| Problem Chosen | 2023 MCM/ICM Summary Sheet | Team Control Number |
|---|---|---|
| Y | | 2331406 |

# Used Sailboat Price Prediction For Hong Kong Broker


## Summary

As the development of the Hong Kong sailboat transaction market, a sailboat broker in Hong Kong asked us for advice about pricing for used sailboats to help him better understand the sailboat market and make informed decisions. To fulfill his task, we need to develop a model that explains the listing price of each sailboat and the effects of sailboats' features in specific regions.

For task 1, after data pre-processing, we use **Random Forest (RF) and Gradient Boosting method(GBRT),** respectively, and use **Ensemble Learning** to combine the two models. The model fits quite well with R-square = 0.8566, and the predicted price is quite a suit for the true price as 2/3 of the prediction results fall into the 95% confidence interval of its sailboat variants.

For task 2, we figure out the **regional effects** regarding listing prices by applying the **Point Bi-serial Correlation test and their significance.** Also, we use **Partial Dependence Plots (PDP)** to judge that the regional effect of “Caribbean” is consistent across all variants, with the other two regions being inconsistent.

For task 3, we select a **subset** from the given spreadsheet and find comparable data on Hong Kong from the website. Then we model the **regional effect of Hong Kong** and find that they are not the same for monohull sailboats and catamarans by **PDP** again.

Finally, we write a memorandum including our model, results, and suggestions for the broker. We hope this memorandum will be a valuable reference.



**Keywords:** sailboats, regional effect, RF, GBRT, ensemble learning, PDP

# 1. Introduction

## 1.1 Problem Background

In recent years, being affected by the COVID-19 pandemic, more people in Hong Kong have chosen sailing as a substitute for entertainment, which makes used sailboats popular. However, due to the economic change and global economic trend of the sailboat market, how to grasp this business opportunity and price them appropriately to maximize benefits has been a big problem for brokers.

## 1.2 Problem Analysis

A broker came to us for a report about the pricing of used sailboats based on previous listing price data from 2005-2019 from three regions: the Caribbean, the USA, and Europe.

To achieve this requirement, we must:

- Get additional features about sailboats, relevant countries' GDP information by year, and comparable listing price data from the Hong Kong market.
- Pre-process the data provided by COMAP officials and do a data stitch with the data we've collected.
- Make a listing price prediction model and test its precision by sailboat variants.
- Analyze the practical and statistical importance of the listing price prediction model regarding the regional effect.
- Model the regional effect of Hong Kong and test its result with respect to catamarans and monohull sailboats, respectively.
- Write a memorandum including our model, precision, and results for the broker as a reference for pricing.

Our modeling framework can be illustrated as shown in Figure 1:

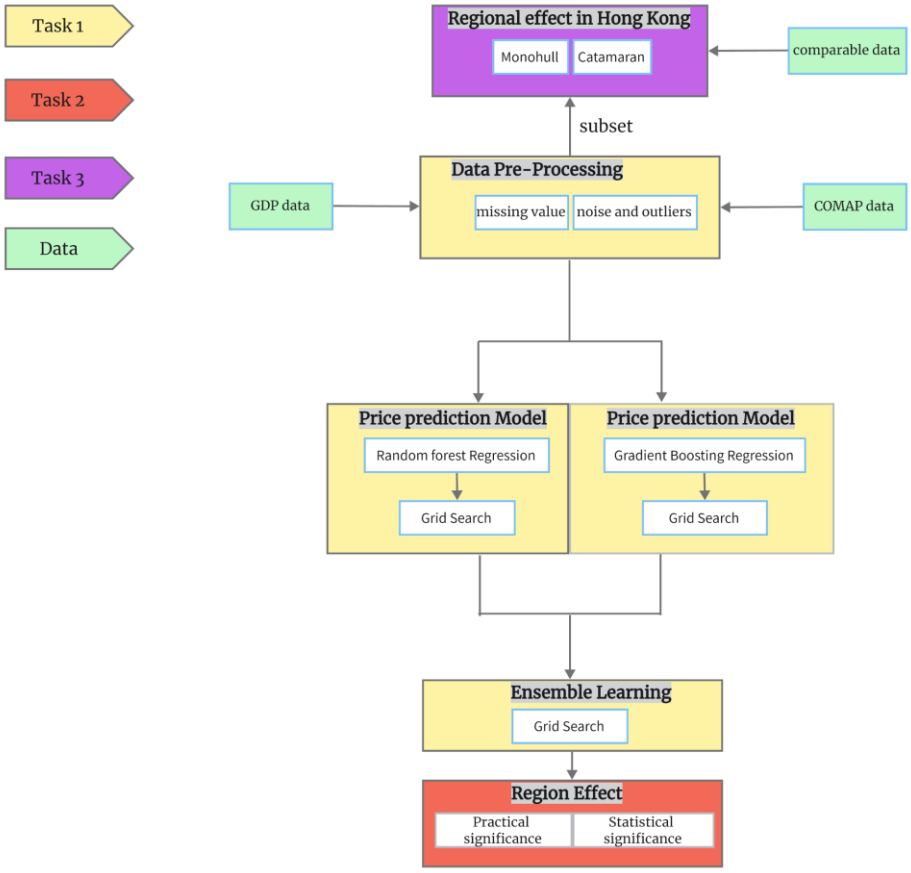



Figure 1: Modeling Framework

## 2. Model Assumptions and Notations

### 2.1 Assumptions

- The data provided in this problem is valid and reliable.
- The GDP data collected from the online database is valid and reliable.
- The inflation of the U.S. dollar can be ignored.

Although the inflation rate of the U.S. dollar has objectively existed, we assume that all the data provided are collected at the same time, which ensures that the value of the US dollar is consistent.

- All the features are independent of each other.

By the visualization result of the heat map, the strongest correlation between two different features is 0.21, which can be ignored in our regression.

- The variance of the listing price is constant across all values of sailboat features.
- The errors (residuals) are normally distributed and have constant variance.
- The price prediction model, due to the over-complexity, neglects the effect of international politics and maritime trade.
- Also, this model neglects economic factors, like the price-changing influence of the complementary objects.
- This model neglects the influence of material re-using times on transaction price.

## 3. Task 1: Model Building and Precision Discussion regarding sailboat variants

### 3.1 New Data Search

In this section, we generated the economic data of all countries, states, and regions, which is gross domestic product (GDP) per capita, from https://databank.worldbank.org/home.aspx. For the conditions in the USA, we found the perspective data from https://www.bea.gov/tools/. We feel it challenging to find GDP data from some tiny regions, such as York in Europe. Therefore, we drop all six empty samples.

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Make | Variant | Variant_Capital | Variant_number | length | Geographic Region | Country/Region/State | Listing Price | Year | GDP | type |
| 2 | Alubat | Ovni 395 | Ovni | 395 | 41 | Europe | France | 267233 | 2005 | 34768.176 | Monohulled Sailboats |
| 3 | Bavaria | 38 Cruiser | Cruiser | 38 | 38 | Europe | Croatia | 75178 | 2005 | 10634.236 | Monohulled Sailboats |
| 4 | Bavaria | 38 Cruiser | Cruiser | 38 | 38 | Europe | Croatia | 66825 | 2005 | 10634.236 | Monohulled Sailboats |
| 5 | Bavaria | 38 Cruiser | Cruiser | 38 | 38 | Europe | Croatia | 54661 | 2005 | 10634.236 | Monohulled Sailboats |
| 6 | Bavaria | 38 Cruiser | Cruiser | 38 | 38 | Europe | Croatia | 53447 | 2005 | 10634.236 | Monohulled Sailboats |
| 7 | Bavaria | 38 Cruiser | Cruiser | 38 | 38 | Europe | Greece | 91101 | 2005 | 22560.147 | Monohulled Sailboats |
| 8 | Bavaria | 39 Cruiser | Cruiser | 39 | 39 | Europe | Greece | 66748 | 2005 | 22560.147 | Monohulled Sailboats |
| 9 | Bavaria | 42 Match | Match | 42 | 41 | Europe | Croatia | 78945 | 2005 | 10634.236 | Monohulled Sailboats |
| 10 | Bavaria | 42 Match | Match | 42 | 41 | Europe | Croatia | 58297 | 2005 | 10634.236 | Monohulled Sailboats |
| 11 | Bavaria | 42 Cruiser | Cruiser | 42 | 42 | Europe | Croatia | 112906 | 2005 | 10634.236 | Monohulled Sailboats |
| 12 | Bavaria | 42 Cruiser | Cruiser | 42 | 42 | Europe | Italy | 95961 | 2005 | 32055.092 | Monohulled Sailboats |
| 13 | Bavaria | 42 Cruiser | Cruiser | 42 | 42 | Europe | Italy | 94746 | 2005 | 32055.092 | Monohulled Sailboats |
| 14 | Bavaria | 42 Cruiser | Cruiser | 42 | 42 | Europe | Italy | 91102 | 2005 | 32055.092 | Monohulled Sailboats |
| 15 | Bavaria | Cruiser 46 | Cruiser | 46 | 46 | Europe | Croatia | 101063 | 2005 | 10634.236 | Monohulled Sailboats |
| 16 | Bavaria | Cruiser 46 | Cruiser | 46 | 46 | Europe | Croatia | 91106 | 2005 | 10634.236 | Monohulled Sailboats |
| 17 | Bavaria | Cruiser 46 | Cruiser | 46 | 46 | Europe | Croatia | 84907 | 2005 | 10634.236 | Monohulled Sailboats |
| 18 | Bavaria | Cruiser 46 | Cruiser | 46 | 46 | Europe | Croatia | 84786 | 2005 | 10634.236 | Monohulled Sailboats |
| 19 | Bavaria | Cruiser 46 | Cruiser | 46 | 46 | Europe | Greece | 133651 | 2005 | 22560.147 | Monohulled Sailboats |
| 20 | Bavaria | Cruiser 46 | Cruiser | 46 | 46 | Europe | Greece | 109338 | 2005 | 22560.147 | Monohulled Sailboats |

Figure 2: Input data

### 3.2 Data Observation

We first draw some graphs to analyze some apparent features' effects. For example, we found that sailboats in the Caribbean tend to have lower listing prices than those in other regions almost every year. Also, the length of the sailboat seems to have a weak positive relationship with the listing price. So these two features, 'Geographic Region' and 'length,' may have a significant influence on the listing price.

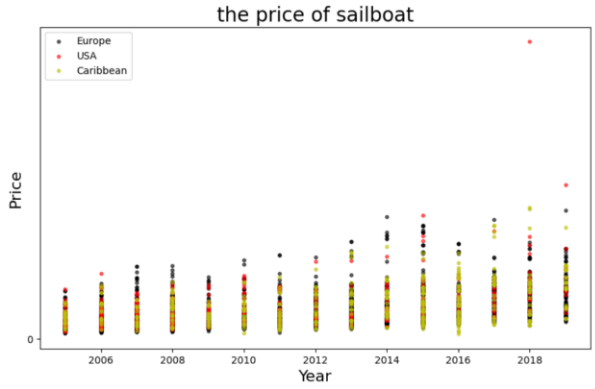



Figure 3: Price-Year scatter plot

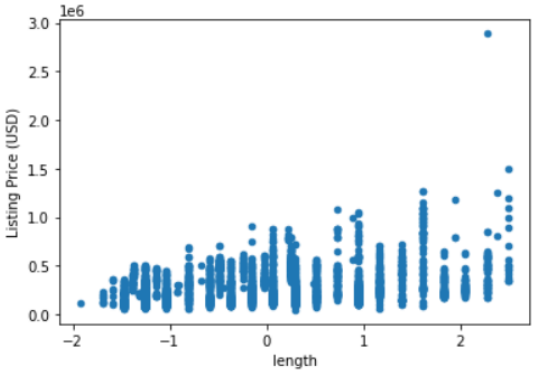



Figure 4: Price-Length scatter plot

### 3.3 Data Cleaning

#### 3.3.1 Missing Value

Since the number of missing values is just six and all the values are in column 'Country/Region/State,' we can not find proper data to fill in the missing value. Thus, we dropped all the missing values. The remaining number of data is 3494.

#### 3.3.2 Handling Duplicates and Merging

We find ten duplicates in the Monohull sailboat sheet and 72 duplicate data in the Catamaran sailboat sheet. We remove those duplicates and merge those two sheets as a single one, making “type” a dummy variable that labels those Catamaran sailboats denoted as “1” and monohull sailboats denoted as “0”.

### 3.4 Data Preprocessing

#### 3.4.1 Other Dummy Variables

For the text feature “Geographical Region” and discrete numerical feature “Year”, we replace them with dummy variables to fit them into our regression model.

#### 3.4.2 Other Text Features

We replace the text feature “Make” and “Variant” with a new variable “MV_expensive” and “Country” with “C_expensive” containing three values: {-1, 0, 1} based on their impact on listing price: those categories that have a listing value greater than *mean + std* denoted as “1”, others lower than *mean - std* denoted as “-1”, and remained ones denoted as “0”, since these outliers can significantly affect our analysis and modeling.

Using a label encoder, we transform the text feature “Country/Region/State” into a numerical value.

#### 3.4.3 Repeated Data

There are some data with all features the same but different listing prices, we think that dropping them directly will affect the accuracy of our model, so we combine them into one data with avg(listing prices) as its new listing price. However, the result has not reached our expectations. Maybe because some repeated data has a bias on listing price so the mean is the best expression. So we decided not to use it in our final model.

### 3.4.4 Data Normalization

We use *StandardScaler()* in Python to normalize our data so that it has a mean = 0 and standard deviation = 1. Normalization will ensure the equality of influence of all features and make our models more robust to outliers.

### 3.4.5 Feature Importance

First, we draw a heat map of the correlation matrix for the data by selecting the 10 most important features, as shown in Figure. Then we compute the feature importance and remove those features with an importance value less than 0.0008 to avoid useless information included.

| | | |
|---|---|---|
| 1) | type | 0.397350 |
| 2) | length | 0.269084 |
| 3) | Variant_number | 0.104993 |
| 4) | GDP | 0.046261 |
| 5) | 2018 | 0.041927 |
| 6) | MV_expensive | 0.031302 |
| 7) | 2019 | 0.030149 |
| 8) | 2017 | 0.017180 |
| 9) | Country_digit | 0.015228 |
| 10) | C_expensive | 0.009754 |
| 11) | 2016 | 0.006966 |
| 12) | 2006 | 0.005871 |
| 13) | 2015 | 0.005743 |
| 14) | 2008 | 0.004033 |
| 15) | 2007 | 0.003324 |
| 16) | 2005 | 0.003168 |
| 17) | 2009 | 0.002079 |
| 18) | 2011 | 0.001700 |
| 19) | 2014 | 0.001075 |
| 20) | 2013 | 0.000804 |
| 21) | USA | 0.000743 |
| 22) | Europe | 0.000477 |
| 23) | 2010 | 0.000326 |
| 24) | Caribbean | 0.000244 |
| 25) | 2012 | 0.000220 |

Figure 5: feature importance

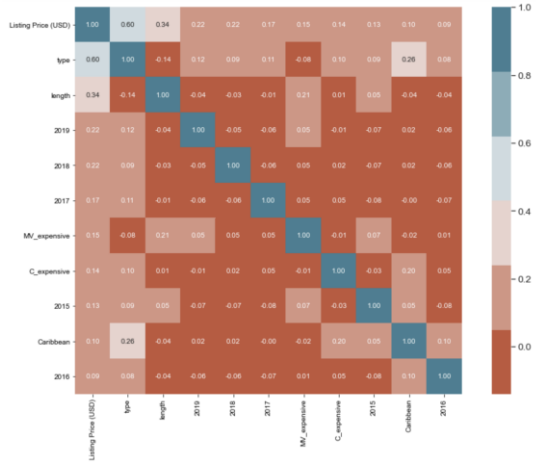



Figure 6: Heatmap of correlation matrix

## 3.5 Model Selection

### 3.5.1 Gradient Boosting Regression

In the gradient boosting regression model, we first split the data into a training set and a test set with a 5:1 proportion. Then we apply a grid search algorithm to figure out the parameters with the best performance scores, which provides a parameter list with several choices of possible values and use cross-validation to find the best ones. We use the returned best parameter to fit our data and predict the result. However, the thorough model is easy to overfit, we also try other models later.

The best performance on test data gives an R square of 0.8492. The comparison between the predicted listing price and true value and their difference is visualized in Figure 7 and 8. We can see that there are still some predictions having high differences from true values.

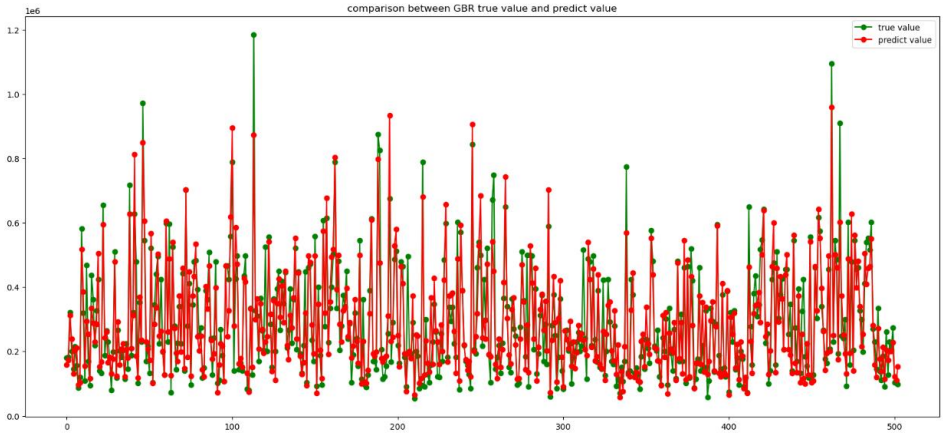



Figure 7: GBRT predict value and true value

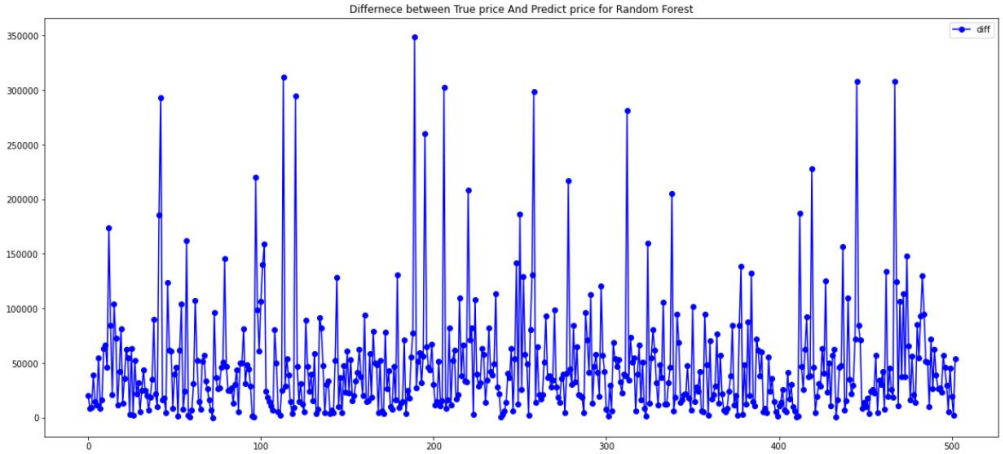



Figure 8: GBRT difference between prediction and ground truth

| dataset\method - R square | GBRT |
|---|---|
| Original data | 0.8419 |
| Data after feature selection | 0.8492 |

Table 1: Evaluation of GBRT

### 3.5.2 Random Forest Regression

In the random forest regression model, similar to the above, we apply a grid search algorithm to figure out the parameters with the best performance scores and fit our data to predict the result. The best performance on test data gives an R square of 0.8652, which is an

improvement compared with Gradient Boosting Regression. And the comparison between the predicted listing price and true value is visualized in Figure. There are still some errors, making space for further improvement of the model.

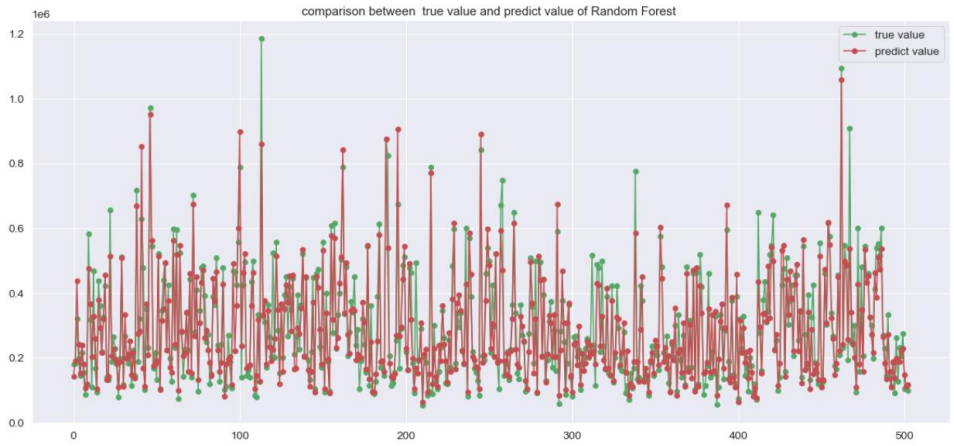



Figure 9: RF predict value and true value


Figure 10: RF difference between prediction and ground truth

| dataset\method - R square | GBRT | RF |
|---|---|---|
| Original data | 0.8419 | 0.8652 |
| Data after feature selection | 0.8492 | 0.8513 |

| | |
|---|---|
| R | 0.8652 |
| MAE | 41161 |
| MSE | 4342524502 |
| RMSE | 65897 |

Table 2: Evaluation of RF

### 3.5.3 Ensemble Learning

In this model, we combine Gradient Boosting Regressor (GBDT) with Random Forest Regressor(RF) to get better results and improve model diversification. Specifically, we use

GBDT in our first layer and RF as our second layer in a two-layer fully connected neural network. We still use a grid search algorithm to tune our hyperparameters. The final R square on test data we get is 0.8699, which is proved to be better than the result of a single GBDT or RF model.

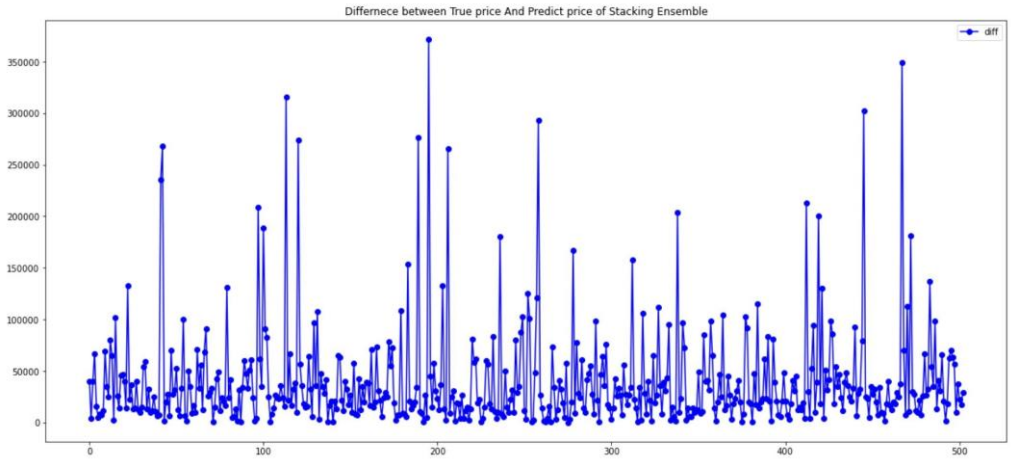



Figure 11: Ensemble learning difference

### 3.6 Results and Precision

To test the precision of our model, we use confidence interval(CI) as our evaluation metric. For each variant of sailboats, we compute a 95% CI from true data and check whether our predicted value falls in this interval. The precision of sailboats' listing price with variant i is computed by:

$$Precision_i = \frac{number\ of\ variant_i's\ listing\ prices\ that\ fall\ in\ 95\ percent\ CI}{number\ of\ sailboats\ with\ variant_i}$$

Formula 1: precision of price prediction for a certain sailboat variant

Overall, there are 105 variants' precision values. The result for all variants' CI lengths is shown in the following figure:

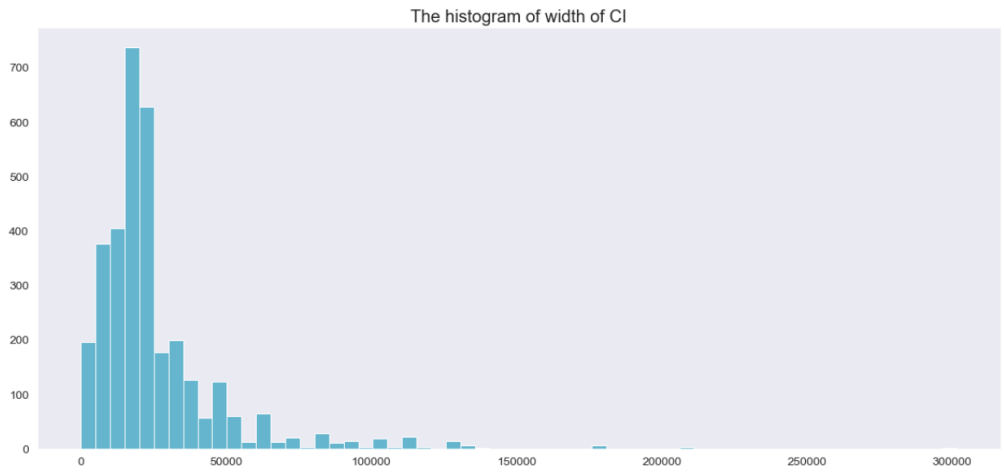



Figure 12: Histogram of all widths of 95% confidence interval

All precision values:

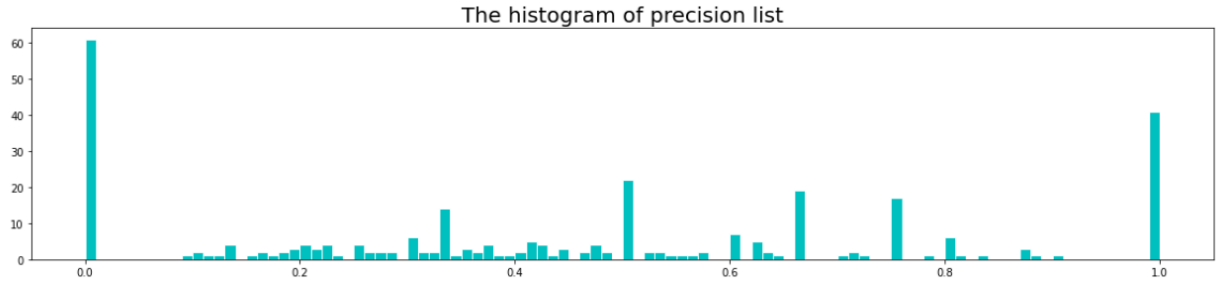



Figure 13: Histogram of all precision values

Among the precision values for all variants, the largest one is 1 with frequency = 35, which means that for 35 variants, we have predicted its listing price 95% correctly by judging CI. There are also some low precision values near 0. The cause of such a result may be a lack of original data for this variant of the sailboat, which makes our training inaccurate. We use the sample mean as our general precision. Overall, the general **precision** is **0.5026** with **R square 0.8699**.

## 4. Task 2: Analyzing Regional Effect of Model

### 4.1 Effect of the Regions on Sailboat Listing Price

To examine the influence mode of a single variable on the prediction result while controlling other features as observed values, we apply partial dependence plots(PDP) to show the marginal effect of regions on the model's predicted listing prices. The formula is listed here:

$$\hat{f}_{xs}(x_S) = E_{xC}[\hat{f}[(x_S, x_C)]] = \int \hat{f}(x_S, x_C) dP(x_C)$$

$$P(x_C) = \int p(X_{all}) dx_s$$

$$\hat{f}_{xs}(x_s) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}[(x_S, x_C^{(i)})]$$

Formula 2: Partial dependence plots

| | |
|---|---|
| $x_S$ | the features related with region:<br>"Europe", "USA", "Caribbean" |
| $f(x_S)$ | predicted listing price under each different $x_S$ value |
| $x_C$ | other features except those contained in $x_S$ |

Table 3: Notation of all parameters

The plots are shown as follows:

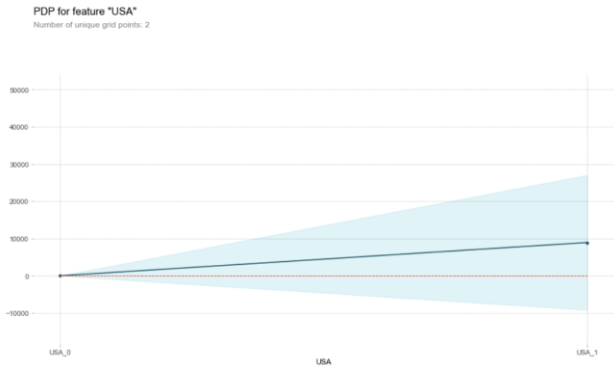



Figure 14: PDP of USA

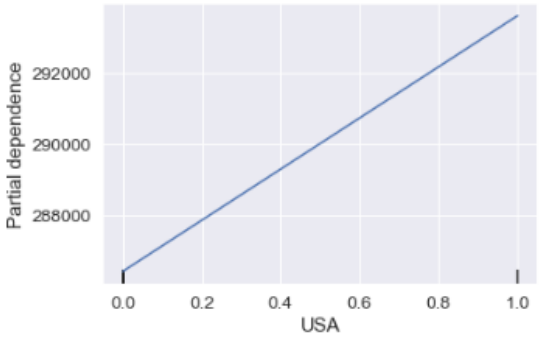



Figure 15: PDP of USA

We can see that the listing price increases when the value of the USA increases. The blue range on USA_1 is the confidence interval of listing price change between USA = 0 and USA = 1 when keeping other variables the same. Since the USA is a dummy variable with only a "0" or "1" value, sailboats in the USA tend to have a higher listing price than those not in the USA, so it has a **positive regional effect**.

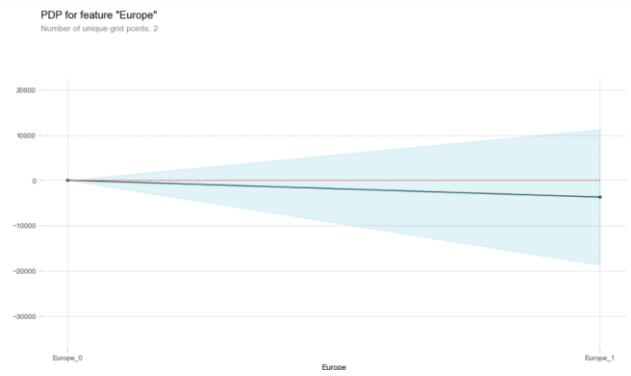



Figure 16: PDP of Europe

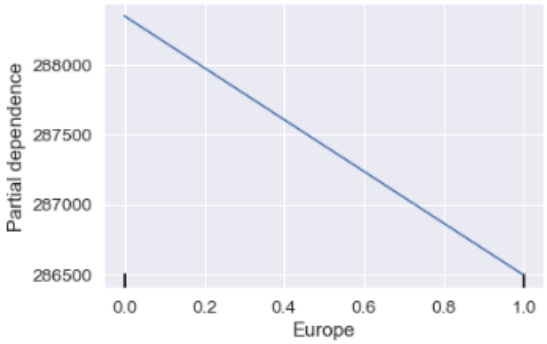



Figure 17: PDP of Europe

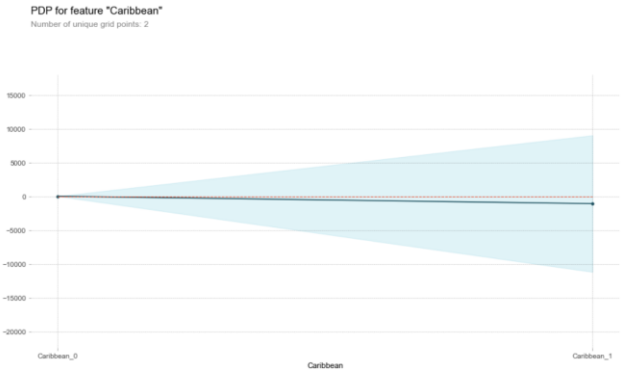



Figure 18: PDP of Caribbean

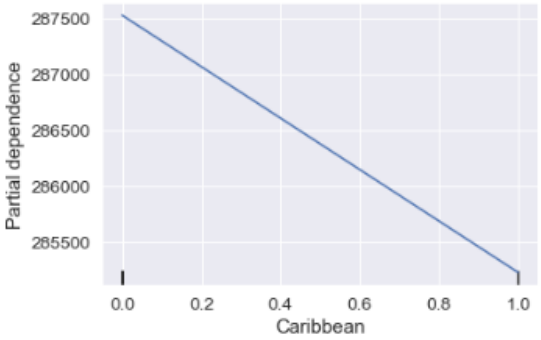



Figure 19: PDP of Caribbean

For the Caribbean and Europe, it's clear to see that both of them show a **negative effect** on listing prices, with Europe relatively **stronger** than the Caribbean, which means that sailboats in Europe tend to have the lowest listing price.

Therefore, buying a sailboat with the same class may be different in those three regions. For a customer, we'd like to recommend you buy sailboats in the Caribbean or Europe. For sellers, selling sailboats in the USA is a better choice. Due to this difference in price, the broker can earn a price spread.

## 4.2 Testing Consistency and Significance of Regional Effect on Different Sailboat Variants

### 4.2.1 Significance: Point Bi-serial Correlation(PBC) Test

Based on the property of the two variables: one of which is a binary variable, while the other being a continuous variable, we try to find a more appropriate method to analyze the correlation. The Pearson correlation number is computed by:

$$r = \frac{\bar{x}_p - \bar{x}_q}{S_x}\sqrt{pq}$$

Formula 3: Pearson correlation number

| | |
|---|---|
| $p$ | proportion of one type of regional binary variables |
| $q$ | $1-p$ |
| $\bar{x}_p$ | the mean value of one type of the continuous variable |
| $\bar{x}_q$ | the mean value of another type of the continuous variable |
| $S_x$ | the stantard deviation of the continuous variable |

Table 4: Notation of parameters

The result of the PBC is shown in the table:

| | USA | Europe | Caribbean |
|---|---|---|---|
| Correlation | 0.0257 | 0.0197 | -0.0509 |
| Pvalue | 0.1366 | 0.2544 | 0.0032 |

Table 5: PBC Test Result

We can see that the “Europe” and “USA” variables showed a positive correlation and “Caribbean” showed a negative correlation with sailboat variants. Also, since the p-value of the “USA” and “Europe” tests is more significant than our significance level of 0.05, we reject the null hypothesis, which says that the two variables do not correlate with each other. Those two regional effects are **statistically significant** but **practically insignificant**. For the “Caribbean” regional effect, it appears as **insignificant** generally.

### 4.2.2 Consistency: PDP Method

Since our model is not a linear regression model, we could not use a traditional hypothesis test to test the model’s practical or statistical significance. To do this, we will fit separate models for each sailboat variant and compare the effects for three regional dummy variables using PDP to generate 2D plots:

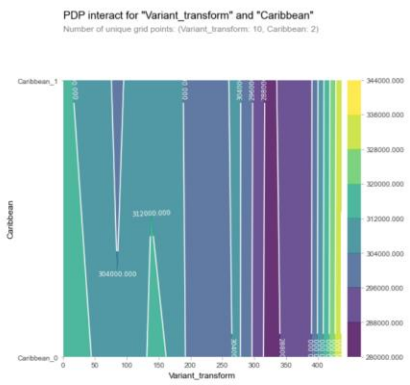



Figure 20: PDP of Caribbean

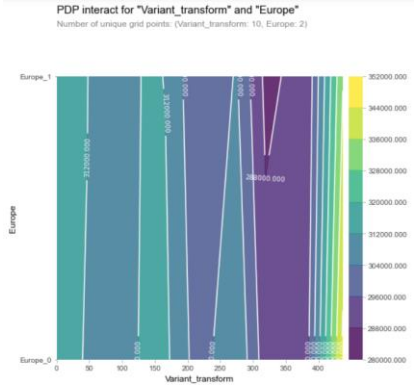



Figure 21: PDP of Europe

The white lines separate the graph by sailboat variants, each of which represents a level of the listing price. The greater the slope, the more inconsistent the region effect has been among all kinds of sailboat variants. In Figure 20, we can see that almost all white lines are vertical except five of them, which implies that the "Caribbean" region effect can be considered **consistent**. The "Europe" region effect appears more **inconsistent** than the "Caribbean" in Figure 21.

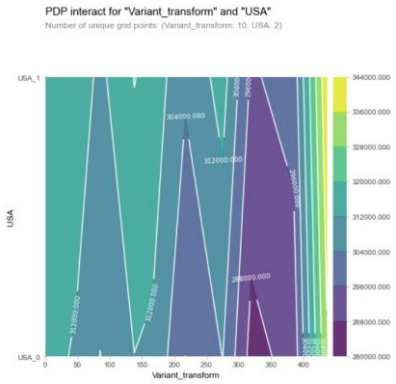



Figure 22: PDP of USA

For the USA-Variant plot, the white lines show apparent slopes far away from vertical lines. So we regard the "USA" region effect as **strongly inconsistent** with sailboat variants.

# 5. Task 3: Analyzing Regional Effect on the Hong Kong Market

## 5.1 Hong Kong Data Scratching

We collected data about Hong Kong used sailboats listing prices from the websites https://www.yachtworld.com/boats-for-sale/region-asia/country-hong-kong/category-sail/ and https://www.boats.com/boats-for-sale/?country=hong-kong&class=sail by searching relevant tags. We got 89 rows of data with detailed information on sailboats. A preview of collected data is shown:

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | type | Hull_Material | fuel_type | length | Country/Region/State | Make | Variant | Year | Listing Price | GDP |
| 2 | Catamaran | Fiberglass | Diesel | 42.17ft | Hong Kong | Bali | 4.2 | 2023 | 631,728 | 52132.076 |
| 3 | Racer/Crui | Composite | Diesel | 54.99ft | Hong Kong | X-Yachts | X-55 | 2009 | 380,362 | 30593.991 |
| 4 | Catamaran | Fiberglass | Diesel | 45.73ft | Hong Kong | Fountaine Pajot | Tanna 47 | 2023 | 1,118,056 | 52132.076 |
| 5 | Catamaran | Fiberglass | Diesel | 61.58ft | Hong Kong | Fountaine Pajot | Samana 59 | 2023 | 3,190,536 | 52132.076 |
| 6 | Racer | Composite | Diesel | 31.17ft | Hong Kong | FarEast | 31R carbon edition | 2021 | 140,000 | 49865.353 |
| 7 | Racer | Composite | Diesel | 52ft | Hong Kong | Judel and Vrolijk | Tp52 | 2006 | 299,000 | 28028.157 |

Figure 23: Input data of Hong Kong market

## 5.2 Data Preprocessing

We merge those data from Hong Kong with the given spreadsheet and apply similar methods for processing data as 3.3. The “Geographical region” of newly collected data is set as “Hong Kong.”

## 5.3 Result of Regional Effect of Hong Kong

We performed model training using our merged data and got prediction results using the model obtained from task 1. The performance of regression has been improved after adding Hong Kong data to the original data set. The new R square generated is 0.8779, surprisingly. So we also find the confidence interval of this model. However, the result of CI is not as good as the R-square since the precision is only 0.32.

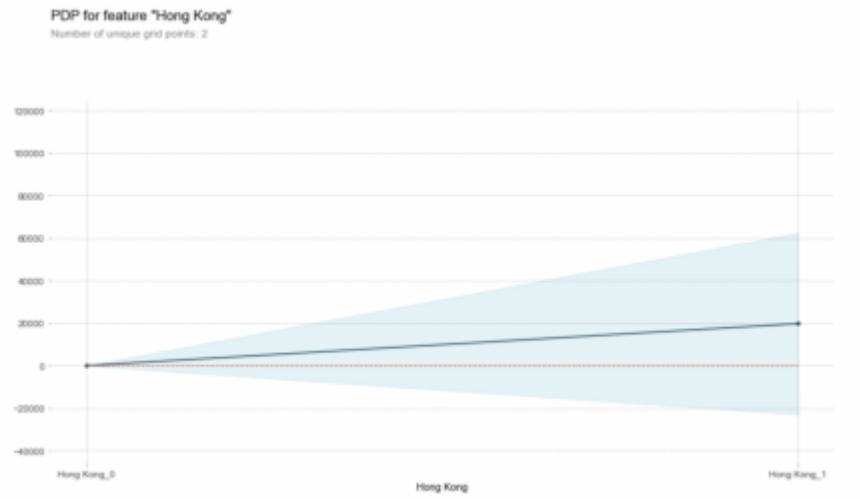



Figure 24: PDP of Hong Kong

We test for regional effects of Hong Kong using PDP from Figure 24 shown above, which illustrates that Hong Kong has a **positive regional effect** on the listing prices of sailboats.

## 5.4 Comparison Between Catamaran and Monohull Sailboats

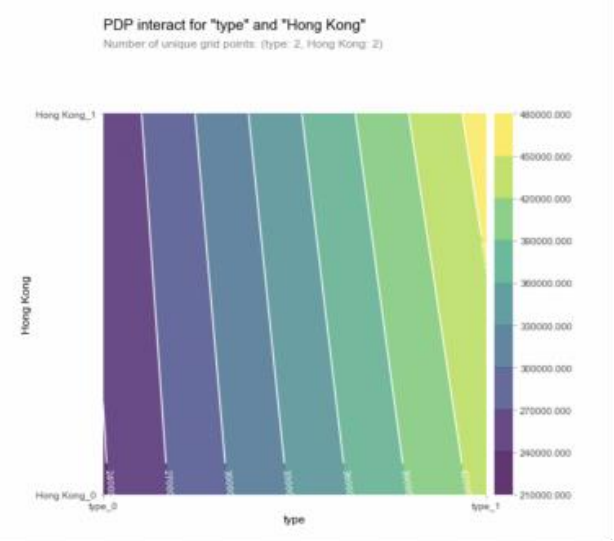



Figure 25: PDP interact of Hong Kong and type

Similar to the previous 2-D PDP plot in 4.2.2, we use PDP again and get the plot with respect to two different types: Catamarans with "type" = 1 and "Monohull" with " type" = 0. Figure 25 shows that the regional effect of Hong Kong varies in two types of sailboats, so they are **not the same**.

## 6. Conclusion

In this report, we develop a mathematical model that explains the listing price of each sailboat in the provided spreadsheet. We collect economic data from the website, including the GDP and Hong Kong sailboats data sets, to extend our data. We then use ensemble learning to combine a random forest regression model with a gradient-boosting regression model to train our data and make predictions. We also examine the effect of the geographic region on listing prices and determine whether any regional product is consistent across all sailboat variants. Sailboat brokers and enthusiasts can use our analysis to better understand the Hong Kong sailboat market and make informed decisions when buying or selling sailboats.

## 7. Discussion

• During the data cleaning, we find that the number element in the "Variant" columns has some relationship with its next column, "Length (ft)," which represents the length of the boat in feet. Some examples like

| Variant | Length (ft) |
|---|---|
| Ovni 395 | 41 |
| 42 Match | 41 |
| 4.1 | 41 |

Table 6: Sample of three types of Variant corresponded with Length

We can see that for a ten-digit number length, number elements in 'Variant' can vary from a hundred-digit number to a single-digit number. However, by multiplying 10 or being divided by 10 and rounding it off, we find that it is quite close to the number in Length. The biggest difference between our processed number in Variant and Length is 3. This may provide the viewers with a convenient way to get information about a sailboat's length without searching for details.

• During the feature selection process, we check the heat map and found that the correlation between the "Type" and "Caribbean" variables is 0.26, which implies that one type of sailboat among those two counts is significantly higher than the other. This corresponds with the real-world sailboat market, which is quite an interesting discovery and may be useful for enthusiasts of sailboats.

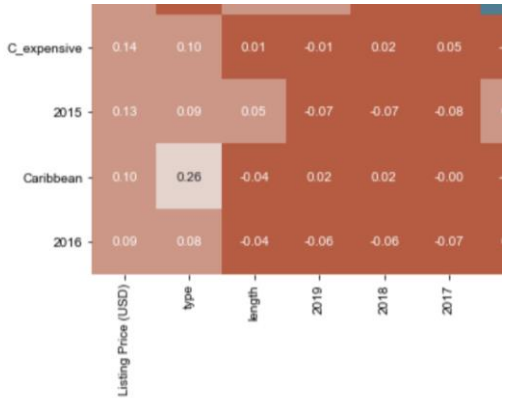



Figure 26: Local picture of heatmap

• During the feature selection process, we find that the feature importance of some years is much larger than those of others. For example, the dummy variable "2018" has a value of 0.041927 in terms of importance, while "2013" has only a value of 0.000804, which differs a lot from the former one. This may be associated with some economic effects of the sailboat market. In some years, sailboat listing prices may be affected greatly by international politics and regional

economic changes. Also, the price-changing influence of the complementary objects on a certain year can also influence sailboat prices.

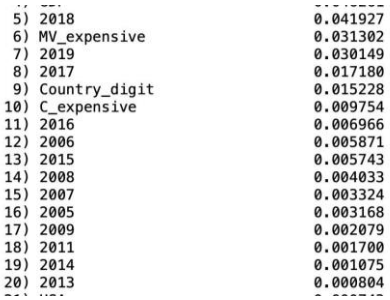



Figure 27: Local picture of feature importance

## 8. Evaluation of Models

### 8.1 Strength

- First, based on the traditional MLR model, we combine the RF and GBRT model, which is more complicated and more suitable for the pricing regression models. We improve the behavior of the R-square value and other evaluation methods.
- Second, the ensemble stacking method harnesses the capabilities of a range of well-performing models on a classification or regression task and makes predictions that perform better than any single model in the ensemble. Also, as the ease of overfitting to complicated models, the ensemble stacking method combining other models can improve the diversification in different data, such as our model has a surprisingly better performance on the Hong Kong market.
- Third, except for correlation analysis on features, we use Partial Dependence Plots (PDP) to do the analysis of causality so the explanation of our model is more readable.
- Fourth, data normalization with the special method brings a quicker convergence rate and higher precision of iterations.

### 8.2 Possible Improvements

- In terms of long-term prediction, we lack information on recycling and incinerating prices. As a result, the prediction of these listing prices completely based on current “Make”s may lead to big errors, especially when there exist new “Make”s in the market.

- The model is relatively time-consuming because our model uses a grid search algorithm to get precise parameters.
- The model in this paper is based on discrete data for the year 2005-2019, but not months or days. This data problem prevents the author from applying the time series model.

## 9. Memorandum for the Broker

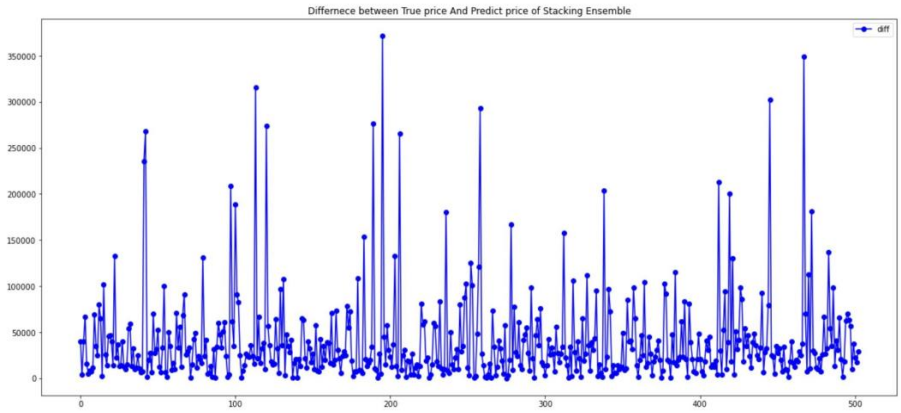
Dear broker,

As people become more enthusiastic about the sport, the Sales of sailing boats in Hong Kong are booming. We are glad to offer our pricing model and make informed suggestions for your decision. Below is the whole process of analysis and the important suggestions.

**Raise attention to perspective Features**: Through the analysis of different features of sailboats, such as variants, length, country, year, GDP, type, and regions, we found some interesting correlations. The length of the boats and the digit part of the variant are similar to each other, so when you make your decision, you can refer to the price of some other sailboats with similar lengths or the digit part of the variant.

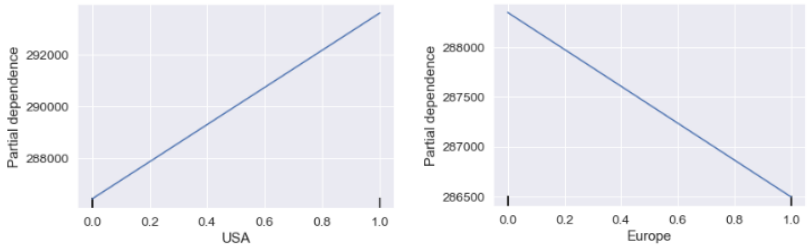
**Do a pre-processing on original data**: Some texture information, like regions and counties, is transformed into numerical values during the data pre-processing. For some data, with all features the same but the listing price different, we take the average of all these listing prices and then replace the original values. We also normalize the data with StandardScaler in python, which brings a quicker convergence rate and higher precision of iterations. This method could provide in-time predictions on listing prices which is essential in emergent decision cases.

**Apply multiple regression algorithms on ensemble learning**: You can apply random forest and gradient boosting methods and ensemble learning for their combination. All these methods are for better evaluation and prediction. The resulting listing price prediction model has an R-square precision value of 0.8779, considering the assumption of neglecting economic and political issues, which significantly explains the data.




**Make sailboat trade across regions**: To examine the effect of regions during the data pre-processing effect of regions, we apply partial dependence plots (PDP), which clearly show the marginal impact of the region. The result indicates a positive correlation between the USA and the listing price while a negative correlation for other regions. With further testing of regional effect consistency, the coefficients show invariance through all variants.

So you could buy sailboats from regions like Europe or the Caribbean and then sell them to regions like the USA. This gap in the price-region correlation brings profits.




**Set optimal price according to the balance between the over-prediction and under-prediction cost:** To balance the risk of over-prediction and under-prediction concerning their cost, we suggest you to implements the “suggestion price”s with the distribution of each variant, respectively. These prices make the possible cost for wrong prediction the least, and the same for one extra unit of 2-case prediction errors.

| | |
|---|---|
| over-prediction unit cost | $\alpha$ |
| under-prediction unit cost | $\beta$ |

$$z = \text{stats.norm.ppf}\left(\frac{\alpha}{\alpha+\beta}\right)$$

$$\text{suggestion price} = \text{mean} + z * \frac{sd}{\sqrt{n}}$$

Sincerely,

Team 2331406

## References

Bowen, Y., Buyang, C., Research on Ensemble Learning-based Housing Price Prediction Model. Big Geospatial Data and Data Science (2018) 1: 1-8.

Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.

Jun Hao, Qianqian Feng, Jiaxin Yuan, Xiaolei Sun, Jianping Li, A dynamic ensemble learning with multi-objective optimization for oil prices prediction, Resources Policy, Volume 79, 2022, 102956, ISSN 0301-4207

Kornbrot, D. (2014). Point biserial correlation. Wiley StatsRef: Statistics Reference Online.

Polikar, R. (2012). Ensemble Learning. In: Zhang, C., Ma, Y. (eds) Ensemble Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-9326-7_1

Rigatti, S. J. (2017). Random forest. Journal of Insurance Medicine, 47(1), 31-39.

Vermeulen, A.F. (2020). Supervised Learning: Advanced Algorithms. In: Industrial Machine Learning. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-5316-8_5

## Appendices

1. Code

Ensemble model

[18]:
```
## GridSearch
lr = GradientBoostingRegressor()  #0.85
sclf1_gs = StackingRegressor(regressors=[clf11, clf21],
                             meta_regressor=lr)
# params = {
#     'meta_regressor__alpha': [0.1, 1.0, 10.0],
#     }
params = {'meta_regressor__learning_rate': [0.1, 0.3, 0.5, 0.7],␣
 ↪'meta_regressor__max_features': range(1, 5),
              'meta_regressor__subsample': [0.1, 0.3, 0.5, 0.7],␣
 ↪'meta_regressor__n_estimators': range(100, 401, 100)}
# params = {'meta_regressor__learning_rate': [0.1, 0.3, 0.5, 0.7],␣
 ↪'meta_regressor__max_features': range(1, 5),
#         'meta_regressor__n_estimators': range(100, 401, 100)}
# lr = Ridge() #0.85
grid = GridSearchCV(estimator=sclf1_gs, param_grid=params, cv=5, refit=True)
grid.fit(train_x1, train_y1)
pre_y_stack1 = grid.predict(test_x1)
scores = performance_metric(test_y1,pre_y_stack1)
print("R-square: %0.8f (+/- %0.2f)"
       % (scores.mean(), scores.std()))
```