选择的问题	2023 MCM/ICM 总结表	团队控制号
Y		2330869

扬帆远航，发现更多

摘要

帆船的价值随着其自身的老化和市场条件的变化而变化。为了更好地了解帆船市场，我们建立了以下三个模型：基于随机森林的预测模型，使用非参数方法的有限混合模型，K-均值聚类子集选择

对于问题1，通过对帆船数据的探索，我们发现特征变量和目标变量之间存在着很强的复杂性。结合数据集中多特征变量的特点，我们选择用K-Fold Cross-Validation建立**随机森林**算法模型。它可以达到较高的**预测**精度，并有效避免模型的欠拟合和过拟合。我们通过该模型得到了对二手帆船价格影响最大的两个特征：帆面积和排水量。

对于问题2：在数据探索过程中，我们发现不同地区的帆船平均价格存在明显差异，如图6所示。为了研究区域对价格的具体影响，我们考虑了三个区域特征：平均货物吞吐量、GDP和海岸线长度，并利用基于EM**算法的非参数方法**建立了一个**有限混合模型**。与传统的模型不同，它可以有效地纠正数据中区域异质性造成的偏差。我们对问题2的结论是：GDP是区域效应的主要原因。

对于问题3：我们使用改进的K-means模型来获得一个具有以下特征的帆船子集
高的信息含量。然后，在讨论香港对帆船价格的区域影响时，我们继续使用模型1。我们发现，香港对单体帆船和多体帆船有不同的影响。

对于问题4，由于我们在建模时不仅考虑了与二手帆船本身有关的多种因素，而且还讨论了地理、政策和需求等方面的问题，因此我们的模型具有很强的适应性。它可以用于我们研究范围之外的地区。

此外，我们对所使用的两个主要模型进行了敏感性分析，结果客观地表明，我们的模型对参数变化不敏感。

关于问题5，我们总结了通过模型得到的结论，并为香港的经纪人编写了一份关于二手帆船的定价报告。该报告通俗易懂，结构合理，并包括便于理解的视觉辅助工具。

关键词预测；随机森林；K-折交叉验证；K-均值；有限混合模型；EM算法；非参数方法

内容

1 简介

帆船运动是一项多样化的运动，包括几种竞争形式，由不同的帆船联合会和游艇俱乐部负责管理。这些比赛项目包括船队内部、双人之间或团队之间的比赛。

这有助于确保一个公平的竞争环境，并为所有参赛者创造公平和令人兴奋的比赛。

帆船可以被认为是一种奢侈品，因为它们的购买和维护费用很高，通常需要大量的财政资源和时间。一些大型帆船的价格可达数百万美元，维护和保养的费用也非常昂贵。此外，帆船需要专业知识和技能来操作，所以可能需要雇用船员或支付培训费用。因此，帆船所有权往往被视为一种奢侈品，只有更富裕的人才能负担得起。然而，对于普通消费者来说，也有更小、更便宜的帆船可以购买和享受。

购买二手帆船可以是一个更实惠的选择，因为二手船的价格往往比全新的船要低很多。购买二手帆船也可以是一个更环保的选择，因为它减少了建造新船所需的资源和能源。

1.1 问题 背景

我们专注于一个更实际的问题，即在二手帆船销售中如何制定定价策略，这是在一个高度竞争的市场上。

目前，我们的实际情况是，出售二手帆船的市场非常复杂，需要重新考虑船和出售地区的问题。就船而言，我们需要考虑三个因素：成本、船的历史和船体状况。就销售地区而言，有三个因素需要考虑：地区人口、纬度和经济发展。

1.2 问题的重述

在这个问题上，我们得到了在欧洲、加勒比海和美国销售的部分帆船的销售数据。为了更好地了解帆船市场，为二手帆船定价，在下面的文章中，我们将：

1. 收集信息，为所提供的数据扩大有用的预测因素和样本量。
2. 根据扩大的数据集，建立一个合适的回归模型来预测表中每艘帆船的上市价格，并评估模型的准确性。
3. 建立模型，探索和解释区域对帆船上市价格的影响，讨论区域效应在实际和统计意义上的一致性。
4. 依靠一些方法来选择一个有信息量的子集，收集香港市场的可比销售数据，从中获得香港的区域影响。
5. 借助于发现更多有趣和有信息量的推论或结论

获得的模型结果。

6. 为香港的帆船经纪人提供一份简单易懂的二手帆船定价报告。

1.3 我们的工作

这个题目对我们提出了多种要求。我们的工作主要包括以下内容：

1 基于我们扩大的帆船数据，我们建立了一个随机森林模型来实现对二手帆船价格的准确预测。

2 在非参数情况下，基于EM算法的有限混合模型被设计出来，以分析区域效应对价格的影响。

3 通过改进的K-means算法选择含有信息的帆船子集，并按照模型一进行具体分析。

4 我们试图找到其他可靠的信息来解释我们的模型结果，发现了一些有趣的结论和推论。

5 我们整合了通过模型得到的结论，为香港的帆船经纪人准备了一份报告。

为了更直观地展示我们的工作流程，流程图见图3。

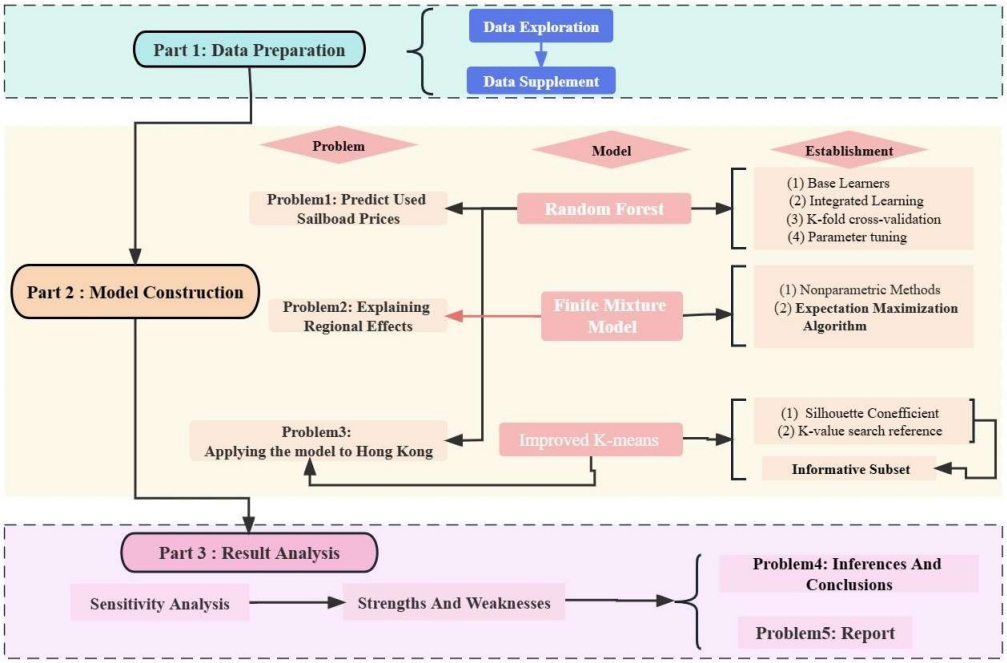

图1：工作流程

2 假设和理由

1. 传销者很细心、可靠、不骗人

仔细是指经销商在向他人购买时，会仔细检查船的健康状况，并根据状况制定合理的销售价格。不欺骗是指经销商在向消费者出售船只时不会隐瞒船只的真实状况。可靠是指经销商将提供售后

如果在购船后出现问题，向消费者提供服务。当经销商满足这三个特征时，就意味着消费者买船是一个完全的商业交易，不涉及人性的游戏。这个假设对我们的模型至关重要，它完全基于心理学。

2. 区域之间的差异，只考虑两个因素：经济和平均货物吞吐量

一般来说，中低纬度地区由于温度适中，是航海的较好地点。高纬度地区的天气除了经常出现极端天气外，主要特点是极易波动。作为一种奢侈品，帆船的购买和维护费用极其昂贵，所以在销售帆船时必须考虑经济因素。人口往往与一个地区的经济密切相关。一般来说，人口众多的地区，整体经济相对繁荣。在考虑地区差异时，这三个因素无疑是三个基本因素。

3 记号

本文所使用的关键数学符号列于表1。

表1：本文中使用的符号

符号	说明
$g_\theta(x_i)$	随机半参数EM算法的密度函数
$K(\cdot)$	内核密度函数
h_{jl}	组件和块密度估算的带宽
$f_{jb_k}(\cdot)$	密度函数Φ_{ij}
	一个n×m矩阵
$H(D)$	信息熵
$H(D\|A)$	条件熵 信息增益 剪
$g(D,A)$	影系数
$S(i)$	

4 数据 说明

我们通过多种数据来源补充了参数栏，Beam(ft), Draft(ft), Displacement(lbs.), Sail Area(sq ft), Average cargo throughput(ton), GDP(USD billion), GDP per capita(USD), Engine Hours, Coastline(km), 并通过网站信息补充和修改了数据集中缺失和问题的数据。

表1：数据、数据库网站和数据类型

数据库名称	数据库网站	数据类型

帆船数据	https://sailboatdata.com/	帆船
游艇世界	https://www.yachtworld.com/	帆船
红花海	https://saffron-marina.com/	帆船
美国人口普查局	https://data.census.gov/	地区
国际联合会 货运商协会	https://fiata.org/	地区
世界银行	https://data.worldbank.org/	地区
世界经济论坛	https://cn.weforum.org/	地区
Kaggle	https://www.kaggle.com/	地区
香港特区政府统计处	https://www.censtatd.gov.hk/	地区
游艇世界	https://www.yachtworld.com/	地区

表2：数据描述

	年	挂牌价	长度	桴鼓相闻	草案	摆放位置	帆船区	平均货物吞吐量	国内生产总值	人均国内生产总值	发动机小时数	海岸线
平均值	2010.38	226305.17	45.27	13.99	6.78	26476.9	1053.76	3.5E+07	1061.8	33338.61	9.62	5175.42
性传播疾病	4.05	144641.8	4.77	1.08	0.87	7956.23	267.99	5.2E+07	1078.41	20499.77	4.05	4360.76
闵行区	2005	45000	36	9.5	3.94	6393	516	5.5E+04	0.8	4871	1	0
25%	2007	139000	40.25	13.08	6.33	19621	861	1.9E+06	57.8	13933	6	121
50%	2009	190303.5	45	13.94	6.75	25353	1032	2.0E+07	650	30459	11	4964
75%	2014	267054.25	49	14.73	7.22	31085	1191	5.2E+07	2005	44494	13	7600
最大	2019	1885229	56	16.73	11.58	61900	2314	3.0E+08	3361	[illegible]	15	25148

寻找和处理缺失值是一个耗时的过程。我们用图1中的矩阵图来描述数据完整性的一般形状。从图中可以看出，货物吞吐量、GDP和人均GDP严重缺失，所以我们删除了缺失值的行。在分析过程中，我们发现一些帆船的挂牌价格有异常值。因此，我们对具有相同属性的船只进行了分类，并取其平均值，重新计算出离群值的影响。经过这两个步骤，2363组原始数据被缩减为1839组处理后的数据，为我们后续的分析提供了基础。我们的数据描述见表2。

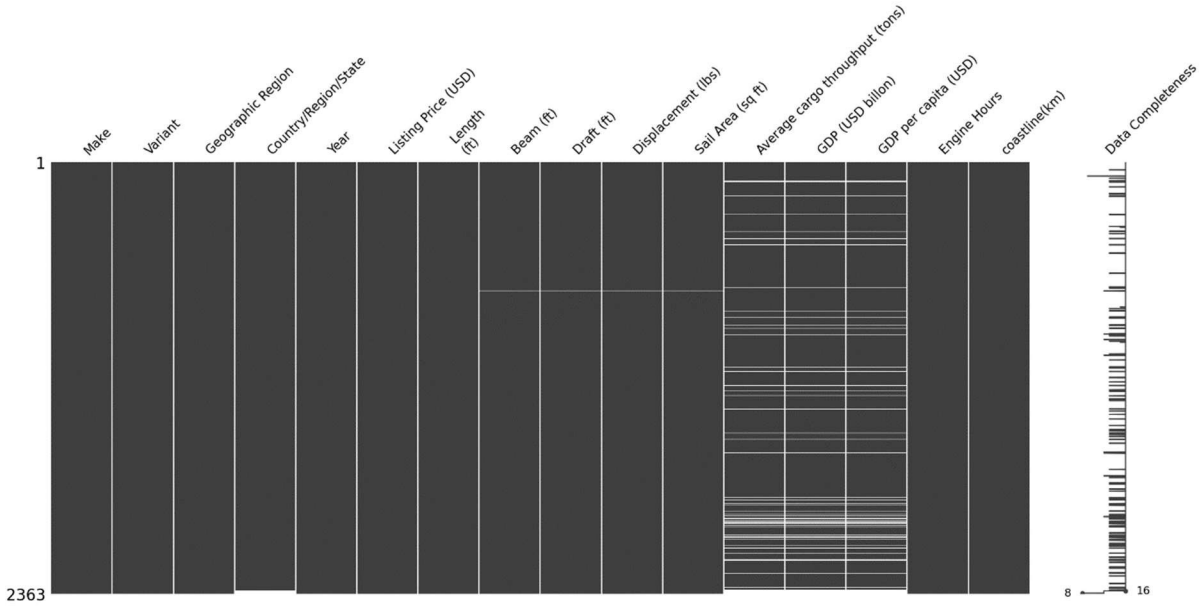

图2：缺失值处理

在对数据进行预处理后，用挂牌价绘制了图2所示的直方图，拟合曲线近似于正态分布，表明数据预处理过程的正确性。同时，如图3所示，价格具有区域效应。在图4中，可以看出帆船销售的频率随长度、销售价格和年份的变化而变化。

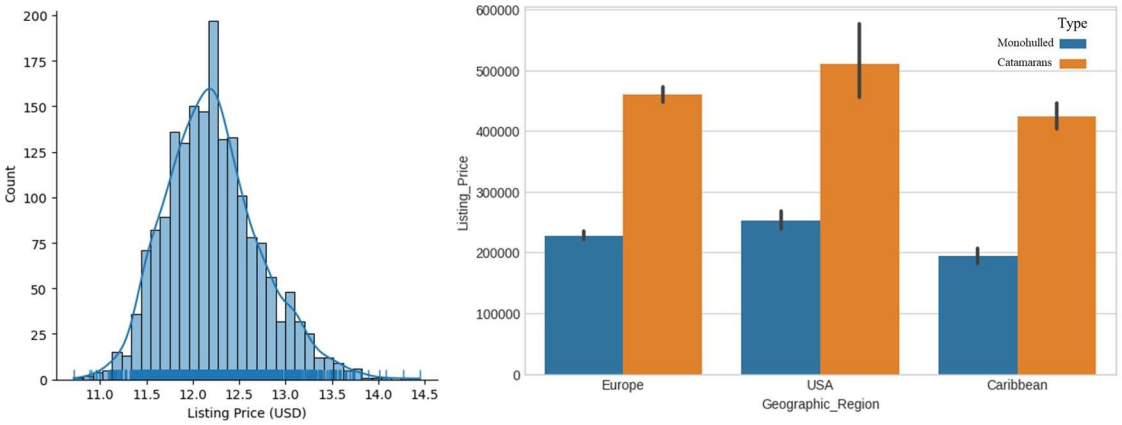

图3：直方图和拟合曲线（左）以及区域差异（右）。

总的来说，许多因素对帆船的销售价格有影响。因此，我们在图5中用热图直观地显示了不同因素与价格之间的关联性。由此可以得出结论，排量、帆面积、宽度、吃水、长度、制造、GDP和价格有比较强的关联性。因此，我们进一步分析这些因素。

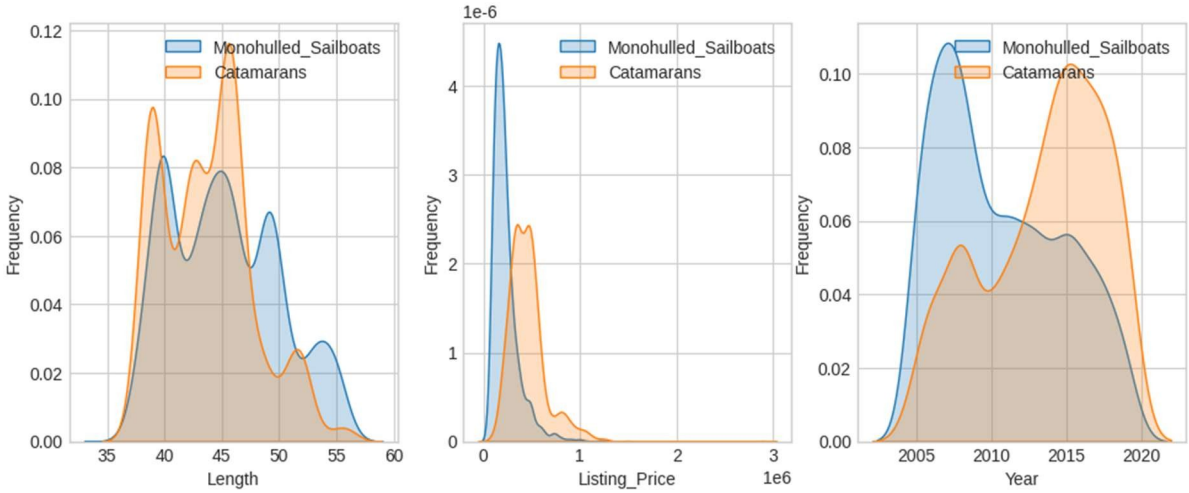

图4：销售状况

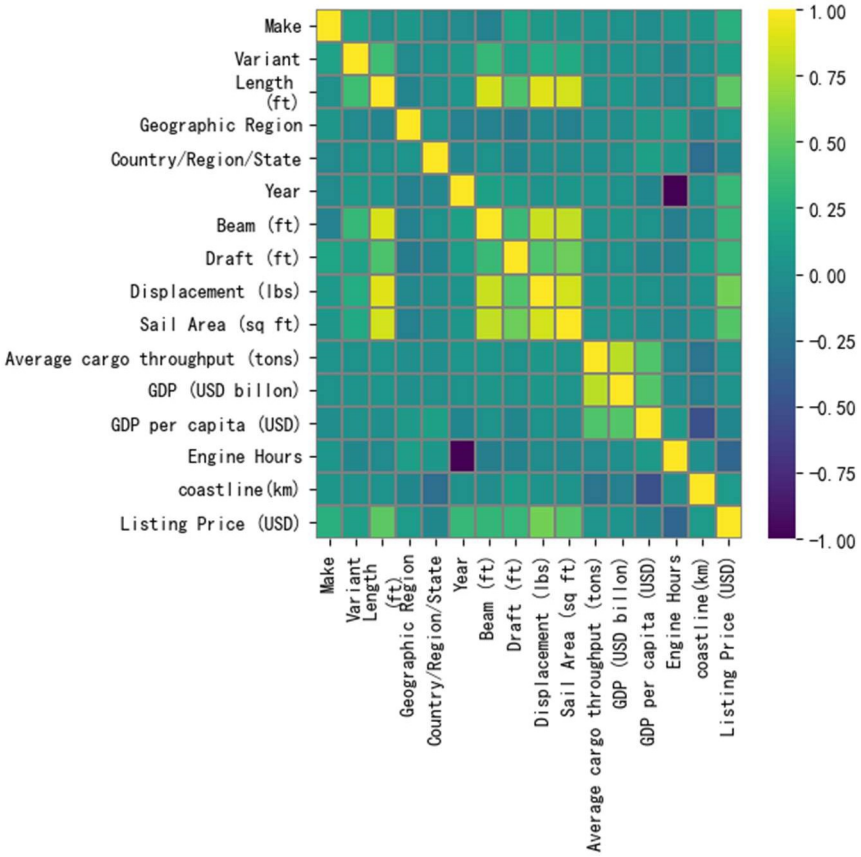

图5：关联性

在研究区域效应之前，我们分别将美国不同地区的平均帆船价格可视化。从图6可以看出，美国东海岸不同地区的帆船价格存在明显差异。基于此，我们初步认为，帆船价格数据具有区域异质性，在建立模型时需要考虑到这一点。

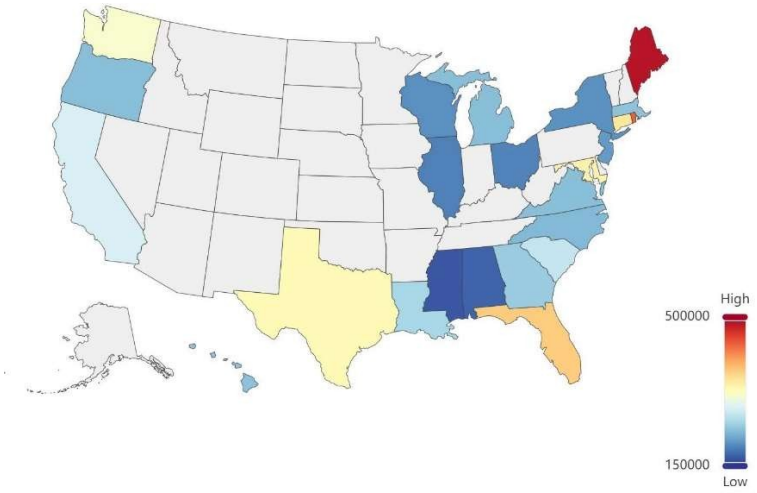

图6：帆船的平均价格（美国）

5 基于随机森林的预测模型

根据问题1的要求，需要选择所有的帆船特征变量，并据此构建一个帆船主要特征对挂牌价格的定量预测模型。然后用这个模型来预测测试集中选定的样本的帆船挂牌价格，用适当的指标比较样本的真实值来评估模型的准确性。挑战在于选择预测模型的算法，并改进模型的最佳参数以达到最佳精度。

首先是预测模型的选择。在数据描述部分，我们发现各特征变量具有多样性和复杂性，特征变量之间存在非线性现象，特征变量与目标变量（帆船上市价格）之间的关系较为复杂，因此，使用简单的线性模型必然无法收敛，无法达到较高的预测精度，我们优先采用非线性预测算法建模。同时，考虑到本题中特征变量的数量对于统计模型来说较大，对于深度学习来说较小，容易造成模型精度的欠拟合或过拟合，我们认为选择传统机器学习的算法建模更为合适。

在传统的机器学习算法中，对于回归问题，单一的回归模型仍然难以处理复杂的变量关系，反过来，我们很容易想到集合学习算法--随机森林[1]，用于建立回归模型。随机森林是一种基于集合学习的监督式机器学习算法，它可以结合不同类型的算法或同一算法多次使用，考虑到数据集的特点，对这个问题再合适不过了。

5.1 模式的建立

对于这个问题，随机森林的基础学习者不需要太复杂、

否则，容易出现过度拟合，可以选择决策树模型，这也有利于根据实际意义解释模型结果。

决策树模型以树状结构呈现，其过程类似于我们在现实生活中的做法，即在得出最终决策之前，对数据提出一系列的问题。决策树的精妙之处在于如何用最少的问题获得答案，即要求每个决策对最终决策的解决方案做出最大的贡献，决策树模型通过以下信息增益标准选择属性来实现这一点：

1. 数据集D的特征变量Ci的信息熵H（D），即特征变量Ci的信息的不确定性。这可以用公式1来表示：

$$H(D)=-\sum_{k=1}^{K}\frac{|C_k|}{|D|}\log_2\frac{|C_k|}{|D|} \tag{1}$$

2. 计算数据集D上特征变量A的条件熵H(D|A)，即以确定特征变量A为条件的信息不确定性，可用公式2表示：

$$H(D|A)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|} \tag{2}$$

3. 计算信息增益g(D, A)，即学习特征变量A后信息不确定性的降低，可以用公式3表示：

$$g(D,A)=H(D)-H(D|A) \tag{3}$$

4. 选择信息增益最大的特征变量作为子节点，然后递归调用上述方法，得到所有子节点

可以看出，决策树实际上是一种贪婪的算法，即它们优先考虑具有更大信息增益的特征变量。

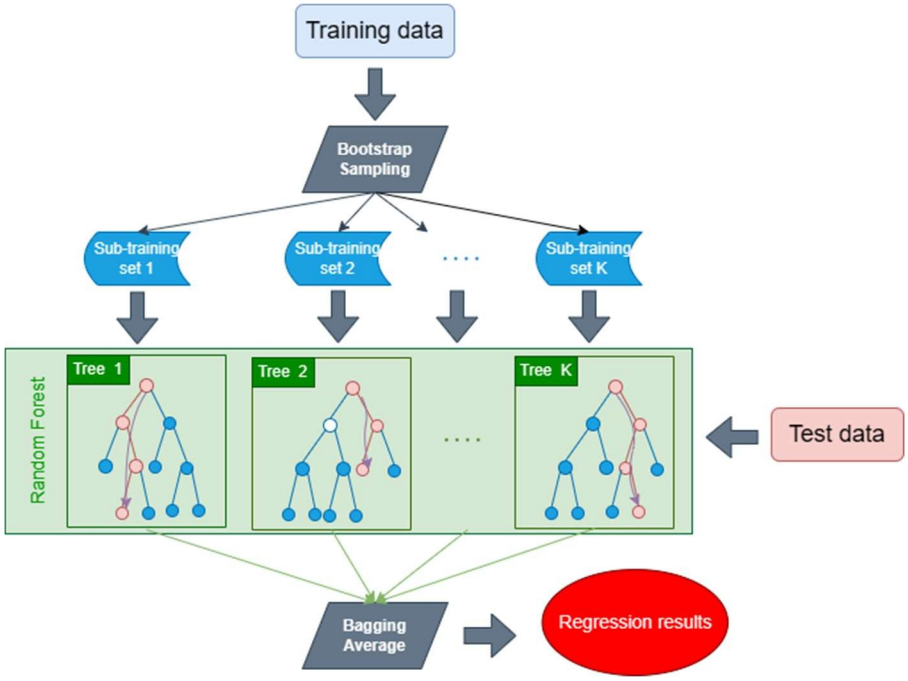

图7：随机森林算法的示意图

随机森林整合了多个决策树来获得结果，类似于现实生活中多人投票决定的场景。其基本原理是，首先通过使用Bootstrapping对数据进行反复抽样，从有限的数据集中选择一个N个样本的子集，然后基于这N个样本构建一棵决策树。回归是基于决策树结果的小变量，因此h1(x)代表其中一棵决策树的预测值，然后通过平均决策树的预测值得到随机森林回归的预测值。该算法的工作原理如图7所示。

5.2 模式的解决方案

5.2.1 对培训/测试 集的分析

表3：训练集测试集数据比较

特征变量	平均		准	
	训练 值	测试	训练 则	测试
长度(英尺)	45.3	45.25	4.78	4.85
地理上的雷金	1.10	1.14	0.50	0.53
梁(英尺)	13.97	14.00	1.08	1.11
吃水(英尺)	6.8	6.73	0.91	0.88
排水量（磅）。	26519.13	26713.65	8005.34	8321.34
帆的面积(平方尺)	1054.58	1055.24	274.07	261.60
⋮	⋮	⋮	⋮	⋮
国内生产总值(亿美元)	1073.29	1136.12	1075.56	1143.99

问题要求确定每个帆船品种的估计准确度，所以样本的选择需要包括每个帆船品种。根据样本总数和帆船种类数量之间的关系，测试集的大小约为样本总数的80%，这符合最重要的80/20定律，即只占约20%。

考虑到我们将使用建立的随机森林预测模型对测试集进行预测，我们首先对训练集和测试集的特征数据的均值和方差进行了统计，结果见表3。可以发现，测试集和训练集的数据分布基本一致，说明测试集能够很好地代表每艘帆船的特征。

5.2.2 K-Fold Cross- 验证

由于数据集中的特征变量较多，样本量较大，在模型调优过程中会发现，只有当参数的值（如基数学习者的数量、最大特征数、最大叶子节点数等）较大时，才能获得较好的预测结果，但同时，大模型也容易出现过拟合。为了提高模型的稳健性和准确性，验证集被用来解决这个问题：

然而，简单地将数据集分为三部分：训练集、验证集和测试集，实际上只是使模型对验证集（与原始训练集相关的较小的子集）更加收敛，这又导致了模型的过度拟合，没有实现交叉验证的想法。这反过来又导致了模型的过度拟合，无法实现交叉验证。因此，我们在划分集合的时候加入了随机性的想法。

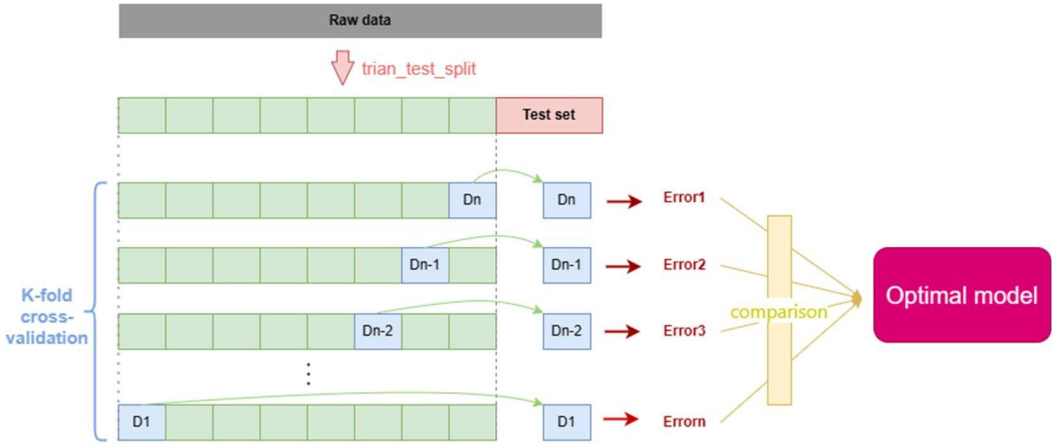

图8：数据划分过程

K-折交叉验证法如图8所示。首先，将所有的数据集分成N份，然后将其中一份作为验证集，每次不重复，而将其他n-1份数据作为训练集，用于训练模型和调整参考，因此我们需要训练N个模型，每次训练使用不同的训练集和验证集，最后计算N个模型的评价指标，并据此选出最佳模型。

5.2.3 参数调整和测试结果

我们使用随机搜索和个体搜索相结合的方法来自动对问题进行超参数调整。

起初，我们并不确定参数的大致位置，所以我们可以先在整个空间中随机取值，得到最佳超参数的大致位置，然后在该数量的超参数周围取一个较小的范围，逐一搜索网格。调整参数以达到最佳模型。模型结果显示，各特征变量的权重在图9中显示。

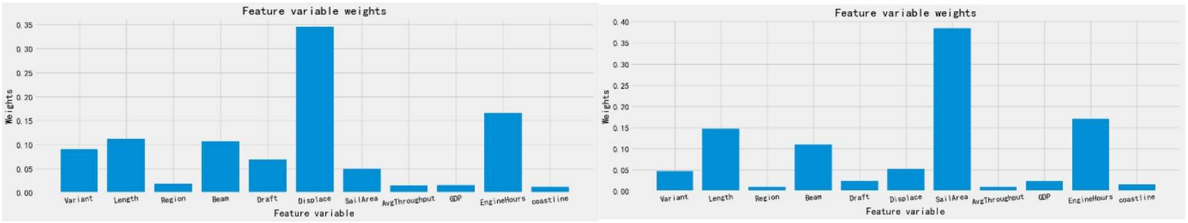

图9：单体船（左）和双体船（右）的特征权重

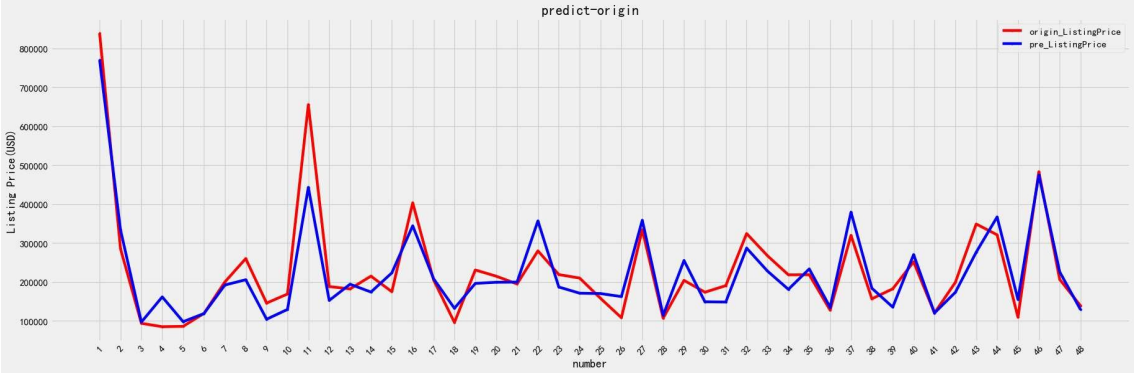

图10：实际结果与预测结果的对比图

将之前划分好的训练集输入到上述训练好的模型中，得到的预测结果的准确率超过80%。图10显示了原始测试集中的目标变量和预测的目标变量之间的比较。

6 使用非参数的有限混合模型 方法

由于我们使用的数据来自不同的国家和地区，即使是同一产品也会因为地区和经济的差异而有不同的交易价格，这使得我们的数据具有离散性。我们原本计划使用泊松模型进行数据分析，但经过研究，我们了解到传统模型在假设中往往没有考虑到数据的异质性，导致模型的偏差。因此，经过综合考虑，我们采用非参数方法建立了有限混合模型来解决这个问题。

6.1 模式的建立

本节重点讨论非参数多变量有限混合模型，它是方程4中提出的随机半参数EM算法[2] 的扩展，允许每个成分和坐标的不同分布。值得注意的是，如果密度

$f_{jk}(\cdot)$ 函数不取决于 ，它们不仅是有条件独立的，而且是同分布的。我们还假设Xi的坐标是有条件独立的，各块坐标是相同分布的。具体来说，我们可以把第k个坐标所属的块表示为b_k，其中1≤b_k≤B，B是这类块的总数。用这个符号，方程可以表示如下：

$$g_\theta(x_i)=\sum_{j=1}^{m}\lambda_j\prod_{k=1}^{r}f_{jb_k}(x_{ik}) \tag{4}$$

在本节中，我们将使用指数、、、和来分别指代一般的个体、成分（亚种群）、坐标（重复测量）和区块。因此，我们总是有1≤≤n，1≤≤m，1≤≤r和1≤≤B。为了估计模型（7），我们采用EM算法，其中包括一个E步骤和一个M步骤。这产生了一个加权的非参数核密度估计，表示如下：

$$f_{jl}^{t+1}=\frac{1}{nh_{jl}C_l\lambda_j^{t+1}}\sum_{k=1}^{r}\sum_{i=1}^{n}p_{ij}^{(t)}I\{b_k=l\}K\left(\frac{u-x_{ik}}{h_{jl}}\right) \tag{5}$$

$K(\cdot)$ 其中是内核密度函数，h_{jl} 是 j第1个成分的带宽，而块密度估计值，C_l 是第l个块中的坐标数。对于任何实数u为每个分量 $j \in \{1, \cdots, m\}$ 和每个块定义 $l \in \{1, \cdots, B\}$

我们有某些方程4的实例[3]，其中某些$f_{jb_k}(\cdot)$ 密度被认为是相同的，除了位置和比例的变化。这种情况被称为半参数化，因为对每个$f_{jb_k}(\cdot)$，就需要确定一个未知的密度，沿着 有许多位置和比例参数。为了说明问题，方程（6）确立了这一想法。

$$f_{jl}^{t+1} = \frac{1}{nh_{jl}C_l\lambda_j^{t+1}} \sum_{k=1}^{r} \sum_{i=1}^{n} p_{ij}^{(t)} I\{b_k = l\} K\left(\frac{u - x_{ik}}{h_{jl}}\right) \tag{6}$$

其中 $l = b_k$，为一般的k。

为了拟合方程6，mixtools软件包在spEM函数中提供了一种算法。npEM函数也需要更新所有的$f_{jb_k}(x_{ik})$。 i，j，以及x.spEM算法更新了一个名为Φ的n×m矩阵，它存储了$f_{jb_k}(x_{ik})$的当前值。

$$\Phi_{ij} \equiv \phi_j(Xi) = \prod_{k=1}^{r} f_{jb_k(x_{ik})} \tag{7}$$

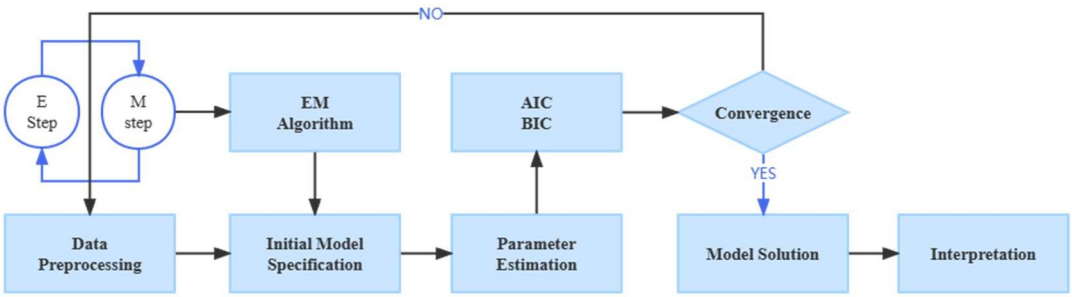

图11：有限混合模型的过程

6.2 解释区域影响

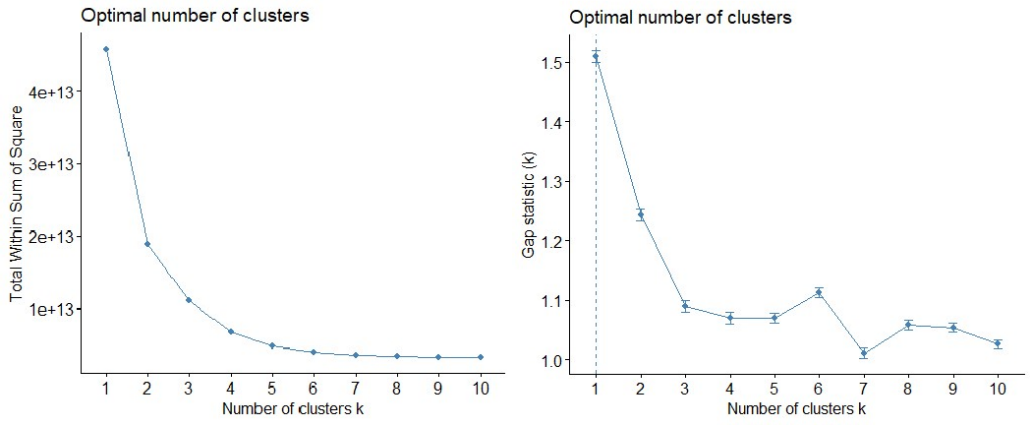

图12：寻找最佳的集群数量k

在数据预处理部分，我们得到了几个与价格有密切关联的变量。由于这个模型是为了解释区域因素对价格的影响，我们引入了三个变量：上市价格、GDP和海岸线。我们用有限的

混合模型，并确定区域效应的统计意义。

首先，我们进行K-Means聚类，并在图12（左）中绘制了一个图表来选择k的值。我们在曲线上寻找与某一k的总和值相对应的点，在那里，聚类数目和总和值之间的关系开始出现曲线或平缓。当图表显示出一个肘形时，它通常表示理想的聚类数量。从图中可以看出，肘部形状出现在k=4时。因此，我们选择k=4。同时，我们用不同的方法再次估计k的值，通过比较聚类数量和统计学差异之间的关系。从图中可以看出，当k=1时，差距统计量最大，这与前面的图中的结果相矛盾。通过统计方法，我们最终确定k=4为聚类的数量，在这个数值下构建的模型可以更好地解释我们的结果。

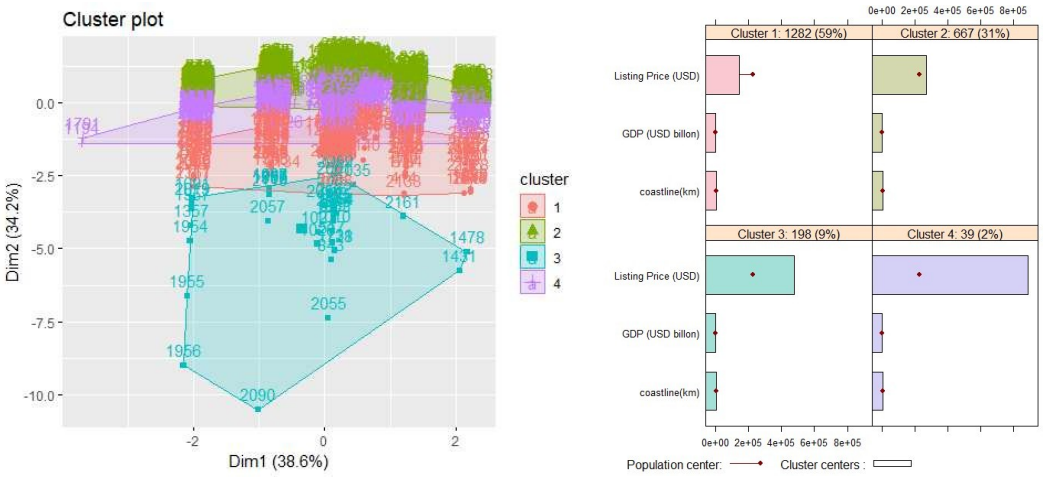

图13：聚类结果

表4：各群组变量的平均值

群体	挂牌价	国内生产总值	海岸线
1	135684.6	987.6517	5111.487
2	880757.9	1063.6125	6302.950
3	246913.3	1139.8484	5046.368

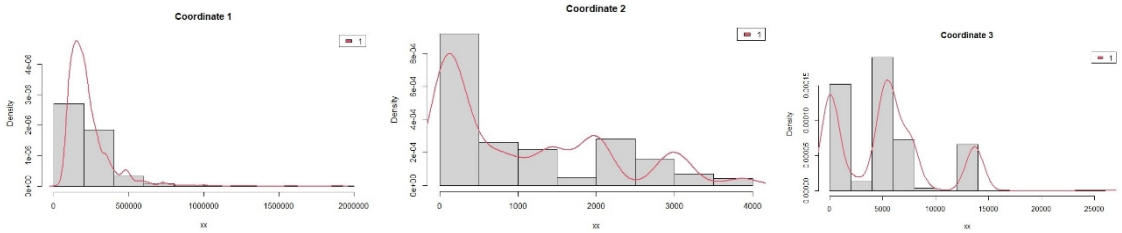

图14：估计的成分密度

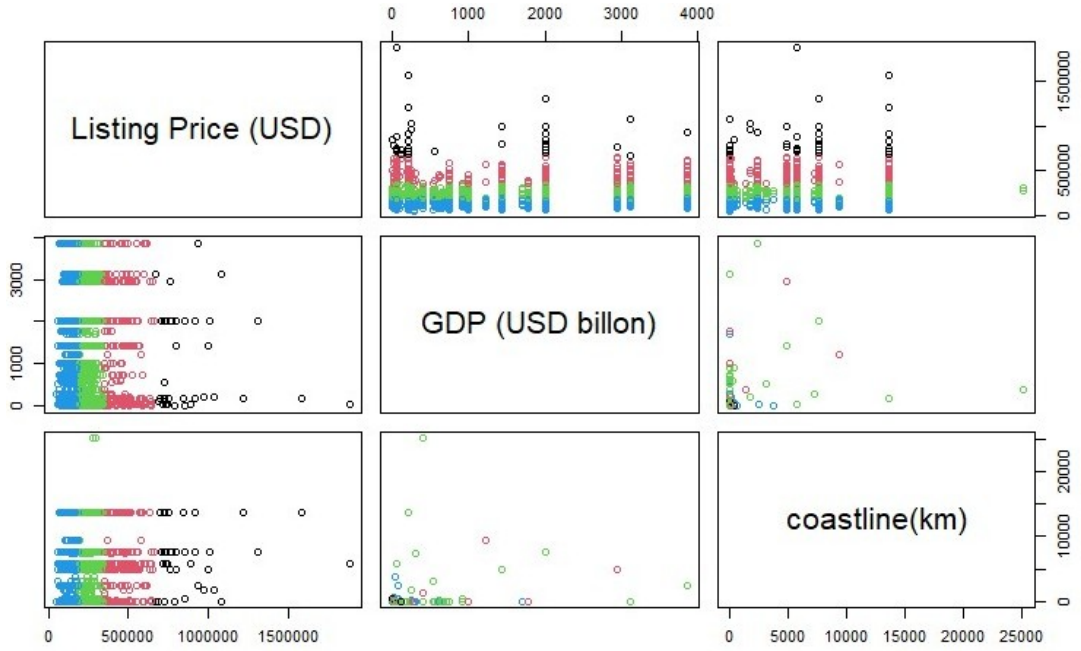

图15：配对图

7 K-均值聚类的子集 选择

7.1 模特在香港的作用 香港

对于问题1中建立的随机森林模型，图16显示了单壳船数据集的代表性决策树修剪后的决策，我们进一步分析了模型的细节。

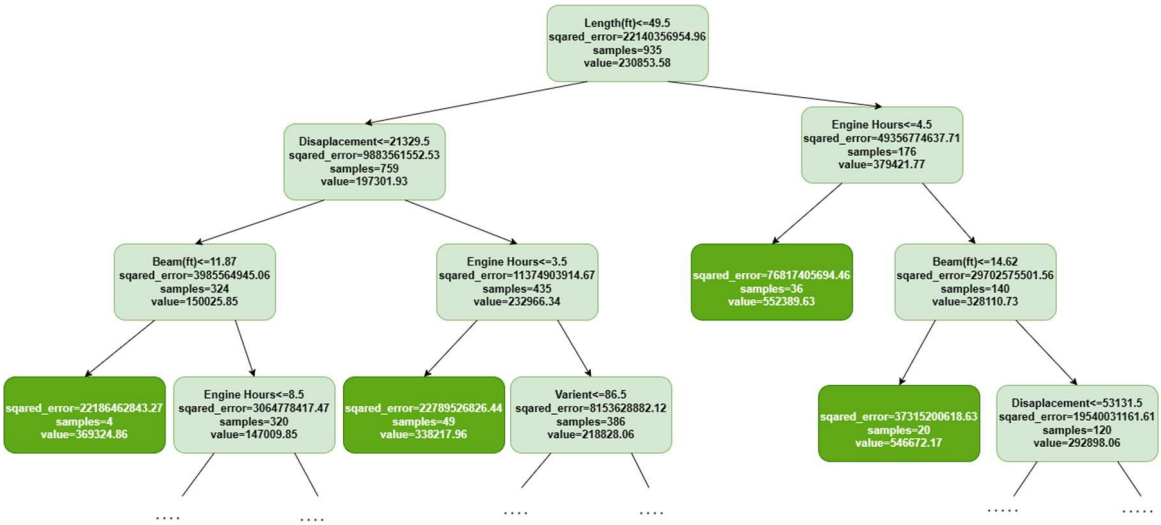

图16：单个船体的代表性决策树的部分结果图

从图15的结果可以看出，对于单体船的挂牌价格，船舶的排量、长度和发动机小时数更具有确定性；相应地，在从双体船数据集生成的随机森林中，帆面积、长度和发动机小时数更具有确定性。

船舶的发动机小时数更具有决定性。总之，对于单体船和双体船，决定其价格的因素与区域相关的特征变量的相关性较小。另外，在问题2的分析中，我们也有类似的结论，即排量、帆面积分别对单体船和双体船的挂牌价格起着决定性作用。因此，我们对特定地理区域的建模对香港市场的帆船定价有一定的指导意义。

为了使帆船子集尽可能地代表整个数据集，我们尝试了一些方法（改进的K-均值算法、排序点识别聚类结构（OPTI CS）、密度峰聚类（DP C）等。）算法（OPTI CS）、密度峰值聚类（DP C）等），最后选择了改进的K-means算法模型，它的特殊性在于它能自我搜索最佳聚类K值，并在优化初始聚类中心和剪影系数的基础上使用快速聚合方法。

然后，我们将聚类算法的结果作为香港上市价格数据的可比子集，这些数据来自上述数据集中表格中的网站的相同船只。

7.2 模式的建立

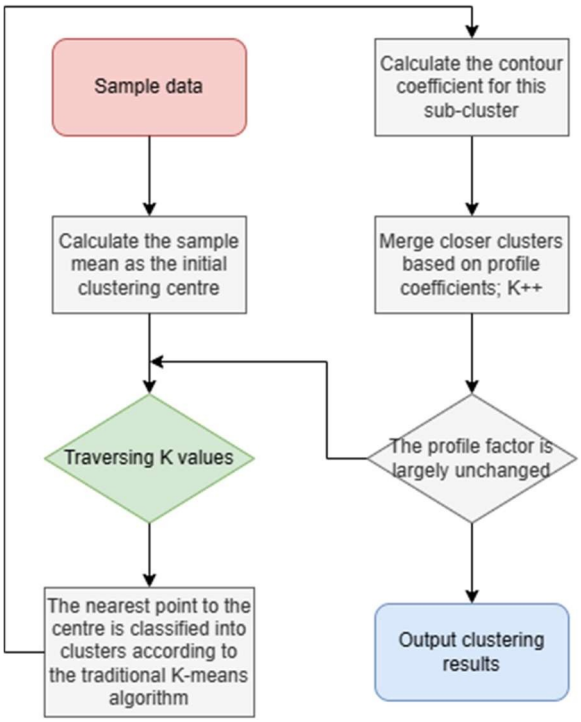

图17：基于初始聚类中心和剪影系数的K-means聚类算法

K-means聚类算法属于无监督学习的范畴，是一种迭代重定位的算法。算法中的k代表k个聚类，手段代表每个聚类中数据值的平均值作为聚类的中心，也就是说，聚类是由每个类的质心来描述。其基本思想是通过迭代使聚类结果所对应的损失函数最小化，找到一种对k个聚类的中心划分，其中损失函数可以定义为聚类中任何一点与中心点之间的距离之和。

该算法的核心步骤是寻找聚类中心的数量K。传统的K-means聚类算法的初始簇中心是随机选择的，簇数K也是人为确定的，一旦初始簇中心和K确定，整个聚类算法的迭代过程就由算法本身来计算，但传统的K-means聚类也存在一个问题，即由于初始簇中心和K的数值不同，每次聚类的结果也不同。为了解决这个问题，我们采用剪影系数法和K值最优搜索法。为了解决这个问题，我们采用了剪影系数法和K值最优引用法，我们认为所有数据的均值点应该是最接近聚类中心的，所以用均值点来初始化聚类中心。该算法的工作流程也可以用图17表示。

7.2.1 剪影系数

剪影系数是评价聚类效果好坏的一种方法，它包括内聚程度和分离程度。假设数据集中的一个元组Xi属于聚类Ci，那么凝聚度ai的计算方法是Xi与同一聚类中其他点的平均距离，代表Xi与同一聚类中其他点的不相似程度；分离度bi的计算方法是Xi与离Xi最近的另一个聚类C的平均距离，以及聚类C中所有的点，代表Xi与最邻近的聚类C的不相似程度。 最后的轮廓系数表示为公式8：

$$S(i)=\frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (8)$$

剪影系数的范围是[-1,1]，越接近1代表内聚和分离的程度都比较好。

7.2.2 K值搜索参考

k值的范围是[2,n] ，n代表数据点的数量。由于从这个问题来看，n的值太大了，我们可以用大的步长在大的范围内确定一个小的范围，然后用小的步长在小的范围内确定一个小的范围，找到最佳的k值，k值的评价函数是剪影系数，图18是我们搜索参数的过程。

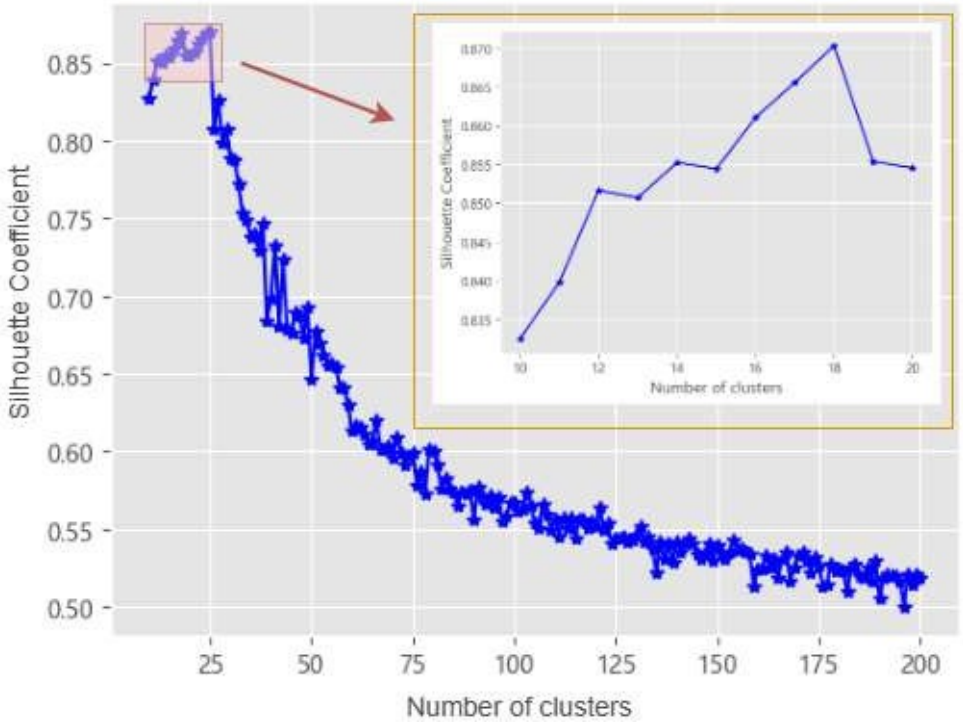

图18：集群的数量

然后我们得到了一个信息量极大的18个数字，并据此在网站上找到了相应船只在香港的挂牌价。

7.2.3 区域影响

使用开发的两个模型对香港进行了初步预测，并对香港和其他地区相同帆船的上市价格进行了比较，结果见表5：

表5：清单价格比较表

在香港上市的价格		双体船	
价格 (香港)	价格 (其他地方)	价格 (香港)	价格 (其他地方)
240000	95961	685000	633646
543375	479824	560000	539753
1770000	1311872	538500	534784
⋮	⋮	⋮	⋮

可以看出，对于双体船来说，各地区的挂牌价格相对稳定，但对于单体船来说，价格相对波动较大，香港市场可以对单体船和双体船的市场库存和价格进行相应调整。

为什么只有香港地区的单体船有这样的区域效应？我们经过查找资料发现，这种价格稳定的差异其实与近年来船舶漏油和海洋环境保护的问题有关。一些地区对单体船的使用采取了严格的规定，因为一旦发生事故，溢油和环境污染的风险更高。这导致了全球对双体船需求的增加。对单体船的严格规定在单体船贸易中产生了一些风险，这也导致了市场上的价格不稳定。

为单体船。

此外，我们从很多《香港海洋问题研究报告》中发现，香港的海洋环境保护队伍素质比较强，海洋环境保护的科研和咨询工作比较扎实。燃油污染[4]，规定如果因事故在香港造成污染损害，有关船主应承担赔偿责任。这也在一定程度上增强了单壳船在香港的区域效应。

8 敏感度 分析

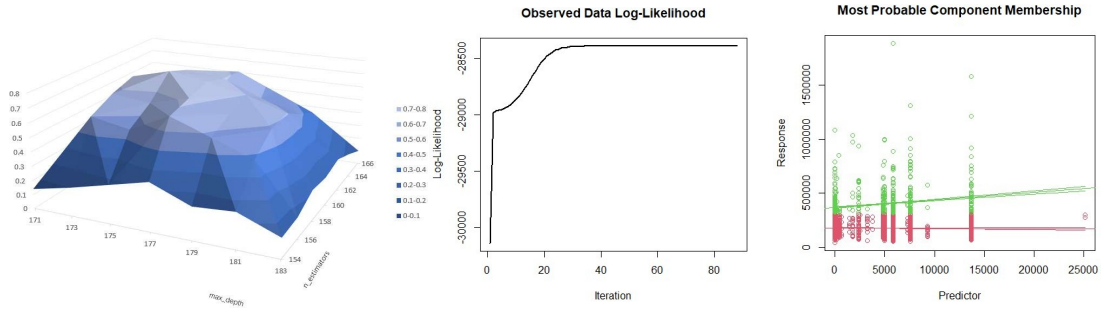

图19：敏感度分析的数字

EM算法是一种迭代算法，用于估计有限混合模型的参数。在每次迭代中，该算法都会更新成分参数的估计值，以及给定数据中属于每个成分的概率。当迭代次数达到87次时，就达到了收敛。曲线从一个较低的值开始，稳步上升，因此，它对数据是一个很好的拟合。在图18中，对于绿线对应的观察值，由于一些观察值处于较高的概率区域，而其他许多观察值处于较低的概率区域，这表明混合模型可以有效地将不同的数据点分配给不同的组，显示了良好的模型拟合。从图中可以看出，当迭代次数在25次左右时，曲线由陡峭变为接近水平，说明经过25次以上迭代的拟合模型已经非常接近收敛了。因此，可以得出结论，我们建立的有限混合模型是非常不敏感的，具有很强的普适性，代表了杰出的模型性能。

图19显示了随机森林模型预测挂牌价格准确性在n_estimators和max_depth下的趋势。从图中我们可以看出：当n_estimators=160，max_depth=175时，挂牌价格准确率可以达到较好的效果。此外，当n_estimators和max_depth分别从160和175向各个方向扩散时，模型对预测挂牌价格的准确性的影响是最小的（这一点从图的中间部分的山顶比较平坦就可以看出）。因此，该模型能够应用于不同类型的数据集，并具有很强的概括能力。

9 模型评估和进一步 讨论

9.1 优势

1. 随机森林算法对数据集的噪声不敏感，这有利于建立一个稳健的模型，而且使用一组不相关的决策树可以有效地防止模型的过度拟合。
2. 我们使用了基于EM算法的有限混合模型，它考虑到了数据的区域异质性，使我们的模型具有更好的预测性能。
3. K-means算法具有相对的可扩展性和高效性，我们对该模型进行了改进，得到了一个全局最优的模型，具有很好的聚类效果。

9.2 弱点

1. 随机森林模型很复杂，它们比其他类似算法需要更多的时间来训练。
2. 由于时间和数据集有限，我们缺乏对与区域相关的更多维度的探索。如果我们能获得更多与区域相关的特征数据，我们的模型将更加可靠，并具有更强的可解释性。

9.3 进一步 讨论

通过随机森林模型，我们可以使用对数据集更有针对性的数据集选择方法来构建决策树。基础学习者可以与其他学习者联合起来，使模型更加稳健

10 推论

我们的模型显示，排水量和帆面积对二手帆船的价格有最明显的影响。在单变体帆船中，排水量对其价格的高低具有最高的影响权重。而在双变体帆船中，帆面积的影响权重最高。通过查阅大量的文献，我们发现上面提到的两个帆船特征与影响帆船速度的两个重要参数完全对应。

首先是帆船的排量-长度比（D/L），其计算方法如下：

$$D/L = \frac{(Displacement/2240)}{(0.01 \times LWL)^3} \tag{9}$$

等式9说明，对于不同型号的帆船，等式9说明，对于不同型号的帆船，排水量比LWL（水线长度）更能决定参数的大小。排水量-长度比的意义在于，相对于水线长度，帆船越轻，帆船的速度潜力越大，特别是在排水模式下[5]。同时，排水量/长度比越低，帆船在航道上就越不稳定，对超载也就越敏感。因此，买家如果有需要使用单一变体帆船进行比赛，那么排水量

将是一个重要的考虑因素，这解释了它的突出地位。

其次是帆船的帆面积-排量比（SA/D），其计算方法如下：

$$SA/D = \frac{Sail\ Area}{(Displacement/64)^{2/3}} \quad (10)$$

虽然没有一个单一的数字可以概括每一种不同类型的帆船的性能，但SA/D提供了一个相对健全的方式来讨论帆的功率和重量之间的关系，这种关系决定了船的加速性、机动性和性能能力的许多方面[6]。SA/D也通常被用来衡量帆船达到最大速度的难易程度。在这个参数中，帆面积比排水量更能决定这个参数的大小。一般来说，帆面积与排水量之比低于15的船会被认为是航行能力不足；高于15的值则表示有相当好的性能。

由于时间成本等问题，我们没有关于帆船速度的信息，但根据以上两个参数公式，我们可以得到以下**推断**：**买家在购买帆船时主要考虑帆船的速度，单变体帆船的速度受排水量的影响较大，而双变体帆船的速度受帆面积的影响较大。**

参考文献

[1] Cap.605 燃油污染（责任和赔偿）（2023）第1（5）条。

[2] Benaglia T , Chauveau D , Hunter D R , et al. mixtools : An R Package for Ana- lyzing Mixture Models[J].Journal of statistical software, 2009, 32.

[3] Wang Q, Nguyen T. T., Huang J. Z., Nguyen T. T. An efficient random forests algorithm for high dimensional data classification[J].Advances in Data Analysis and Classification, 2018, 12(4)：953-972.

[4] McLachlan G J, Lee S X, Rathnayake S I.Finite mixture models[J].Annual review of statistics and its application, 2019, 6: 355-378.

[5] BoatQuest。https://www.sailmagazine.com/boats/comparing-design-ratios。于2023年4月2日访问。

[6] 帆船的生活。https://www.lifeofsailing.com/post/what-is-sail-area-displacement-

报告

首先，根据我们收集的数据，帆船销售在香港非常强劲，复合增长率相对较高。"就其规模而言，香港可能是全球游艇交付的第一大目的地"，香港的销售Charles Massey[3 https://www.sevenstar-yacht- transport.com/news/hong-kong-is-top-estination-for-yacht-deliveries]说。因此，在香港出售二手帆船是一个明智的、具有前瞻性的选择。根据我们的分析，我们发现二手帆船的销售价格与两个因素有关，即船的特性和区域效应。我们将为您详细介绍我们的分析结果。

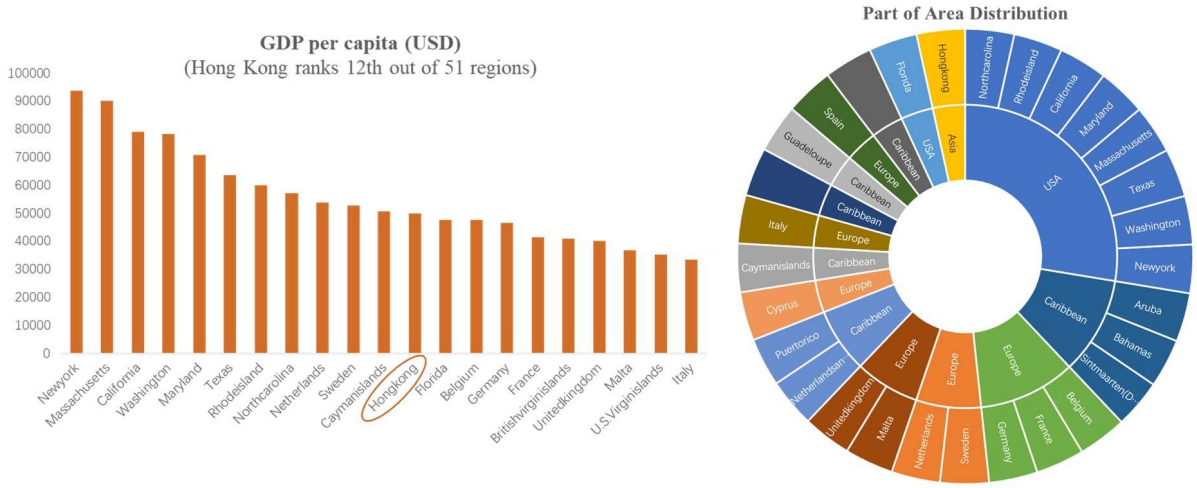

图19：香港的市场情况和地区分布图

区域因素：找出能影响香港地区帆船价格的主要因素，如经济状况、市场需求和帆船的可用性。

根据我们收集的数据，我们将区域效应分为三类：GDP、海岸线长度和平均货物吞吐量。然而，在我们的分析中，我们发现GDP是影响二手帆船销售量的主要因素。我们了解到，香港的GDP是...，在所有地区中排名...。因此，香港的二手帆船有一个很大的市场。

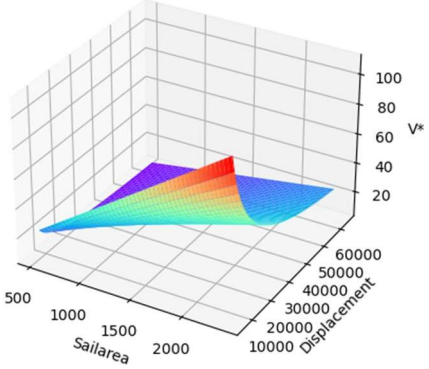

图18：速度、帆面积和排量的三维关系

在船的特性方面，通过研究，我们发现，帆面积和排量是影响帆船价格的主要因素。这两个因素与帆船的速度直接相关，所以我们推测，消费者在购买帆船时主要考虑的是速度。影响帆船价格的次要因素是长度，它与帆船的甲板面积有关，说明消费者在购买帆船时不仅考虑速度，还考虑与舒适度有关的甲板尺寸。因此，在定价时，我们建议售价应充分考虑帆面积、排水量和长度这三个因素的综合影响。

另一点不容忽视的是，在我们的分析中，我们发现单体帆船的预测是不稳定的，容易受到各种因素的影响，而双体船的预测则非常稳定。这是由环境和政策因素造成的。一旦单体帆船受损，将对环境产生比较严重的影响，而对于双体船来说，这种影响要小得多。因此，一些政府会利用政策来减少单体帆船的使用。因此，我们建议在销售中重点销售双体船。

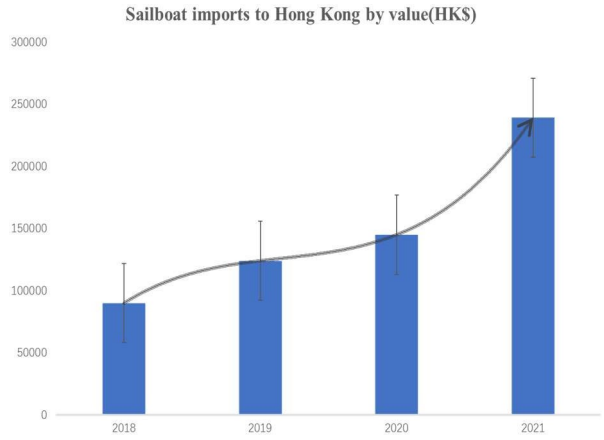

图19：香港的帆船进口情况

新数据显示，2021年，香港进口的娱乐用摩托艇（不包括船外机船）同比增长超过45%，从2020年的近20亿港元增至29亿港元。帆船的进口值从2020年到2021年上升了56%，从1.47亿港元到2.3亿港元。这两个类别的进口价值远远超过了过去十年中的任何一年。

根据分析，我们得出结论，香港是一个具有巨大潜力的二手帆船销售市场，这是一个明智的决定。在确定售价的时候、帆的面积、位移和长度因素应该是应充分考虑，定价应符合逻辑和合理。为了实现利润最大化，建议重点销售双体船作为主要产品。但是，作为卖家，也要注重提供良好的客户服务，包括售后服务，以提高客户体验。