions, intermediaries, etc. This makes our model too complex and deviates from the purpose of constructing a mathematical model. To address this issue, we propose some reasonable assumptions that can help us reduce the complexity of the model, making it more oriented towards a mathematical model rather than a social one, and improving the generalizability of the model.Our proposed assumptions are as follows:

1. The purpose of the transaction between the two parties is to maximize their respective interests, i.e., both the buyer and the seller pursue maximum benefit.
2. The price of a sailboat depends solely on the value determined by its own physical characteristics.
3. Sailboats of the same model have identical physical characteristics and do not change due to geographical reasons.
4. All sailboats begin to be used from the moment they are manufactured, and are used to the same extent each year, during which time sailboats are not damaged.

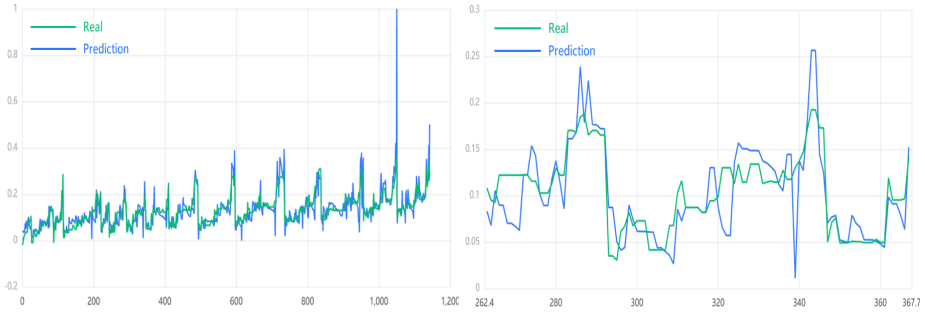
4.1 Question 1

Based on the data we collected, we selected Listing Price (USD), Year, LWL (ft), Beam (ft), Draft (ft), Displacement (lbs), Sail Area (sq ft), and GDP as variables and performed a multivariate linear regression analysis. Figure9,10 shows the regression's performance of monohull sailboats and catamarans respectively. As you can see, the performance are satisfied, meaning that linear regression successfully fits the excel data.



(a) Global view (b) Partial view

Figure 9: Visualization for Monohulls by multivariate linear regression



(a) Global view (b) Partial view

Figure 10: Visualization for Catamarans by multivariate linear regression

The mathematical results are concluded as follow:

Monohull: $ListingPrice = 0.059 + 0.038 * Draft + 0.043 * GDP + 0.451 * Displacement - 0.33 * Beam - 0.139 * LWL + 0.114 * Year + 0.2 * Length - 0.006 * SailArea$

Catamarans: $ListingPrice = -0.014 - 0.043 * Draft + 0.028 * GDP - 0.097 * Displacement + 0.004 * Beam + 0.092 * LWL + 0.112 * Year + 0.092 * Length + 0.196 * SailArea$

For monohull sailboats, the linear model demonstrates that among these predictors, Displacement is the most important predictor of sailboat listing price, followed by Beam and LWL, while the other predictors such as Draft, GDP, Year, Length, and Sail Area, have smaller coefficient values and therefore have a relatively weaker effect on the sailboat's listing price.

As for catamarans, it seems that Displacement, Beam, and Draft are not particularly useful predictors of listing price, as they have a negative relationship with the response variable. On the other hand, LWL, GDP, Year, Length, and Sail Area all appear to have a positive impact on the listing price, and thus could be useful predictors for sailboat brokers to consider when setting prices for their vessels.

In addition, we use

$$Loss = \frac{|Prediction - Real|}{Real} \times 100$$

as the precision to estimate each sailboat variant's price as prediction criteria, the lower the $Loss$, the accurate the result is. The variant of top accuracy are listed in the following table4.4, where only those whose prediction loss within 1% are included.

Class	Variant	Precision
Monohull	Nordship DSC 38	0.34586
Monohull	Beneteau Oceanis 43	0.35831
Monohull	De Stadt Stainless Steel Cutter	0.57971
Monohull	Jeanneau Sun Odyssey 479	0.75624
Monohull	Hanse 508	0.924996996
Catamarans	Broadblue 435	0.03207
Catamarans	Nautitech 442 Owners Version	0.28517
Catamarans	Chris White Atlantic 48	0.42238

4.2 Question 2

In this part we explore the relationship of "Country/Region/State" and "GDP" with "Listing Price" respectively. We used Spearman's correlation analysis[11] as the tool to obtain the significance level P for each feature, and visualize their results in Figure11. $P_{\text{"Country/Region/State"}}$ and $P_{\text{"GDP"}}$ are both estimated less than 0.05, indicating that both "Country/Region/State" and "GDP" have correlation with "Listing Price".

The practical significance of this data is that in real life, sailboats generally command higher prices in regions with higher GDPs. The statistical significance is that if a sailboat

(a) With Country/Region/State (b) With GDP

Figure 11: Visualization of Spearman Correlation Analysis

is taken to a region with a high GDP, it is more likely to sell for a better price. The difference between practical and statistical significance lies in the fact that one reflects real-life situations, while the other provides an indication of the general trend.

4.3 Question 3

In the previous models, the factor that had the greatest impact on prices among geographic regions was GDP, while the influence of other factors was very small. The geographic and economic environments of Hong Kong (SAR) can be approximated by similar regions in the sample. Therefore, as long as we know the GDP of Hong Kong (SAR), we can use it in the model to estimate prices.

We chose "Bavaria 39 Cruiser" and "Lagoon 450" from Monohulls and Catamarans as the selected subsets. After standardizing the data and incorporating the GDP of Hong Kong SAR, which is 362 billion USD, into the model, the following result was obtained:

Variant	**Year**	**Listing price(Real)**	**Listing price(Prediction)**	**Deviations**%
Bavaria Sport 39	2013	240000	234782.0188	2.174158833
Lagoon 450	2017	685000	619851.253	9.510765985
Lagoon 450	2016	560000	597491.253	6.694866607
Lagoon 450	2012	500000	508051.253	1.6102506
Lagoon 450	2014	538500	552771.253	2.650186258

According to the table above, the predictions of the model are relatively satisfactory. At the same time, we can observe from the table that the closer the year of sailboat construction is to the present, the greater the deviation in predicted price. Through consulting relevant literature, we offer the following explanation for this phenomenon:

(1) Most sailboats in the training sample were constructed a long time ago, so the model performs better for boats manufactured further in the past.

(2) In Hong Kong (Special Administrative Region), sailing has only become popular in recent years, and market demand has increased, which has affected the prices of sailboats and caused a certain degree of fluctuation.

Due to the insufficient data on second-hand sailboats in Hong Kong, we believe that the regional effects on catamarans and monohulls in Hong Kong (Special Administrative Region) are similar around 2012-2014. In the models for monohulls and catamarans, the most important features are the physical characteristics of the sailboats themselves, so the regional effects on catamarans and monohulls are basically the same.

4.4 Question 4

The trading amount of sailboats maintained an upward trend before 2007, but began to continuously decline in 2008. In 2015, there was a brief rebound, but the overall trend remains downward. The reason for this is that in 2008, a financial crisis that affected capitalism occurred, causing the trading volume of sailboats to begin and continue to decline.

Listing Price (USD)	**Year**
537240346	Total
33181459	2005
39928015	2006
53811182	2007
50897108	2008
33262766	2009
28060007	2010
33561620	2011
34500506	2012
32561388	2013
28198113	2014
42686072	2015
32767049	2016
39604640	2017
30440215	2018
23780206	2019

The total trading volume of sailboats in various countries and regions varies greatly. It is not surprising that some smaller countries and regions have lower trading volumes, but it is unexpected that Croatia has consistently maintained high trading volumes and ranks first in total trading volume. We believe that this is because Croatia has a long coastline and is located in an economically developed area, naturally leading to maritime development. With many shipyards, Croatia takes advantage of its geographical position and sells sailboats to various parts of the world, thus creating the highest trading volume.

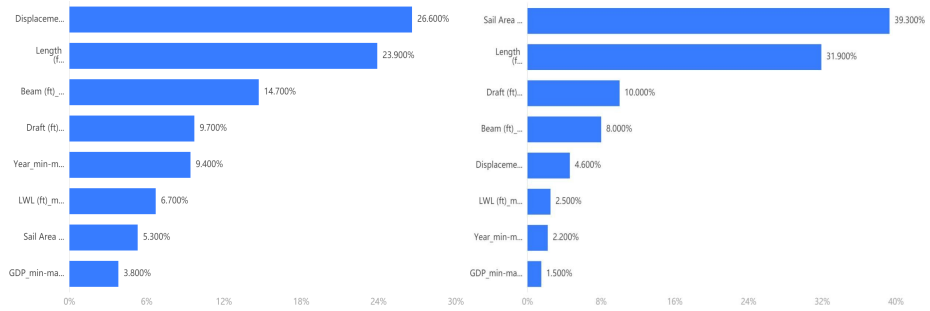
Country/Region/State	Listing Price (USD)	Country/Region/State	Listing Price (USD)
Netherlands Antilles	86000	Michigan	3049500
Cork	95961	Denmark	3129588
West Indies	125000	New York	7453725
Aruba	155000	Germany	11066786
Mississippi	169000	Maryland	11267300
Alabama	176000	British Virgin Islands	11305662
Belize	179000	Florida	14225244
Saint Kitts and Nevis	179000	Netherlands	16261090
Hawaii	710000	Turkey	17278690
Norway	753538	California	20442025
South Carolina	1638700	United Kingdom	24943382
North Carolina	1669700	France	49062112
Ireland	1710823	Greece	50532390
Wisconsin	1814800	Spain	55942598
Antigua and Barbuda	2187396	Italy	72482016
Mexico	2884999	Croatia	78978426

5 Report for the sailboat broker

Dear Hong Kong broker:

As per your request, we have developed a mathematical model to explain the pricing of used sailboats using the provided data. We modeled the prediction as a regression problem, and used machine learning algorithm to explain the listing price of each sailboat variant, including any useful predictors, and to examine the effect of geographic region on listing prices. We offer our analysis and methods which hopefully can help you with your business in Hong Kong.

Our models showed that the specifications of sailboats matters in deciding the prices, and the weight of each features varies according to the ship type. Basing on our data, statistically speaking, monohulls' displacement is tightly bound to its price, while the catamarans buyers take more concern to the sail area. And length has considerate influence in both classes. You can find more details in Figure12. You can integrate this finding with your expertise to draw more useful insights in the future.



(a) Monohull Sailboat (b) Catamarans

Figure 12: Feature Importance Chart for Sailboats

Apart from physical characteristics, the regional traits deserve your due attention. We used our model to have the influence of GDP, unemployment rate, inflation rate tested. And the results demonstrate that GDP information is a helpful feature during the prediction. That's because GDP reflects the abundance of a certain area, which is highly relevant with listing price. Yet, we noticed that sailboat purchases happen among the wealthy group while GDP is a more commonly standard. Therefore, introducing statistics more related to the rich may produce more valuable results.

However, we have to admit that our results have limitation because of the quality of data and representative samples. For example, the subset of our training data where "Country/Region/State" is "Hong Kong" does not exist, and the test data has less than 10 related samples. Thus, strictly whether the prediction can be trusted remains a doubt. Besides, the training label is the records of year 2020, the purchase power varies by year, making the numerical results can not be directly mapped to the current market.

What's more, we proposed a system design which can be used to form a search engine for your company. Figure13 shows the idea: Whenever you want to know how much

can a type sailboat sold in a specific location. You can input the variant name and target market to the searching system. The searching system will maintain 3 separate data bases, each responsible for provide features of certain realistic characteristics given the input. Afterwards, the selective features will be combined into a feature vector and sent to the price predictor. Since the predicted value is biased because of the metrics in price, in order to produce a practical price to refer,we strongly suggested a Price Calibration added ahead of the final prediction.

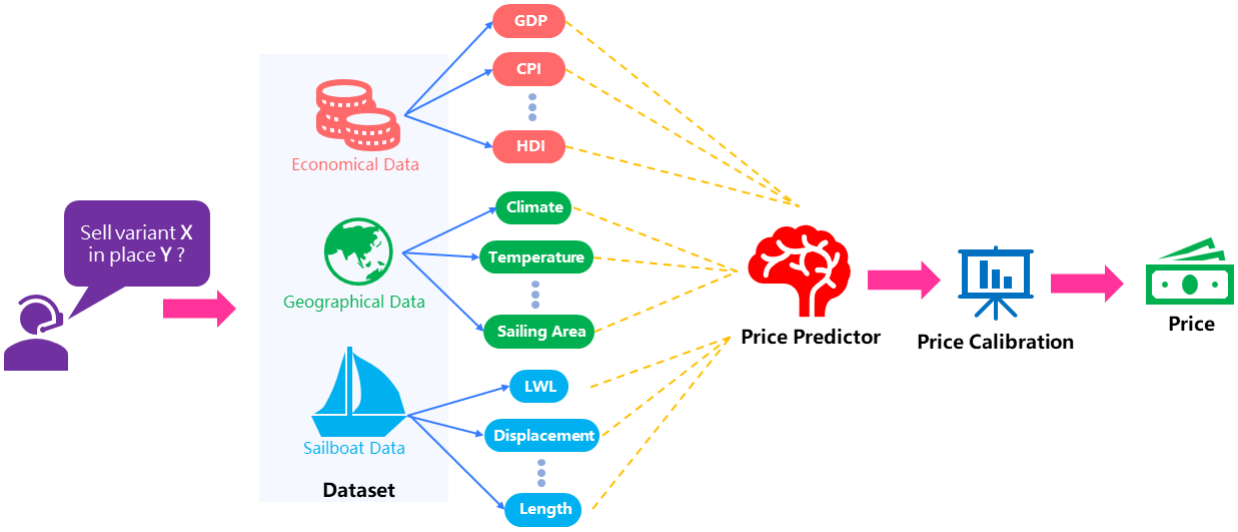


Figure 13: Proposed Price Predicting System

In conclusion, our analysis has shown that both linear regression and the XGBoost algorithm are reliable methods for predicting sailboat transactions. However, due to the sparse data on Hong Kong sailboat transactions, it may be necessary to supplement the data with information from other regions for reference. To address this data gap, we recommend collecting both physical features such as sailboat specifications, and geographical characteristics such as economic data and weather statistics to train the prediction model. We hope that this report will be useful to you in making informed decisions and wish you the best of luck in your business.

Yours sincerely,

Team 2332034.

References

[1] "Introduction of sailboats." https://www.boats.com/on-the-water/sailing-101-sailboat-types-rigs-and-definitions/.

[2] S. Rosen, "Hedonic prices and implicit markets: Product differentiation in pure competition," *Journal of Political Economy*, vol. 82, pp. 34–55, 02 1974.

[3] "Introduction of multivariate linear regression." https://www.hackerearth.com/practice/machine-learning/linear-regression/multivariate-linear-regression-1/tutorial/.

[4] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *CoRR*, vol. abs/1603.02754, 2016.

[5] "Yachtworld website." https://www.yachtworld.com/.

[6] "Introduction of web crawler from wikipedia." https://en.wikipedia.org/wiki/Web_crawler.

[7] J. MacQueen, "Some methods for classification and analysis of multivariate observations," 1967.

[8] "Introduction of linear regression from wikipedia." https://en.wikipedia.org/wiki/Linear_regression.

[9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees (cart)," *Biometrics*, vol. 40, no. 3, p. 358, 1984.

[10] "Machine learning - classification and regression trees (cart)." https://wiki.q-researchsoftware.com/wiki/Machine_Learning_-_Classification_And_Regression_Trees_(CART).

[11] "Introduction of spearman correlation coefficient." https://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient.