Problem Chosen	2023 MCM Summary Sheet	Team Control Number
Y		2332102

Title

Summary

Sailboats fulfill diverse roles, fueling a thriving secondary market. Brokers, facing numerous complex factors, struggle to determine reasonable pricing. A tool to assist in comprehensive evaluations and rational pricing for used boats is urgently needed.

This paper aims to construct a reliable model based on existing datasets, which can provide a reasonable explanation for the pricing of the second-hand sailboat market. It also analyzes the impact of different factors and indicators on prices. Finally, the model will be applied to the second-hand sailboat market in Hong Kong to provide a reasonable and accurate pricing rule.

For Problem(a), …

For Problem(b), …

For Problem(c), …

For Problem(d), …

For Problem(e), …

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

Keywords: Linear interpolation, Adaptive Density-Based Clustering, Heuristic Hierarchical Multiple Regression, Deep Forest Model, Machine Learning, Data Mining, Analysis of Variance.

Contents

1 Introduction

1.1 Background

In our daily lives, sailboats are not only a means of transportation, but also serve as leisure and entertainment, and even for competitive sports. As a result, the growing demand for sailboats has given rise to a thriving boat market, which has gradually developed into a secondary market. In the secondary market, buyers and sellers usually trade through brokers, who play a crucial role in the transaction process.

For brokers, it is essential to be familiar with the used sailboat market, comprehensively consider various factors, and make reasonable pricing for the used sailboats in order to facilitate a successful transaction. However, the factors affecting the price of used sailboats are numerous and complex, with different brands, variants of boats, years, depreciation rates, as well as local consumption levels and geographical environments having significant impacts. The intertwined influences of these complex factors make it difficult to determine the pricing in the used sailboat market, and it is challenging to come up with a reasonable price that takes all factors into account.

Therefore, brokers urgently need a tool to assist them in making more reasonable and comprehensive evaluations of used sailboats, and to make the pricing in the used sailboat market more rational.

1.2 Problem Restatement

Problem (a) :

- Develop a prediction model to explain the listing price of each of the sailboats in the provided spreadsheet.
- Discuss the precision of our estimate for each sailboat variant's price.

Problem (b) :

- Determine whether region has an impact on the price of second-hand boats and explain the effect.
- Discuss whether any regional effect is consistent across all sailboat variants.
- Address the practical and statistical significance of any regional effects noted.

Problem (c) :

- Based on the model, find out how it can be useful in the Hong Kong market.
- Choose one subset and model the regional effect of Hong Kong on each sailboat prices.
- Assess whether the effect is the same for both catamarans and monohull sailboats.

Problem (d) :

- Identify and discuss additional informative conclusions drawn from the data.

Problem (e) :

- Create a one-to two-page report with well-chosen graphics to assist the Hong Kong sailboat broker to understand your findings.

1.3 Our work & Model Overview

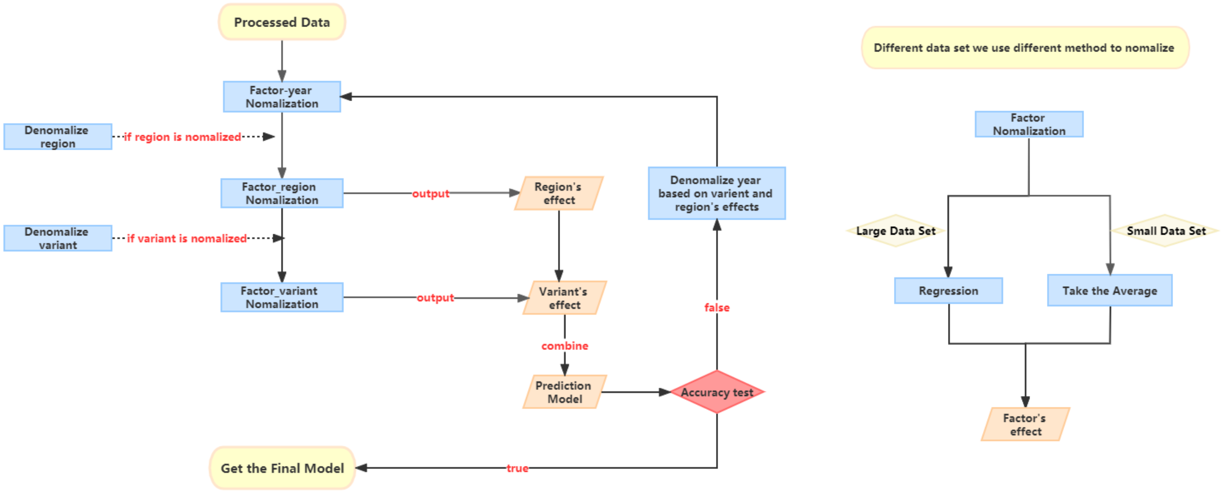


Figure 1: Model Framework

2 Assumptions and Justifications

2.1 Assumptions

To simplify our problems, we make the following basic assumptions, each of which is adequately justified.

- The price of used sailboats is solely determined by the factors in the dataset.
- The factors in the dataset are independent and unrelated.
- The data in the dataset are all real, reasonable, and follow a certain pattern.
- The pricing required by a broker should be reasonable and in accordance with market rules, rather than false pricing.

2.2 Notations

Symbol	Definition
A	the first one
b	the second one
α	the last one

3 Data Exploration

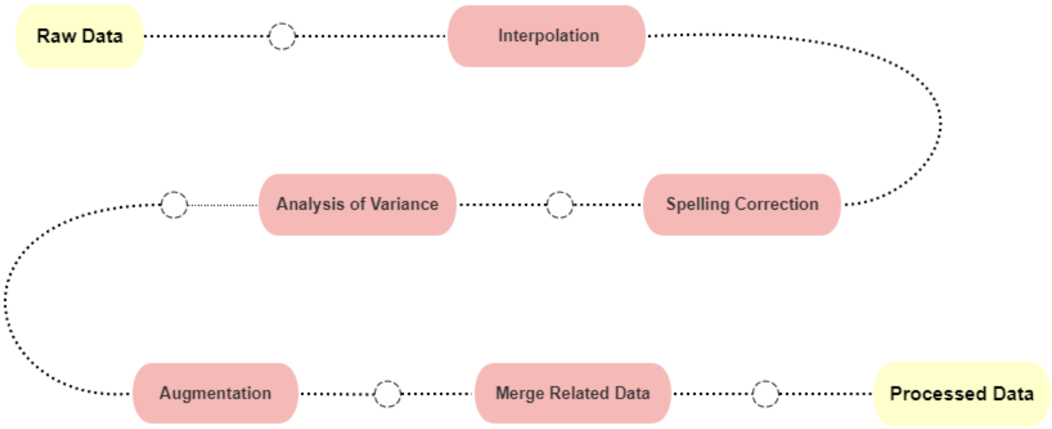


Figure 2: Data Exploration

3.1 Data Cleaning

First of all, we turn the `xlsx` format data sheet into `csv` format. The conversion causes some minor errors like extra spaces and unexcepted characters which can be easily filtered out by text editor and python string operating functions like `strip()` or so.

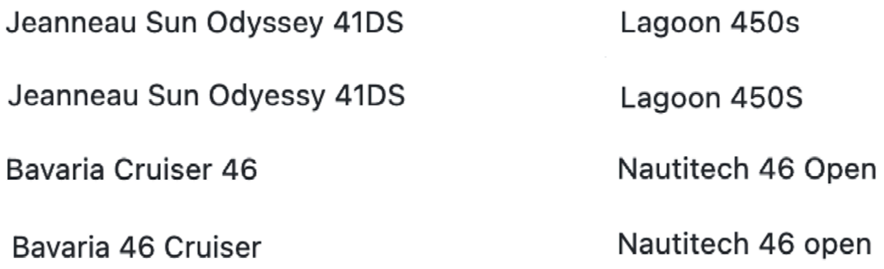
Secondly, we use **Linear interpolation** to fill in the missing data and then apply **Adaptive Density-Based Clustering** for spell correction, resulting in a more complete and accurate dataset than the original.

Further more, we merge **Make** and **Variant** into a single feature, which we refer to as **Variant**.

Finally, we merge the corrected dataset with the accurate one after making necessary modifications.

3.2 Data Augmentation

We collect and organize more additional features of a given sailboat(such as **beam**, **draft**, **displacement**, **cabins**, etc.), and the **2020 per capita GDP data** of the relevant regions, greatly en-



Jeanneau Sun Odyssey 41DS

Jeanneau Sun Odyessy 41DS

Bavaria Cruiser 46

Bavaria 46 Cruiser

Lagoon 450s

Lagoon 450S

Nautitech 46 Open

Nautitech 46 open

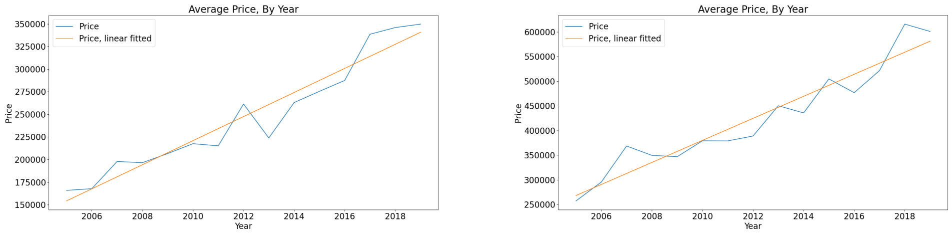
Figure 3: Examples of error data

riching the diversity of the data and expanding the dataset size. These efforts lay a solid foundation for subsequent modeling.

3.3 Data Analysis

There are lots of labels related to specifications, and it is hard for us to discuss their influence on listing price one by one. However, trying to directly preform regression based on these labels by machine learning models can cause **overfitting issues**. Nevertheless, we can still draw some relatively simple conclusions from the data, which provide theoretical support for our subsequent modeling:

- Time: Time is a quasi-continuous and ordered variable, and the price has an obvious linear relationship with time.(see Figure 4 and Table 1)
- Variant and Region: They are obviously related to the price, but due to their unordered and non-continuous nature, further analysis is needed.



(a) Monohulled (b) Catamaran

Figure 4: The relationship between year and price

	Pearson	Spearman	Kendall
Monohulled	0.964	0.989	0.943
Catamaran	0.960	0.971	0.867

Table 1: The correlation coefficient of year and price

4 The Prediction Model of Used Sailboat Prices

4.1 The Establishment of The Model

4.1.1 Model 1 —Heuristic Hierarchical Multiple Regression

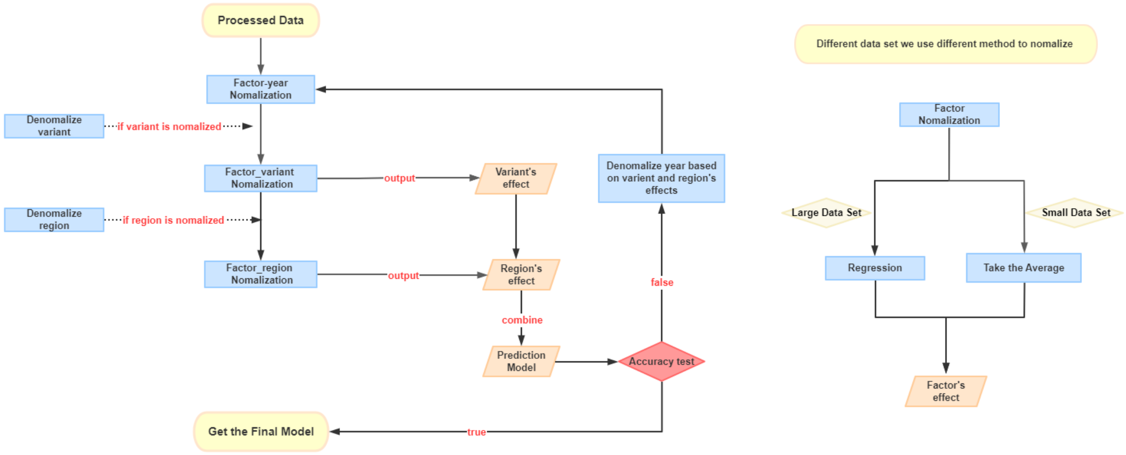


Figure 5: Brief process of Model 1

We propose *Model 1* as shown in Figure 5, The model takes `Year`, `Region`, `Variant` as inputs, , `Price` as output, and performs a heuristic hierarchical multiple regression. Specifically, in each regression layer, one variable is denormalized, and the other two variables are normalized. The obtained `Price` at this time is used as the Effect of this variable. At the same time, the fitting result is used as the baseline for normalization in the next layer and is involved in the regression. This process is iterated continuously until the model trains a precise fitting effect.

It is worth noting that we use a heuristic approach for each normalization. For larger datasets, we perform regression, while for smaller datasets, we take their average. This method greatly improves the accuracy of the model predictions.

4.1.2 Model 2 —Deep Forest Model

However, *Model 1* is somehow incomplete because *Model 1* requires that the variant and region of the predicted data have appeared in the dataset before. However, in actual prediction, the variant and region may be opaque to the model. Taking the third question as an example, *Model 1* can't predict the price of Hong Kong because the region effect of Hong Kong hasn't been evaluated by the model. Therefore, we use the **Deep Forest Model** to establish *Model 2*.

Decision tree is a common machine learning method with two models: classification and regression. At the beginning, we used the basic decision tree for regression, but the model error was large, and the accuracy for price prediction was poor. Therefore, we optimized the model using the deep forest algorithm. Deep forest is an ensemble learning model based on random forest. Unlike traditional decision trees, deep forest trains each tree classifier with a randomized strategy to increase the diversity of the model. Compared with a single decision tree, the deep forest model is more robust.

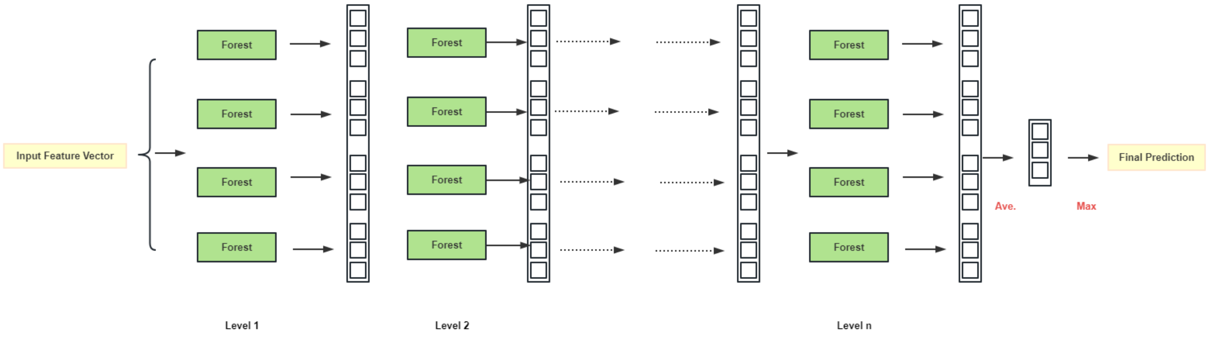


Figure 6: Deep Forest Model

As for *Model 2*, we first expanded the dataset to obtain quantified data for regions and variants. Specifically, for variants, we expanded data such as `draft`, `beam`, `water tank`, `fuel tank`, `cabins`, `displacement`, etc. For regions, we expanded data such as `GDP`, `Gini coefficient`, `Human Development Index`, etc. When dealing with unmarked variants and regions, we first used *Model 1* to obtain the effects of marked variants and regions, and then applied the effects of marked variants and regions to the deep forest. By expanding the data for regression, we obtained the effects of unmarked variants and regions, and thus predicted the price.

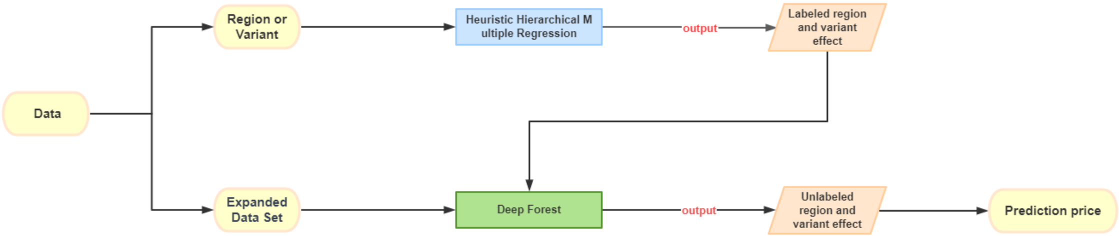

Region or Variant
Heuristic Hierarchical M ultiple Regression
output
Labeled region and variant effect
Data
Expanded Data Set
Deep Forest
output
Unlabeled region and variant effect
Prediction price

Figure 7: Brief process of Model 2

4.2 Model Validation

4.3 Model Accuracy Analysis

Due to the limited data provided for some sailship types, they belong to imbalanced datasets. Therefore, we need to use as much data as possible for training to prevent the loss of small-sample data during training. Given that the leave-one-out method is not affected by random partition, it is effective for small-sample and imbalanced datasets, and can accurately evaluate the performance of the model, we used the **leave-one-out method** to analyze the model's error in the error analysis part, and adopted mean squared error as the performance measure of the model.

Given a set of examples in a prediction task $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$, In which y_i

represents the true label of sample $\mathbf{x}_i$, and mean squared error is represented as:

$$E(f;D) = \frac{1}{m}\sum_{i=1}^{m}(f(\mathbf{x}_i) - y_i)^2$$

5 Regional Effect Analysis

6 The applicability of The Prediction Model in Hong Kong

7 Extended Inferences or Conclusion

8 Further Improvements

9 Strengths and Weaknesses

9.1 Strengths

- First one...
- Second one ...

9.2 Weaknesses

- Only one ...

10 Conclusion

Report

To: Heishan Yan
From: Team 1234567
Date: October 1st, 2019
Subject: A better choice than MS Word: LaTeX

In the memo, we want to introduce you an alternate typesetting program to the prevailing MS Word: **LaTeX**. In fact, the history of LaTeX is even longer than that of MS Word. In 1970s, the famous computer scientist Donald Knuth first came out with a typesetting program, which named TeX …

Firstly, …

Secondly, …

Lastly, …

According to all those mentioned above, it is really worth to have a try on LaTeX!

References

[1] Einstein, A., Podolsky, B., & Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete?. *Physical review*, 47(10), 777.

[2] *A simple, easy LaTeX template for MCM/ICM: EasyMCM*. (2018). Retrieved December 1, 2019, from `https://www.cnblogs.com/xjtu-blacksmith/p/easymcm.html`

Appendix A: Further on LaTeX

To clarify the importance of using LaTeX in MCM or ICM, several points need to be covered, which are …

To be more specific, …

All in all, …

Anyway, nobody **really** needs such appendix …

Appendix B: Program Codes

Here are the program codes we used in our research.

test.py

```python
# Python code example
for i in range(10):
    print('Hello, world!')
```

test.m

```matlab
% MATLAB code example
for i = 1:10
    disp("hello, world!");
end
```

test.cpp

```cpp
// C++ code example
#include <iostream>
using namespace std;

int main() {
    for (int i = 0; i < 10; i++)
        cout << "hello, world" << endl;
    return 0;
}
```