Problem Chosen	2023 MCM Summary Sheet	Team Control Number
Y		2333372

Uncovering the Hidden Patterns in Sailboat Pricing

A Comprehensive Analysis using Proposed Berted-Tabnet

Summary

In this study, we **unveil the intricate factors** that affect the listing price of a sailboat, with a special focus on regional effects. Leveraging an extensive dataset obtained from both official sources and web scraping, we ingeniously **devise a state-of-the-art Berted-TabNet** model to analyze sailboat pricing determinants.

Our meticulous data preprocessing includes the acquisition of **21 additional sailboat features** and regional economic data, followed by rigorous feature selection and data cleansing. Our innovative **Berted-TabNet model outperforms other well-known methods**, such as **XGBoost**, **LightGBM**, and **Elastic Net**, in terms of Root Mean Squared Error (RMSE) and Relative Absolute Error (RAE), showcasing its remarkable prowess in regression tasks.

To further quantify regional effects, we **ingeniously fuse self-attention mechanisms with traditional importance measures to conceive the novel Attentive Feature Importance (AFI)** method. Our results divulge the profound impact of the **Country/Region/State** variable on sailboat pricing, emphasizing the significance of regional factors.

When applied to the Hong Kong (SAR) market, our unconventional model **takes into account not only the GDP impact but also a comprehensive set of geographic indicators**, including the **Coastal Proximity Index (CPI), Infrastructure and Service Index (ISI), Cultural and Recreational Activity Index (CRAI), and Competition Index (CI).** This exhaustive analysis reveals higher sailboat prices compared to other regions, with **a notable difference** in the regional effect between Monohulls and Catamarans. This insight is particularly valuable for the Hong Kong (SAR) sailboat broker, who can leverage our findings to optimize their pricing strategy.

Additionally, we uncover the substantial influence of other factors, such as **Make**, **Time**, and **Rigging Type**, on sailboat pricing. By **integrating text and visual representations**, we **not only present our transformative research results but also provide actionable insights for industry stakeholders.**

In conclusion, our innovative work with the **Berted-TabNet** model has led to a **deeper understanding of the complex interplay between various factors** that shape sailboat pricing and potentially **transform the way we approach the sailboat market.**

Keywords: Transformer Embedding Berted- Tabnet Attentive Feature Importance(AFI) XGBoost LightGBM Elastic Net

Contents

1 Introduction

1.1 Restatement of the Problems

Figure 1: **What factors influence the market price of sailboats? How do factors specific to the Hong Kong(SAR) region affect sailboat market prices?**

Sailing yachts' value is **influenced by factors** such as age and market conditions. Wear and tear reduce value as yachts age. Demand affects values greatly, with high demand increasing prices and low demand causing decreases. Brand, model, maintenance, region, and history also impact value. This paper addresses these issues ,as outlined in Table 1.

Part	Task Overview
I	· Develop a mathematical model to predict the listing price of each of the sailboats based on given data and self-collected data. · Evaluate the precision of our model's estimate.
II	· Utilize our model to explicate the potential influence of region on the listing prices. · Contemplate whether the impact of some of the regions is uniform across all sailboat variants. · Explain the practical and statistical significance of any regional effects observed.
III	· Analyze how to use the regional effect of our model to research the Hong Kong (SAR) market. · Build a model to analyze the regional effect of Hong Kong, using corresponding listing price data from the Hong Kong (SAR) market for a subset of sailboats from the given data. · Analyze whether the effect is common to both types of sailboats.
IV	· Discuss other useful inferences our team in addition to the conclusions already studied.
V	· Write a report to interpret the conclusions of this paper for the Hong Kong (SAR) sailboat broker.

Table 1: Problem Restatement

1.2 Our work

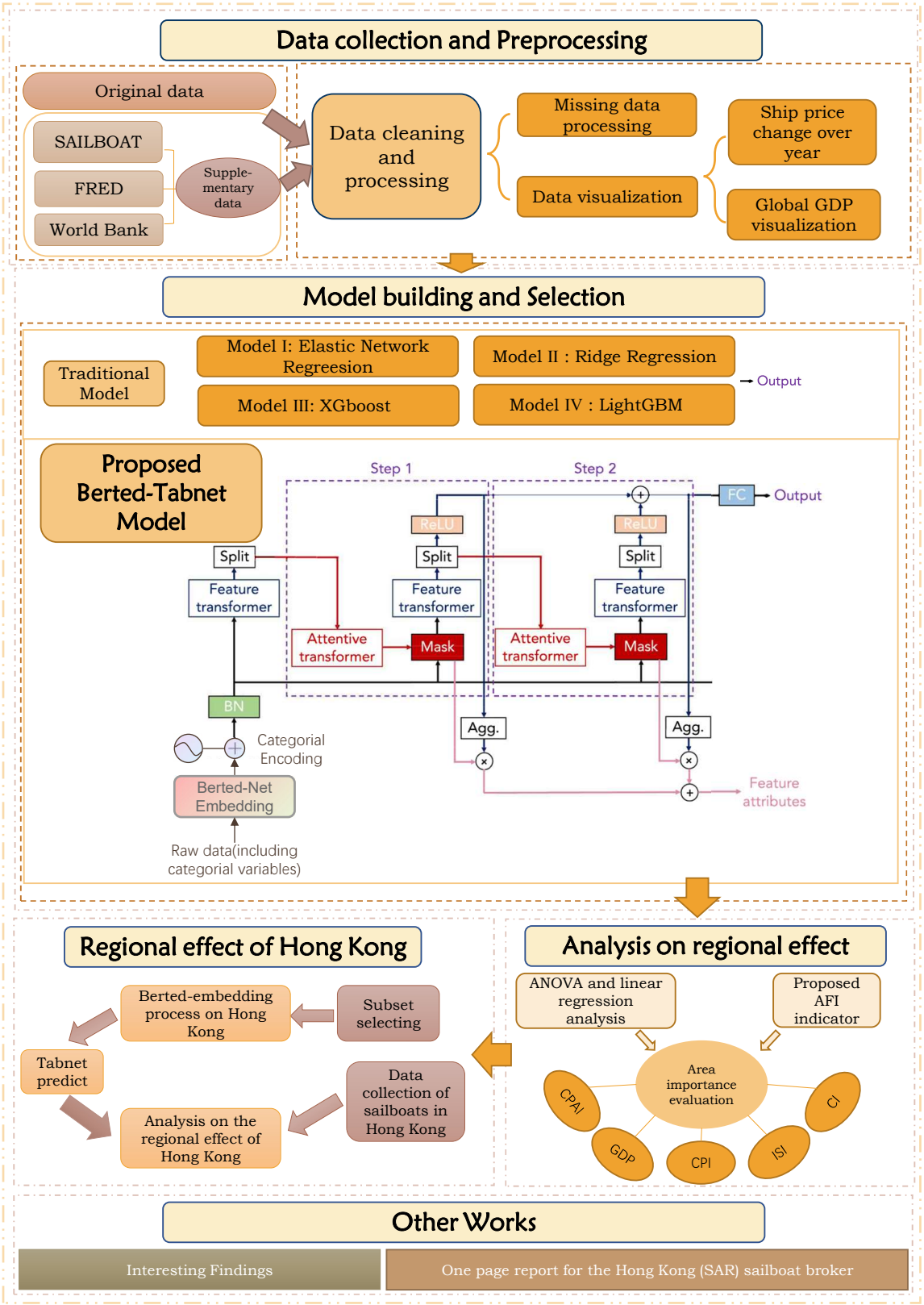


Figure 2: Work Summary

2 Data Preparation

2.1 Assumptions and Limitations

1. **The data provided is representative of the sailboat markets in Europe, the Caribbean, and the USA**.The dataset contains sailboat data for many regions, while some regions have less data. This assumption can help us better analyze the impact of different regions on the forecast results.

2. **The influence of different seasons on the listing price of ships is consistent**. Fluctuations in listing prices may be affected by seasonal factors. For example, during the peak tourist season, the demand for activities such as wading entertainment will increase, which may have an impact on the listing price, but our data only has year data. This assumption can help us ignore the influence of seasonal factors on the listing price.

2.2 Notations and Terminology

The primary notations used in this paper are listed in Table 2.

Variable	Meaning
Beam	Width of a boat at its widest point
Catamarans	Watercraft with two parallel hulls
Displacement	Weight of water displaced by boat
Draft	Minimum depth to float boat
Engine Hours	Running time of boat engine(s)
Headroom	Height to stand in cabin
Hull	Main body of a ship or vessel
Hull Materials	Boat hull materials (e.g. fiberglass)
Listing Price	Seller's asking price
Make	Sailboat manufacturer
Monohull Sailboats	Sailboats with one hull
Rigging	Controls sails and steers with ropes, cables, pulleys.
Sail Area	Total surface area of a boat's sails when fully raised
Variant	Sailboat model name (e.g. "Sun Odyssey 54 DS")
GDP	Total value produced within a country's borders one year
GDP per capita	Per capita GDP

Table 2: Variables in Sailboat Problem

Sailboat performance evaluation involves multiple indicators, such as the Sail Area to Displacement Ratio (SA/Disp.) for potential speed in different wind conditions, Ballast to Displacement Ratio (Bal./Disp.) for stability, Comfort Ratio for motion comfort, Capsize Screening Formula (CSF) for seaworthiness, Bruce Number (BN) for light-air performance, Kaiser Performance Number (KSP) for potential speed based on various factors, and Displacement to Length Ratio (Disp./Len.) for assessing the balance between comfort and speed capabilities.

The calculation formulas and visualizations for these indicators are shown in the following Figure3.

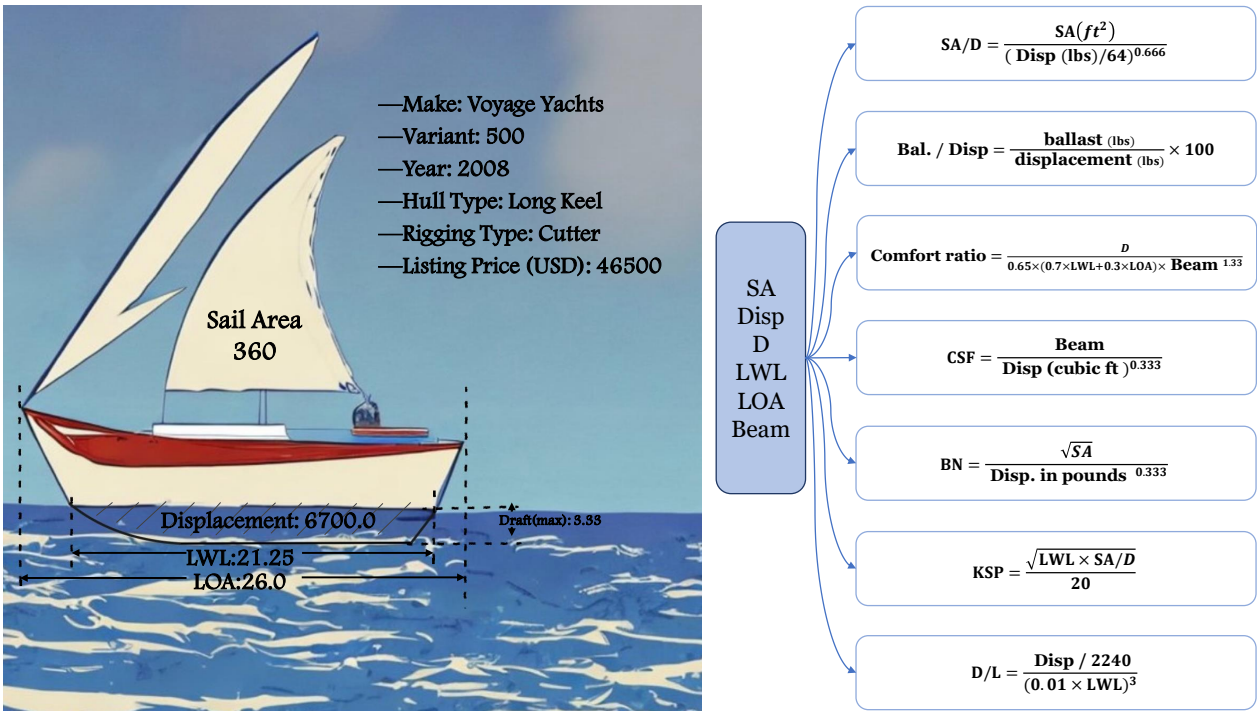
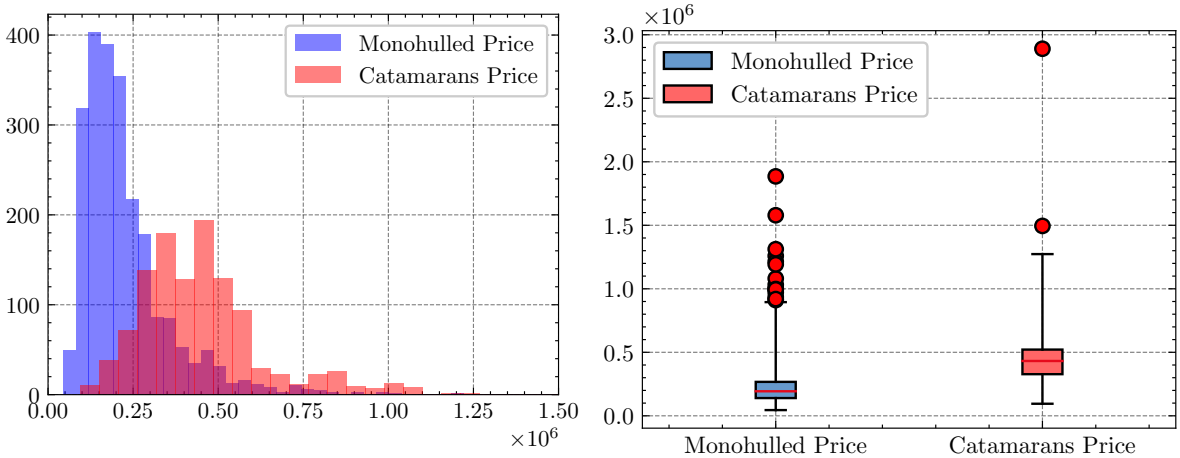


Figure 3: Sailboat Performance Indicator Visualizations

2.3 Data Exploration and Visualization

We perform an analysis on the tabular data by creating histograms and box plots for the listing prices of two types of boats: Monohulled Sailboats and Catamarans. They are depicted in Figure



(a) Histograms of Boat Prices

(b) Boxplots of Boat Prices

Figure 4: Histograms and Boxplots of Listing Prices

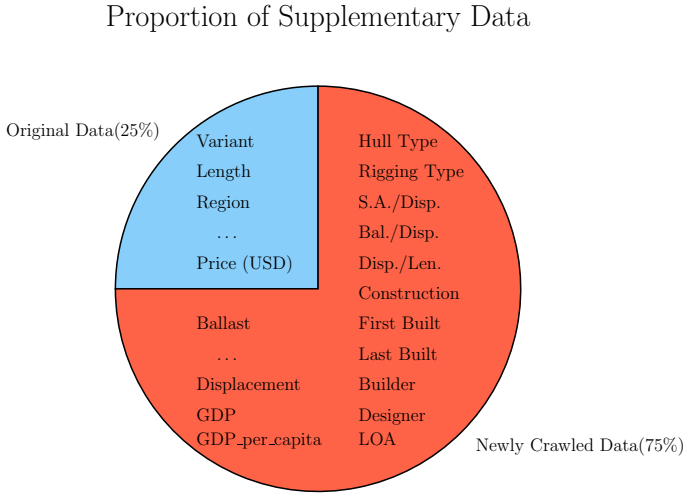
2.4 Data Acquisition and Collection

In our study, we encounter **a severe lack of data variables** in the original dataset. The dataset contains only two numeric arguments, Length (ft) and Year, while the majority of the independent variables including Make, Variant, Geographic Region and Country/Region/State are categorical. This poses a significant challenge for our regression. To better explain the listing price of each sailboat, we need to utilize other resources to expand and collect additional data.

2.4.1 Supplementary Data Scraping for Sailboats and Regional Economic Data

We utilize **BeautifulSoup to parse the HTML content** of sailboat websites, extracting detailed specifications for each unique sailboat. Concurrently, we employ Python packages such as pandas, requests, tqdm, pycountry, and APIs like **Federal Reserve Economic Data (FRED)** API and **World Bank** API to obtain regional economic data, including state-level population and GDP data. We match country names using **ISO 3166-1 alpha-3** codes and handle missing data with 'None' values as necessary.We merged the supplementary data with the original dataset saved it as a new Excel file.

This enhanced dataset provides a more comprehensive and informative foundation for our analysis, allowing us to capture the impact of both sailboat specifications and regional economic factors on listing prices.The following Figure6 presents an example of per capita GDP maps for various regions in 2021, obtained through web scraping.



Data Type	Source	Website
Sailboat data	Sailboat Website	https://sailboatdata.com/
US population and GDP	FRED	https://fred.stlouisfed.org/
Global population and GDP	World Bank API	https://www.worldbank.org/en/home

2.5 Feature Reduction and Data Cleaning

Features of various types of boats are collected from different websites. However, there are lots of **missing values** in these data. Therefore, we need to carefully select relevant features for analysis, taking into account the degree and patterns of missing data. Additionally, we need to remove data from boats that have insufficient data quality. The

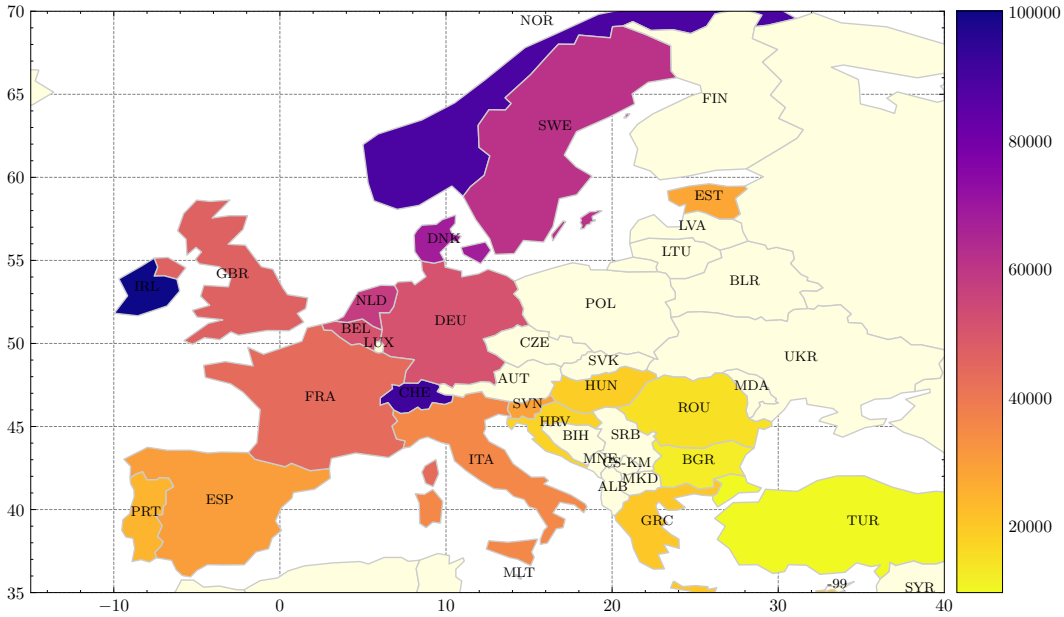


Figure 5: Example of GDP per Capita in Europe for 2021

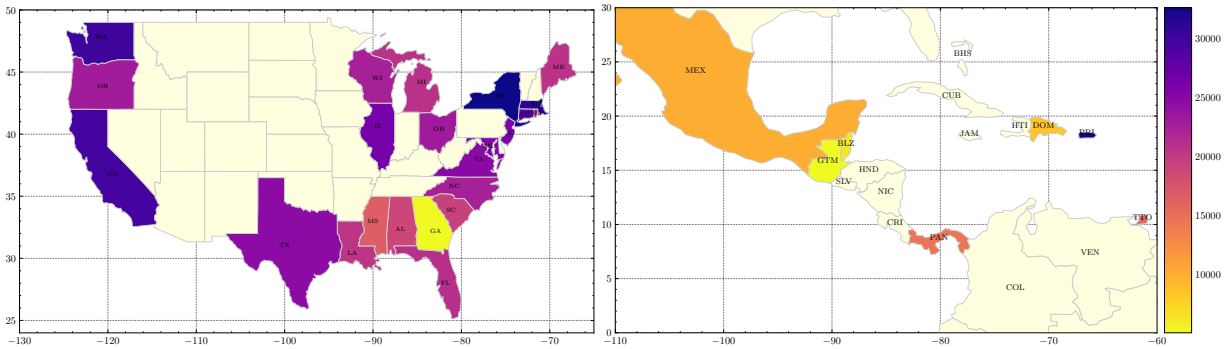


Figure 6: Example of GDP per Capita in continental US, Europe and Caribbean region for 2021

data is characterized by various degrees of missingness, which we also take into account. We list features with more than 10 percent of missing values which are shown in Table 3. It is not difficult to find that the features listed in Monohulled Sailboats are similar to those in the Catamarans, which reflects that these features may be difficult to obtain or not important to most consumers.

After **deleting these features** and removing data from catamarans, a small part of the data is still missing, and we eliminate this part of the monohulled sailboats data. After deleting this part of the data, it is not difficult to find that the increase in the number of features is accompanied by the decrease in the amount of data. But the reduction of this part of the data has little impact on the construction of the model.

After this operation, we have data with a sufficient number of features and a large enough amount of data for subsequent model building on monohulled sailboats.

Hull Type	Variable	Missing Values	Proportion(%)
Monohulled Sailboats	Built	916	39.04
	Draft (min)	916	39.04
	Last Built	814	34.70
	Ballast Type	419	17.86
	Bal./Disp.	264	11.25
	Ballast	264	11.25

Table 3: Statistics of missing values in our collected data for Monohulled Sailboats

3 The Models

3.1 Sailboat Listing Price Model Construction

3.1.1 Comparison and Evaluation of Traditional Regression Methods and Encoding Techniques

In this section, we briefly discuss the limitations of traditional regression methods and encoding techniques for categorical variables. Traditional regression methods such as Elastic Net and Ridge Regression offer good feature selection and can handle high-dimensional data but may not provide the highest accuracy due to their reliance on tuning hyperparameters. Methods like XGBoost and LightGBM can achieve high accuracy, but they are prone to overfitting and require substantial memory consumption.

Considering these limitations, we propose a novel approach using the Berted-Tabnet architecture to predict sailboat listing prices.

3.1.2 Proposed Berted-Tabnet Method for Sailboat Listing Price Prediction

The proposed Berted-Tabnet architecture aims to address the shortcomings of conventional regression methods and encoding techniques. This method combines the power of the BERT embedding layer for efficient encoding of categorical variables and the TabNet deep learning model for accurate price prediction.

Bert Embedding Layer **BERT (Bidirectional Encoder Representations from Transformers)** is a powerful natural language processing model that has demonstrated exceptional performance in various tasks. One of its key features is **the embedding layer, which transforms input tokens into dense vector representations**. This layer enables the model to capture both semantic and syntactic information, making it highly effective for processing complex text data. In our proposed method, we utilize the BERT model's embedding method to transform categorical variables in the data table after web scraping and supplementing the data. The dimensions were increased to 100, making the original data of different categories linearly separable in high-dimensional space. The utilization rate of the vector space $\mathcal{S}$ was significantly improved. The drawbacks of The James-Stein Encoder and leave-one-out encoding methods, which include reliance on training set labels

and high computational expense, have been avoided.This approach offers several advantages, such as reducing the risk of overfitting, providing a more efficient representation of categorical variables, and eliminating the reliance on training set labels for encoding. By leveraging the power of transformer-based models, we can achieve better performance and more accurate predictions in our regression analysis. Figures 7 and 8 illustrate the comparison of one-hot encoding and embedding visualizations.

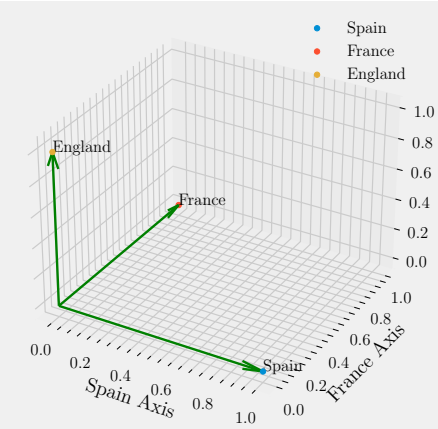


Figure 7: **One-hot encoding** visualization, where we can observe that the words Spain, France, and England are **mutually orthogonal**, while the vast vector space remains **underutilized**

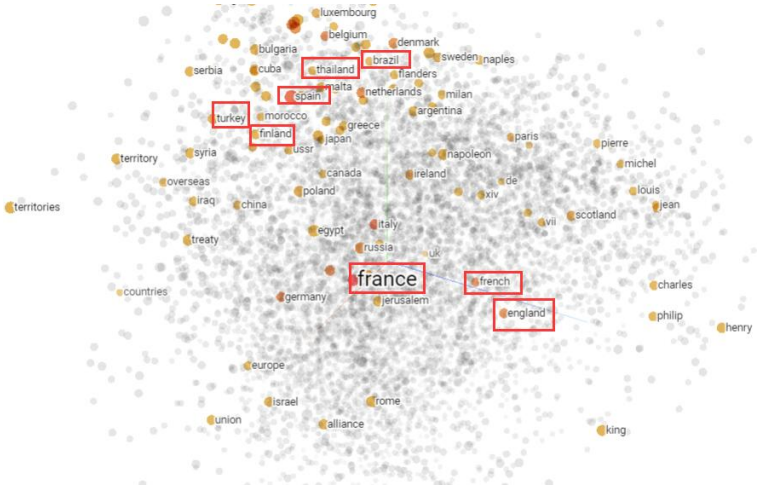


Figure 8: **Embedding** visualization, where categorical variables (using region names as an example) **densely populate the entire vector space**, and countries with greater similarity are closer together. This allows the model's training data to **better reflect true characteristics**, rather than treating all countries equally

Figure 9: Comparison of one-hot encoding and embedding visualizations

Tabnet Self-Attention Part **TabNet is a novel deep learning architecture** that leverages the principles of the Transformer model to **address the challenges of handling both categorical and continuous variables in structured data**. By utilizing **self-attention mechanisms and feature selection**, TabNet is capable of automatically learning meaningful and interpretable patterns from a given dataset, alleviating the need for extensive feature engineering or dimensionality reduction techniques. Additionally, TabNet demonstrates impressive performance in terms of accuracy and interpretability when compared to other deep learning methods designed for tabular data.

The TabNet architecture in our proposed Berted-Tabnet method is composed of two main components: the Feature Transformer and the Attentive Transformer. The Feature Transformer is responsible for learning representations of the input features, while the Attentive Transformer is used for selecting the most relevant features in each decision step. This combination allows TabNet to focus on the most important variables, leading to improved model performance and interpretability. Figure 2 illustrate the TabNet encoder and decoder architecture, respectively. Additionally, Figures 10 showcase the Feature Transformer and Attentive Transformer components.

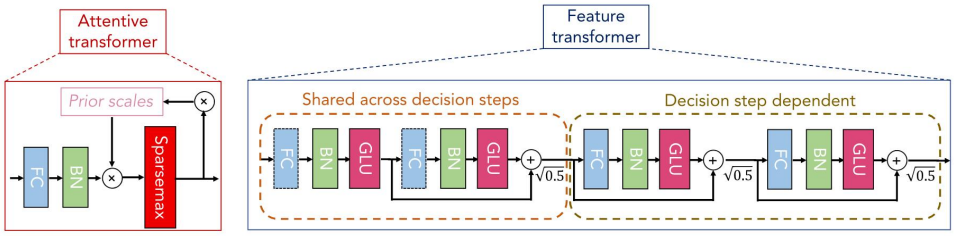


Figure 10: Feature-transformer and Attentive-transformer:The 4-layer network includes feature and attentive transformer blocks, with shared and step-dependent layers. Each layer has a fully-connected layer, batch normalization, and gated linear unit nonlinearity. The attentive transformer uses prior scale information and sparsemax normalization for sparse selection of salient features.

Algorithm 1: Proposed Berted-Tabnet Algorithm

Data: Categorical and continuous features as input
Result: Final output y
Step 1: Apply BERT Embedding
  Process categorical features **using BERT Embedding**, obtaining 100-dimensional vectors
  Combine embedded categorical features with continuous features as input x_0
  Initialize previous split decision d_{-1}
Step 2: TabNet Algorithm
for $k = 0$ *to* $K - 1$ **do**
    Compute masked features: $x_k = x_0 \cdot (1 - M_k)$
    Compute feature transformer: $z_k = \text{FeatureTransformer}(x_k)$
    Compute attentive transformer: $M_{k+1} = \text{AttentiveTransformer}(z_k, d_{k-1})$
    Update split decision: $d_k = d_{k-1} + M_{k+1}$
Compute the final output: $y = \text{Decoder}(z_K)$

3.1.3 Model Performance Analysis and Comparison

In order to evaluate the performance of our proposed Berted-Tabnet method for predicting sailboat prices, we employed the following metrics to assess the results.

RMSE and RAE The Root Mean Square Error (RMSE) is a widely used metric to measure the accuracy of a regression model by calculating the square root of the mean squared differences between the predicted and actual values. The Relative Absolute Error (RAE) is another metric used to assess the performance of regression models. It calculates the relative difference between the predicted and actual values and is useful in comparing the model's performance across different scales of data. The mathematical representation of RMSE and RAE is given by:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \text{RAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}i|}{\sum i = 1^n |y_i - \bar{y}|} \tag{1}$$

where y_i represents the actual sailboat price, $\hat{y}_i$ denotes the predicted sailboat price,

and n is the total number of observations. A lower RMSE value indicates better model performance in predicting sailboat prices.

The final results of our model are shown in Table 4. Due to space constraints, we only present the results for monohull sailboats, while the results for catamaran sailboats are similar. For detailed results data including catamaran sailboats, please refer to the following link: Complete Data Link.

Sailboat HullType	Method	Parameters						
Monohulled Sailboats	Berted-Tabnet	n_Step	Gamma	Momentum	Lambda Sparse	n_Shared	MSE(10^8)	RAE
		3	1.3	0.01	0.001	1	13.0350	0.1952
		3	1.3	0.02	0.005	1	14.0702	0.1981
		3	1.5	0.01	0.005	2	13.0751	0.1931
		3	1.5	0.02	0.005	2	6.9224	0.0809
		5	1.3	0.01	0.001	1	6.2034	0.0800
		5	1.3	0.02	0.001	1	6.7495	0.0805
		5	1.5	0.01	0.001	2	3.5583	**0.0441**
		5	1.5	0.02	0.005	2	4.7365	0.0530
	XGBoost	Colsample bytree	Learning Rate	Depth	n_Estimators	Subsample	MSE(10^8)	RAE
		0.7	0.1	5	300	0.9	9.2345	0.1833
		0.7	0.1	7	200	0.8	10.4373	0.1872
		0.7	0.05	7	300	0.8	9.9541	0.1839
		0.7	0.05	7	300	0.9	5.9049	0.0920
		0.8	0.1	5	200	0.8	6.923	0.0947
		0.8	0.1	5	300	0.8	5.6667	0.0893
		0.8	0.05	7	200	0.8	6.9042	0.0931
		0.8	0.05	7	300	0.8	5.6424	0.0879
	LightGBM	Depth	Learning Rate	Subsample	Colsample Bytree	n_Estimators	MSE(10^8)	RAE
		5	0.2	0.7	0.7	100	14.1137	0.2178
		5	0.2	0.8	0.7	100	14.1137	0.2178
		7	0.2	0.8	0.8	100	34.6816	0.4504
		7	0.2	0.7	0.8	200	34.6816	0.4504
		5	0.1	0.8	0.7	200	10.6051	0.1912
		5	0.1	0.9	0.7	100	10.7208	0.1533
		5	0.1	0.9	0.7	200	10.7208	0.1533
	Elastic Net	Alpha	L1_Ratio		Tolrance		MSE(10^8)	RAE
		0.1	0.3		1×10^{-3}		52.5658	0.5968
		0.1	0.5		1×10^{-3}		51.8409	0.5930
		0.1	0.3		1×10^{-4}		50.2782	0.5840
		0.5	0.5		1×10^{-3}		54.3055	0.6017
		0.5	0.3		1×10^{-4}		54.0667	0.6021
		0.5	0.5		1×10^{-4}		53.6189	0.6012

Table 4: **Evaluation and comparison** of the predictive performance of various models for the price regression of monohulled sailboats. As shown in the figure, our proposed **Berted-Tabnet** method demonstrates a significant advantage, **even surpassing the current state-of-the-art XGBoost model**. The Berted-Tabnet method achieves an **RAE (Relative Absolute Error) metric of 0.0530 and the lowest global MSE (Mean Squared Error).**

Evaluating the Accuracy of Berted-Tabnet for Sailboat Price Estimation As shown in Table 4, our proposed Berted-Tabnet model achieves its best performance with the optimal parameter configuration, yielding a Mean Squared Error (MSE) of 4.73×10^8. This corresponds to a Root Mean Squared Error (RMSE) of approximately 21749.3, which implies that our model's average deviation for sailboat price prediction is only **21749.3 USD**. Additionally, the model achieves a Relative Absolute Error (RAE) of 0.0441, indicating that the average percentage error in price prediction is **merely 4.41%**. This demonstrates that our Berted-Tabnet model can provide highly accurate sailboat price estimations.

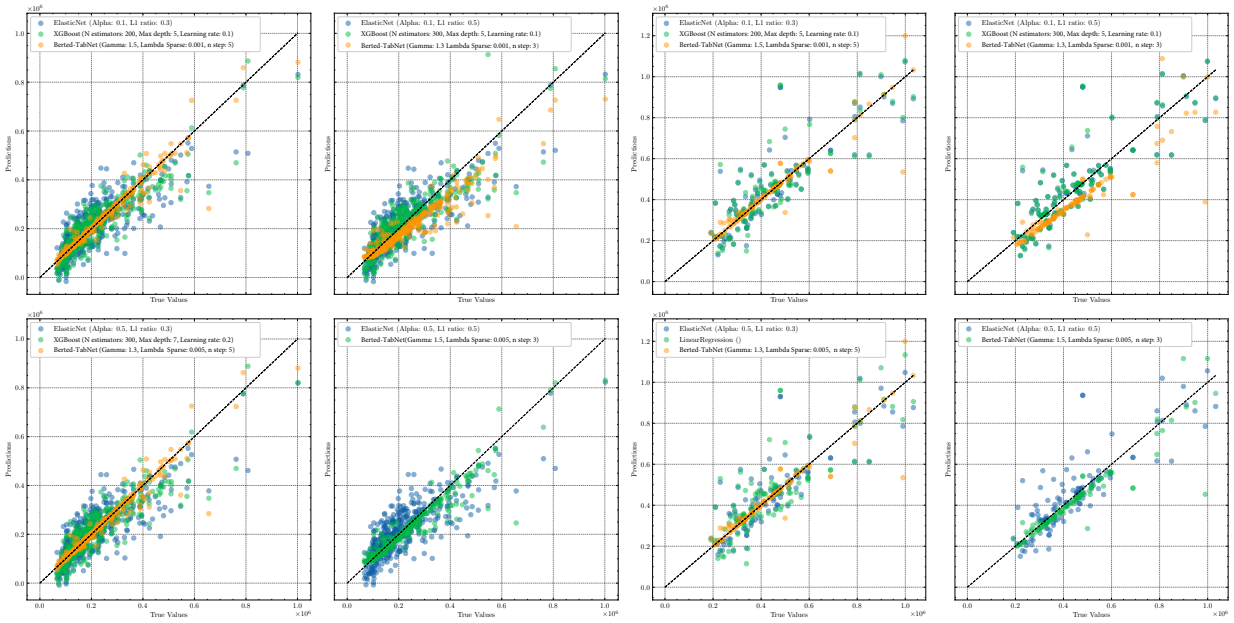


Figure 11: Scatter plots of predicted versus true sailboat prices for various models:The scatter plots show the performance of the proposed Berted-Tabnet method compared to traditional regression methods, such as Ridge regression, XGBoost, and Elastic Net regression. It can be observed that **the bubbles in the scatter plots, representing the predicted and true values for both monohulled and catamaran sailboats, are significantly closer to the $y = x$ line for the Berted-Tabnet method compared to the other methods.** This indicates that the **Berted-Tabnet method provides more accurate predictions**, as the bubbles for other methods are more dispersed.

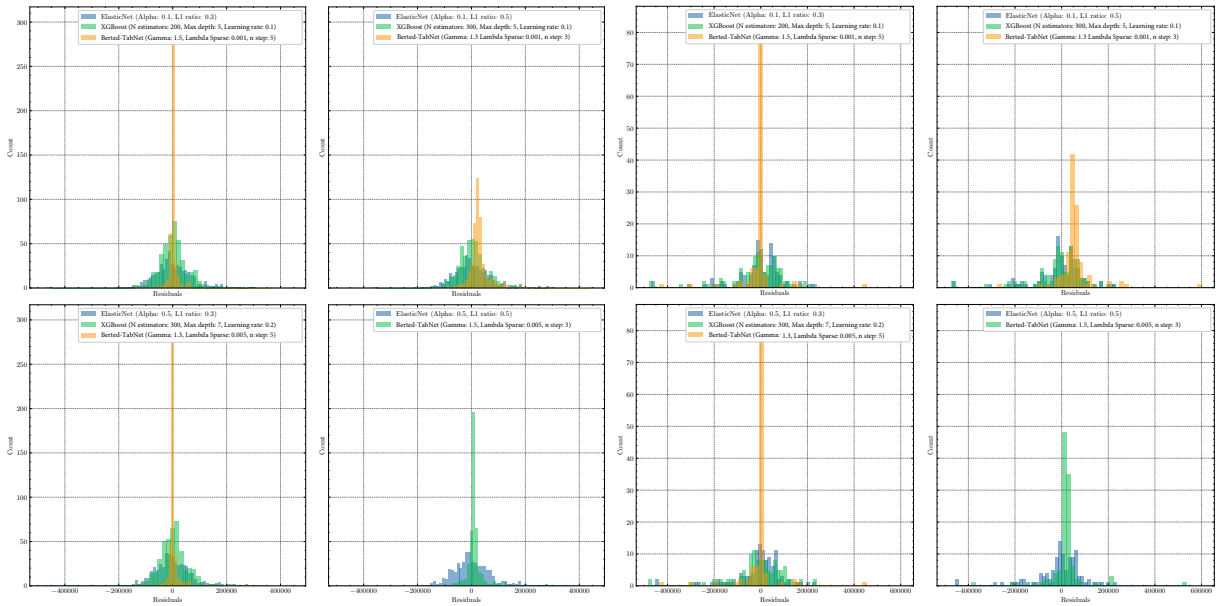


Figure 12: **Residual Distribution Plots of Berted-Tabnet versus Traditional Regression Methods for Monohulled and Catamaran Sailboats**, indicating a better fitting result of Berted-Tabnet compared to the traditional methods.

3.2 Regional Effects on Sailboat Listing Prices

3.2.1 Regional Sailboat Prices and Global GDP Distribution: Visual Comparison

To gain a better understanding of the data, we start by visually analyzing the processed data after enhancing and cleaning the supplemented data. The visualizations display the average prices of boats in different countries and regions, and it becomes apparent that there is a significant variation in average prices across different regions, particularly in Europe.

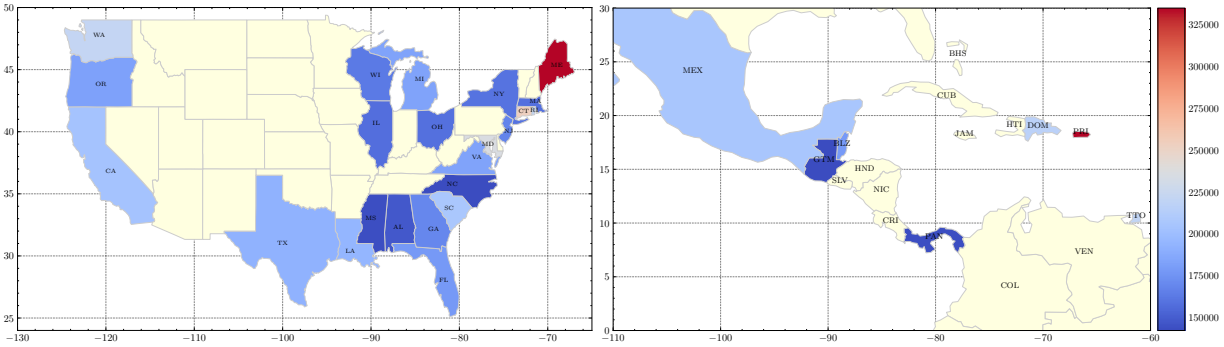


Figure 13: Average Sailboat Price in continental US and Caribbean region

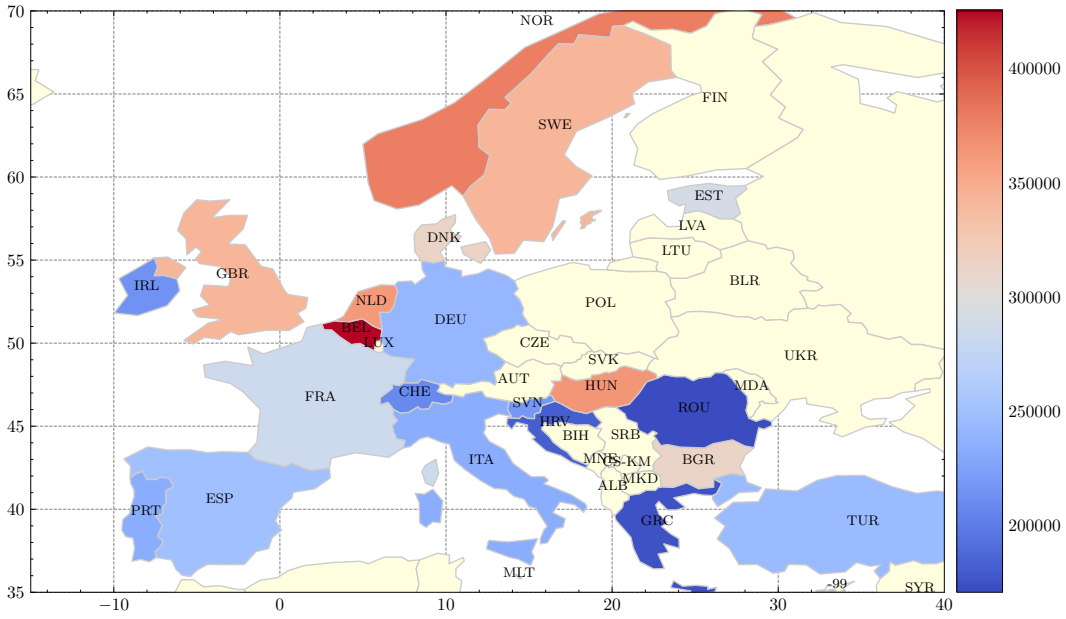


Figure 14: Average Sailboat Price in Europe

Based on the comparison between Figure 13,14 and Figure 6 ,it can be observed that sailboat prices are generally higher in countries with higher per capita GDP. These two factors show a significant correlation. By analyzing the visualized results, we can tentatively conclude that region has a consistent positive effect on sailboat prices across all sailboat variants. Furthermore, our visualizations average the data across all sailboat variants, and therefore **the observed effect is consistent across all sailboat variants**.

Our next step is to investigate the impact of the regional factor on the Berted-TabNet model.

3.2.2 Quantitative Evidence for Consistent Regional Effects

To quantitatively evaluate the consistency of regional effects across sailboat variants, we propose a novel feature importance indicator, named *Attentive Feature Importance* (AFI). This metric combines the self-attention mechanism with traditional importance measures to better capture the interaction between different features.

The methodology for calculating AFI is described in Algorithm 2. By applying a non-linear transformation to the input feature vector x and computing key, query, and value matrices, we can obtain attention weights and output feature matrix O. AFI scores are then calculated for each feature, providing insight into the importance of that feature in determining sailboat prices.

Algorithm 2: Attentive Feature Importance (AFI) Calculation

Data: Input feature vector $x \in \mathbb{R}^d$
Result: AFI scores for each feature
Apply non-linear transformation: $v = \phi(x) \in \mathbb{R}^d$
Compute key, query, and value matrices:
    $K = vW_K$
    $Q = vW_Q$
    $V = vW_V$
Calculate attention scores: $S = \frac{QK^T}{\sqrt{d}}$
Obtain attention weights: $A = \text{softmax}(S)$
Compute output feature matrix: $O = AV$
for *each feature i* **do**
    if *is_categorical(i)* **then**
        $\text{AFI}_i = \frac{\sum_{k=1}^{100} \sum_{j=1}^{d} O_{(i-1)\times 100+k,j}}{\sum_{i=1}^{d} \sum_{j=1}^{d} O_{ij}}$
    else
        $\text{AFI}_i = \frac{\sum_{j=1}^{d} O_{ij}}{\sum_{i=1}^{d} \sum_{j=1}^{d} O_{ij}}$
return *AFI scores*

As shown in Figure 15, our AFI analysis reveals the significant influence of the **Country/Region/State** variable on both Monohulled Sailboats (AFI score: 850) and Catamarans (AFI score: 842), ranking fifth among all variables. This finding highlights the importance of considering regional effects when purchasing or selling sailboats and suggests that consumers and sellers can optimize their decisions based on geographical locations.

The consistency of regional effects across sailboat variants has practical implications for consumers, who can consider regional price differences when purchasing sailboats, and for sellers, who can adjust prices or change selling locations based on different ge-

ographical locations. The AFI scores demonstrate that any regional effect is consistent across all sailboat variants.

3.2.3 Practical and Statistical Significance of Regional Effects

Practical Significance and Impact of Regional Effects on Sailboat Prices The practical significance of regional effects on sailboat listing prices benefits **market participants, such as buyers, sellers, and manufacturers**, in making informed decisions and developing targeted strategies. **Recognizing regional price differences allows buyers to find better deals, sellers to optimize pricing strategies, and manufacturers to adjust production and marketing approaches** according to regional market conditions.

Understanding regional effects on sailboat prices **enables companies to tailor strategies for maximizing profits**. They can offer competitive pricing in regions with lower GDP per capita and cater to specific consumer needs. Additionally, **companies can develop targeted marketing campaigns that resonate with regional preferences and address each market's unique characteristics.**

Sellers may benefit from **identifying locations with higher sailboat prices**, while **buyers** can **look for regions with lower listing prices.** This awareness of regional price differences can **lead to a more efficient sailboat market**, as participants adjust their strategies accordingly. **Policymakers may also need to consider regional effects when designing regulations or incentives** to ensure fairness and competitiveness in the market.

In summary, the practical significance of regional effects on sailboat prices not only informs buyers and sellers but also helps companies make strategic decisions, tailor market-

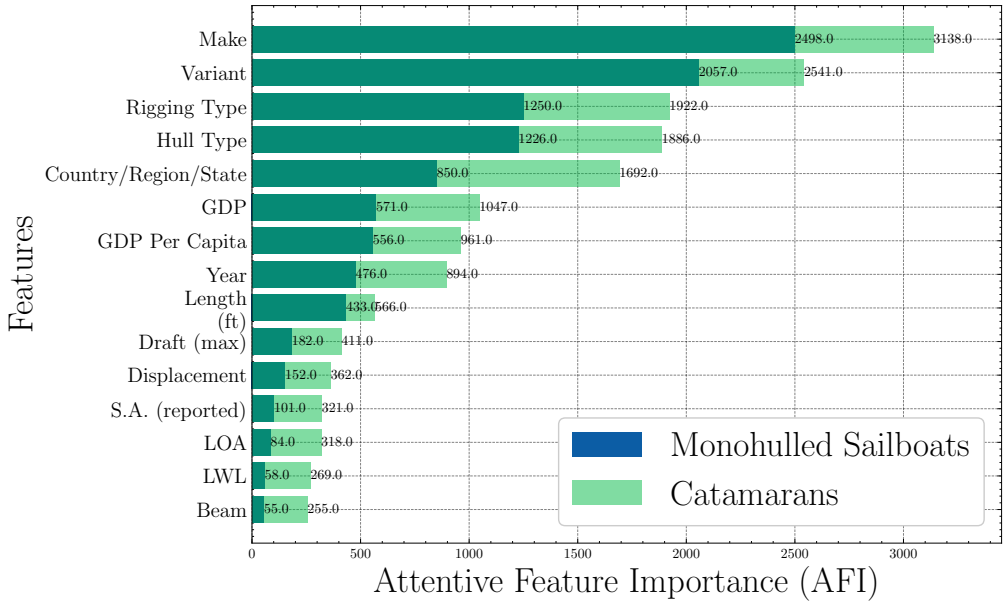


Figure 15: Attentive Feature Importance. The AFI scores reveal the significant influence of the **Country/Region/State** variable on both Monohulled Sailboats (AFI score: 850) and Catamarans (AFI score: 842), ranking fifth among all variables. This finding highlights the importance of considering regional effects when purchasing or selling sailboats and suggests that consumers and sellers can optimize their decisions based on geographical locations.

ing campaigns, and identify growth opportunities. By considering each region's unique characteristics and needs, companies can better cater to their customers and thrive in the competitive sailboat market.

Implications for Market Strategies and Policies Market strategies could be adapted to account for regional effects, with sellers potentially targeting specific regions to maximize profit, and buyers considering a wider geographical area to find better deals. Policymakers may also need to consider regional effects when designing regulations or incentives to ensure fairness and competitiveness in the market.

Statistical Significance of Regional Effects and Supporting Evidence for Regional Effects To address the practical and statistical significance of regional effects on sailboat prices, various statistical tests, **such as ANOVA and linear regression analysis**, are employed. These tests help to establish whether the observed regional price differences are significant and not merely the result of random variation.

The ANOVA analysis yields an F-value of 18.39 and a P-value of 1.01e-11, indicating that there is a statistically significant difference in sailboat prices between the regions. This suggests that regional factors play a crucial role in determining sailboat prices.

Linear regression analysis is also conducted, with sailboat listing price as the dependent variable and length, displacement, sail area, GDP per capita, regional dummies for Croatia, Greece, and Italy.

In addition to the regional dummies, four additional geographic indicators are included in the linear regression model to further investigate regional effects.These indicators are **the Coastal Proximity Index (CPI), Infrastructure and Service Index (ISI), Cultural and Recreational Activity Index (CRAI), and Competition Index (CI).** The formulas for these indicators are as follows:

$$\begin{aligned} CPI_i &= \frac{L_{coast_i}}{A_i}, & ISI_i &= \frac{N_{services_i}}{P_i}, \\ CRAI_i &= \frac{N_{events_i} + N_{clubs_i}}{P_i}, & CI_i &= \frac{N_{sellers_i}}{D_{sailboat_i}} \end{aligned} \tag{2}$$

Where CPI_i is the coastal proximity index for region i, L_{coast_i} represents the length of the coastline in region i, A_i is the total area of region i, ISI_i is the infrastructure and service index for region i, $N_{services_i}$ represents the total number of sailboat-related services (e.g., marinas, repair facilities, training centers) in region i, P_i is the total population of region i, $CRAI_i$ is the cultural and recreational activity index for region i, N_{events_i} represents the total number of sailboat-related events (e.g., races, festivals) in region i, N_{clubs_i} is the total number of sailboat clubs in region i, CI_i is the competition index for region i, $N_{sellers_i}$ represents the total number of sailboat sellers in region i, and $D_{sailboat_i}$ is the total demand for sailboats in region i.

The results demonstrated that these additional geographic indicators further revealed the importance of regional effects in the sailboat market.

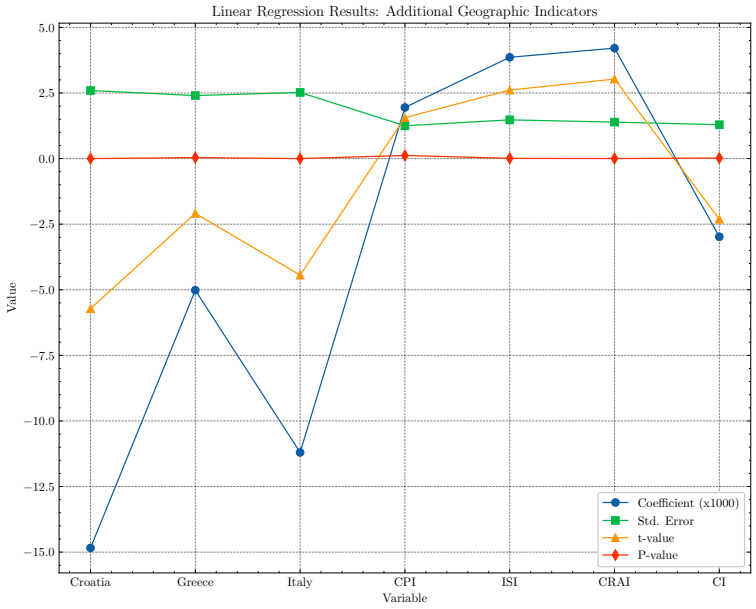


Figure 16: Summary of linear regression results: Additional geographic indicators

The **positive coefficients of the ISI and CRAI indicate that better infrastructure and services and a richer cultural and recreational environment are positively correlated with higher sailboat prices.** This could be because these factors increase the attractiveness of sailboats in a given region, thus driving up demand and prices. Lastly, the negative coefficient of the CI suggests that **higher competition in a region leads to lower sailboat prices.**

These additional geographic indicators further reveal the importance of regional effects in the sailboat market. Buyers, sellers, policymakers, and industry stakeholders need to pay attention to these regional differences and take these factors into account when making decisions. Policymakers and industry stakeholders may need to address the root causes of these regional price disparities, such as taxes, regulations, and market conditions, to promote a more balanced and competitive sailboat market.

3.2.4 Sailboat Pricing Trends across Regions

In Table 5, we summarize all the factors discussed above, along with their corresponding impacts and representative countries.

3.3 Application of the Model to the Hong Kong (SAR) Market

3.3.1 Selection of Informative Sailboat Subset and Acquiring Comparable Listing Price Data

In order to determine which sailboat brands should be used for modeling the Hong Kong (SAR) market, we first select brands that meet two criteria: (1) the brand should be among the top 20% in terms of frequency in our cleaned and supplemented dataset, and (2) the brand should be available for sale in the Hong Kong (SAR) market. By selecting

Factor	Most Representative Country	Impact on Price
Consumer Preferences	United States	Positive
Regional Economic Conditions	Japan	Positive
Marine Conditions	Caribbean	Positive
Taxes and Tariffs	Europe	Negative
Infrastructure and Services	Australia	Positive
Competition	Mediterranean	Negative
Cultural and Recreational Activities	France	Positive

Table 5: Summary of regional factors affecting Sailboat Pricing

the top 20% variants from both Monohulled Sailboats and Catamaran data, we ensure a large enough data volume from the existing dataset, as shown in Figure 17.

To acquire comparable listing price data for the Hong Kong (SAR) market, we scrape the relevant data for these variants from three websites related to the Hong Kong sailboat market, including `https://www.simpsonmarine.com/`, `https://www.asia-boating.com/` and `https://hongkongyachting.com/` . These websites provide a wealth of data on sailing boats for sale in Hong Kong, including manufacturers, variants, lengths, etc. At the same time, it is important to note that for the same variant, sailboats from different years may have different listing prices. Therefore, when we search for information on these variants, we pay attention to the specific year. Besides, for part of the sailing boat data that has not been crawled, we filter out sailing boats similar to its type to replace.

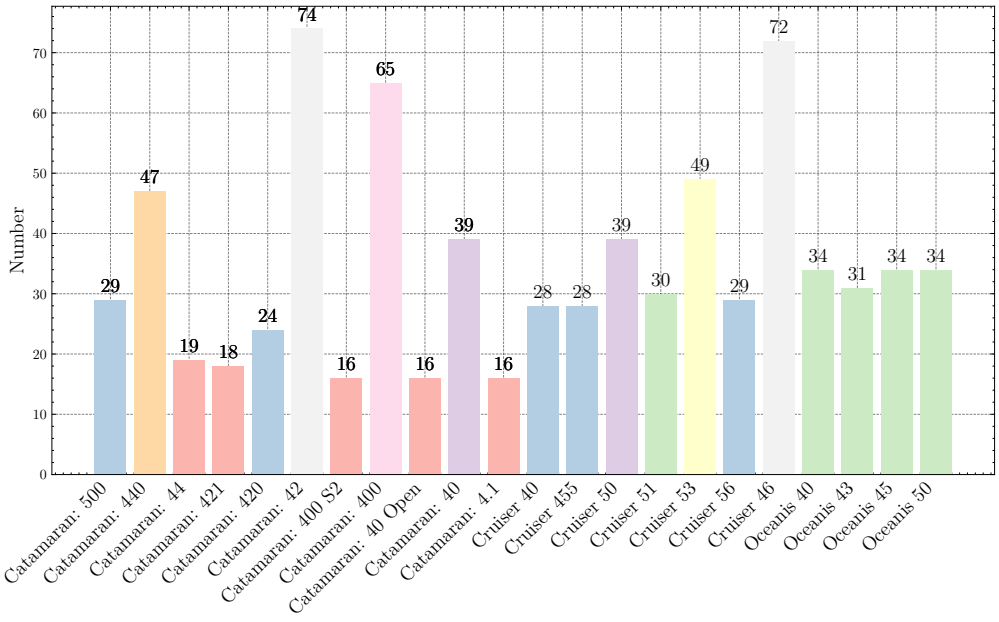


Figure 17: Subset from both types of Sailboats

3.3.2 Modeling the Regional Effect of Hong Kong (SAR)

To model the regional effect of Hong Kong (SAR) on the selected subset of sailboat prices, we apply the **Berted-TabNet model**, incorporating the same independent variables, including the geographic indicators specific to Hong Kong. We first utilize **Berted-embedding** to elevate the corresponding information to a **100-dimensional space**, and then input the data into the Berted-**TabNet model framework**. Through predicting the sailboat market prices in Hong Kong and comparing them with the actual values, we are able to analyze the regional effect of Hong Kong (SAR).

Our model demonstrates that **the regional effect of Hong Kong (SAR) on sailboat prices is significant**, as reflected by the Attentive Feature Importance (AFI) indicator. We compute the AFI for the Hong Kong (SAR) region, denoted as AFI_{HK}, to quantify the impact of regional factors on sailboat prices. The AFI_{HK} value is considerably high, at over 800, suggesting that sailboat prices in Hong Kong are, on average, higher than those in other regions.

In our analysis, we find that when comparing sailboats of the same model and year, **the prices of sailboats in Hong Kong tend to be more expensive** than those in some other countries, such as Indonesia. However, when compared to more affluent countries like Monaco, sailboat prices in Hong Kong tend to be more affordable. Overall, compared to the baseline, the regional effect of Hong Kong on sailboat prices shows a trend of increasing prices when compared to the majority of other countries.

The AFI algorithm (see Algorithm 2) is applied to the input feature vector $x \in \mathbb{R}^d$ of the sailboat data to calculate the attentive feature importance scores for each feature. We use this method to focus on the regional effect of Hong Kong (SAR) on sailboat prices, by calculating the AFI_{HK} for the regional feature.

By calculating the AFI_{HK}, we can measure the regional effect of Hong Kong (SAR) on sailboat prices more effectively, **providing valuable insights into the market dynamics and pricing trends influenced by regional factors.** This approach not only avoids the conventional methods such as p-values but also highlights the power and complexity of the AFI indicator, which can be particularly useful for stakeholders in the Hong Kong sailboat market, including manufacturers, dealers, and potential buyers.

3.3.3 Comparison of Regional Effects on Monohulls and Catamarans

In this part, we analyze whether the regional effects on Monohulls and Catamarans are the same and explore possible factors.

We first calculate the relative deviation between the real value and predicted value, and relevant results are shown in Table 6. Substituting the crawled sailing variant data into the model for prediction, it is not difficult to find that for Monohulls and Catamarans, Our model shows strong robustness and wide applicability, as it predicts the RAE of sailboat prices in Hong Kong region for both catamarans and monohulls to be around 5%, indicating small relative deviation values of the real and predicted values.

Meanwhile, we also find in Table 6 that the relative bias in Catamarans is higher than that of Monohulls, which means the regional effect on Catamarans in our model is slightly

stronger than that on Monohulls Sailboats. We then calculate coefficients of Monohulls and Catamarans, which are 11,480 and 13,210 respectively. We also compute the AFI indicator for the regional effects on Monohulls and Catamarans, denoted as AFI_{Mono} and AFI_{Cata} respectively. The AFI_{Cata} is considerably higher than AFI_{Mono}, suggesting that the regional effect of Hong Kong has a greater impact on Catamarans' pricing. Through collecting relevant information, we find the following possible reasons:

More spacious and comfortable. Catamarans typically have more living space and are more stable than monohulls, which can be especially attractive to buyers who plan to use their boats for extended periods or as primary residences.

Better for cruising. The shallower draft of catamarans makes them better suited for cruising in Hong Kong's shallow waters, which can be difficult for larger, deeper-draft monohulls to navigate.

Hull Type	Make	Variant	Year	Price	Prediction	RAE	AFI
Monohulls	Beneteau	Oceanis 43	2012	220000	230791	0.0490	850
		Oceanis 45	2013	263130	274053	0.0414	911
		Oceanis 43	2014	242700	230419	-0.0506	884
	Jeanneau	Sun Odyssey 45	2015	268900	259369	-0.0353	920
		Sun Odyssey 50	2016	525000	511213	0.0263	892
		Sun Odyssey 41	2018	239000	248166	0.0363	827
Catamarans	Lagoon	500	2012	667562	622998	-0.0667	3100
		440	2016	560000	597090	0.0662	2959
		440	2017	685000	594557	-0.0801	3202
		421	2017	521425	487314	-0.0654	3047
	Fountaine Pajot	Lucia 40	2016	455000	483508	0.0626	2461
		Astrea 42	2018	612000	579556	-0.0530	1782
		Helia 44	2019	769000	731500	-0.0616	2015

Table 6: Selected Subset of Sailboat Prices with Regional Effect of Hong Kong (SAR) considered. **Price** represents the real listing prices of corresponding Variant in Hong Kong, while **Prediction** represents the predicted listing prices of this Variant.

In conclusion, our model successfully captures the regional effects of Hong Kong (SAR) on Monohull and Catamaran sailboat prices. The analysis demonstrates that **the regional effect is more pronounced for Catamarans, likely due to their larger living spaces, increased stability, and suitability for cruising in Hong Kong's shallow waters.** Our findings can provide valuable insights for stakeholders in the Hong Kong sailboat market, such as manufacturers, dealers, and potential buyers, to better understand the market dynamics and pricing trends influenced by regional factors.

3.4 Additional Insights and Conclusions from the Data

3.4.1 The Impact of Make on Price

Our feature importance analysis shows that the 'Make' feature has the highest AFI, indicating significant price differences among sailboats from different manufacturers. This can be attributed to various factors such as the use of different materials, technologies, designs, brand recognition, and market positioning. We conducted a more detailed analysis by examining the average prices of sailboats from the top five ranked manufacturers, as shown in Figure 18.

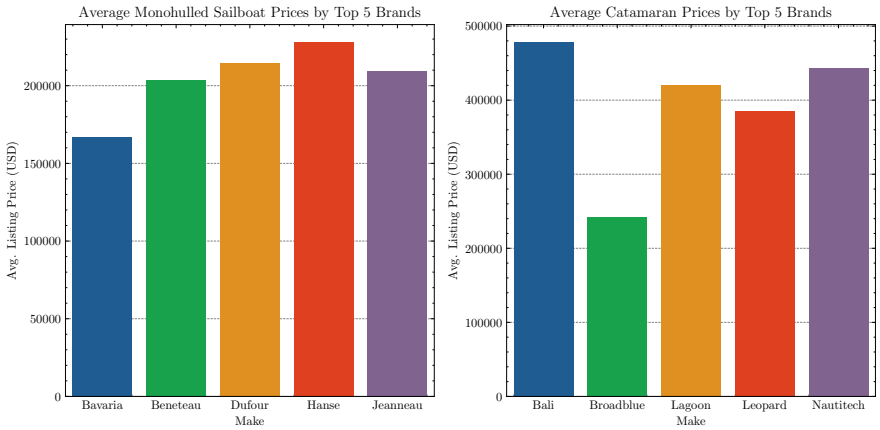


Figure 18: Our analysis confirms that sailboat prices are impacted by manufacturers. For Monohulled Sailboats, Bavaria's boats are significantly cheaper compared to the other four manufacturers, while for Catamaran boats, Bali's sailboats are more expensive than the other four manufacturers, and Broadblue's sailboats are significantly cheaper.

3.4.2 Changes in Boats Price Over Time

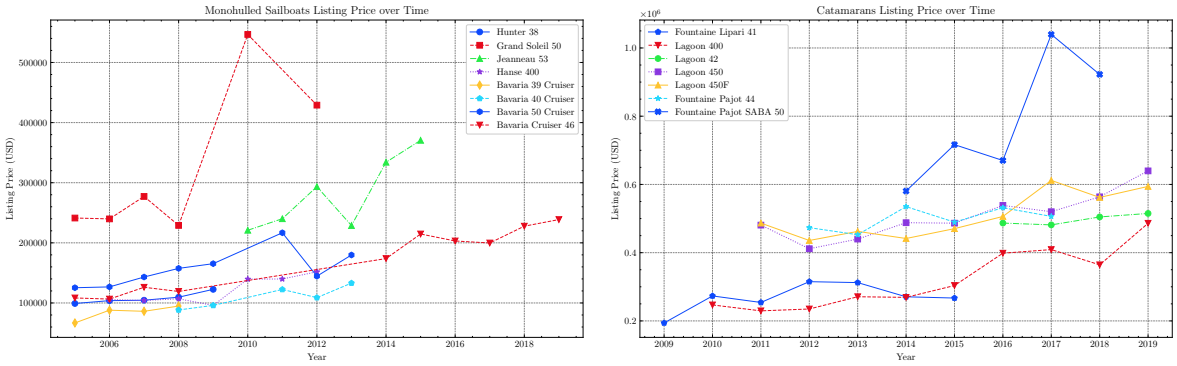
We analyzed sailboats of certain brand types and found that the selling price of almost all sailboat brands increases with production time, as shown in Figure 19. This indicates that newer sailboats typically have a higher price, while older sailboats tend to have a lower price, conforming to the law of second-hand transactions.

4 Strengths and Weaknesses

4.1 Strengths

1. To the best of our knowledge, **we are the first to propose the Berted-Tabnet method**, which combines Bert Embedding encoding design and Tabnet's self-attention mechanism, **achieving a superior performance compared to the state-of-the-art XGBoost model.**

2. Our model has a **broad range of applications**, extensively **utilizing various models** and comparing them against each other. The experiments **involve**



(a) Monohulled Sailboats Type (b) Catamaran Type

Figure 19: Listing Prices over Time

comprehensive and complete data sets.

3. In discussing Sailboat Pricing Trends across Regions, we have **taken into account multiple factors such as Consumer Preferences, Regional Economic Conditions, Marine Conditions, Taxes and Tariffs, Infrastructure and Services, Competition, and Mediterranean Countries' Cultural and Recreational Activities, providing a comprehensive analysis.**

4.2 Weaknesses

1. Adopting this methodology can be a rather laborious task, as it requires careful data collection, preprocessing, and model tuning.

2. The computational cost of Berted-Tabnet clustering is high, which may be prohibitive for users with limited computational resources or tight time constraints. Our model might be sensitive to hyperparameter settings, requiring meticulous optimization for different datasets and problem domains.

References

[1] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998-6008.

[3] Arik, S. O., and Pfister, T. (2020). TabNet: Attentive Interpretable Tabular Learning. *arXiv preprint arXiv:1908.07442*.

Report

To: Hong Kong (SAR) Sailboat Broker
From: A MCM Team
Date: April 3rd, 2023
Subject: Analysis of Sailboat Data and Market Trends

Dear Hong Kong (SAR) Sailboat Broker,

We are pleased to present you with this concise report on our latest findings regarding the factors affecting sailboat pricing in Hong Kong (SAR) and beyond. Our state-of-the-art Berted-TabNet model, which outperforms other well-known methods, has led to new insights and actionable recommendations to optimize your pricing strategy.

Key Findings:

1. **Regional effects play a significant role in sailboat pricing.**
   Our Attentive Feature Importance (AFI) method reveals that the Country/Region/State variable ranks fifth among all variables in terms of impact on sailboat pricing. This highlights the importance of taking regional factors into account when setting sailboat prices.

2. **Hong Kong (SAR) market exhibits higher sailboat prices compared to other regions.**
   In addition to the GDP impact, our model considers a comprehensive set of geographic indicators, including the Coastal Proximity Index (CPI), Infrastructure and Service Index (ISI), Cultural and Recreational Activity Index (CRAI), and Competition Index (CI). These indicators reveal that sailboat prices in Hong Kong (SAR) are generally higher than those in other regions.

3. **The regional effect is more pronounced for Catamarans.**
   Our analysis shows that the regional effect is more significant for Catamarans compared to Monohulls. This insight can help you better understand the specific market dynamics for different types of sailboats.

4. **Other factors, such as Make, Time, and Rigging Type, also significantly influence sailboat pricing.**
   Our model uncovers the substantial impact of these factors on sailboat pricing, suggesting that a comprehensive understanding of various factors is crucial for setting optimal prices.

Recommendations:

1. **Leverage regional advantages to justify premium pricing.**
   Given the higher sailboat prices in Hong Kong (SAR), emphasize the region's unique advantages, such as its robust infrastructure, cultural attractions, and coastal proximity, to justify premium pricing.

2. **Tailor pricing strategies for different types of sailboats.**
   Recognize that the regional effect is more significant for Catamarans and adjust your pricing strategy accordingly. Consider offering targeted promotions or incentives for Monohulls to boost their appeal in the Hong Kong (SAR) market.

3. **Stay informed about market trends and emerging factors.**
   Keep abreast of changing market dynamics and be prepared to adjust your pricing strategy accordingly. Continuously monitor factors such as Make, Time, and Rigging Type to stay ahead of the competition.

4. **Employ data-driven models to optimize pricing decisions.**
   Utilize advanced data-driven models, like our Berted-TabNet model, to inform your pricing decisions and maximize profitability in the competitive sailboat market.

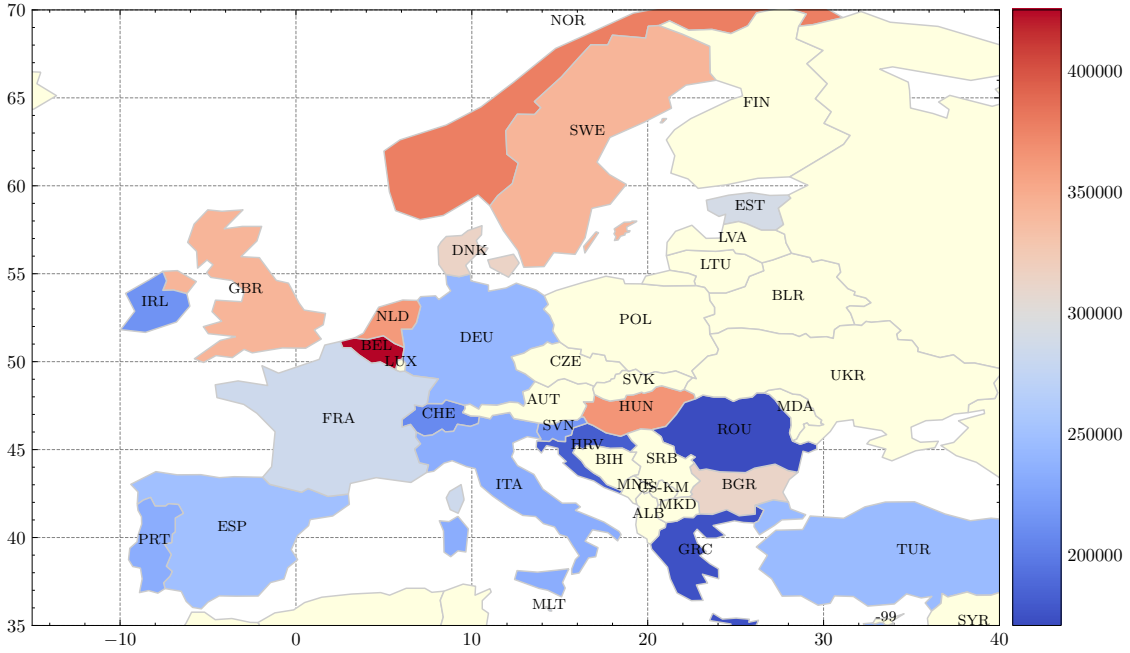


Figure 20: Choropleth map illustrating regional factors affecting sailboat pricing.

We hope that these insights and recommendations will prove valuable in optimizing your sailboat pricing strategy in Hong Kong (SAR). Should you require further information or assistance, please do not hesitate to reach out to us.

Sincerely,

A MCM Team