探究婴幼儿奶粉市场的供求态势

摘要

现有电商平台 846 条关于婴幼儿奶粉的销售信息，每条信息由 11 个指标组成，其中评价量可以从侧面反应顾客对产品的关注度。

对于问题一，需要以评价量为因变量，分析其它变量和评价之间的关系，为商家以及宝妈提供可参考的方向，本题采用了多元线性回归分析法，通过探讨各指标与评价量之间的线性关系，得到相关系数方程，可知自变量**团购价对评价量的影响最大**，团购价每上升 **1%**评价量就对应下降 **1.96%**，其次是**段位和奶源产地**。对于问题二，需要以评价量为因变量，研究影响评价量的重要因素，已知多元线性回归在数据量较大的样本处理中容易受共线性干扰，而随机森林模型由于随机性的引入，不容易陷入过拟合，具有一定的抗噪声能力，精确度较高， 因此本题运用随机森林模型进行分析，确定重要特征，即重要因素，得到影响评价量的重要因素为**团购价**，且仅有此一项为重要因素。

关键字：多元线性回归;随机森林;预测;

一、 问题重述

现有某电商平台 846 条关于婴幼儿奶粉的销售信息，每条信息由 11 个指标组成，其中评价量可以从一个侧面反映顾客对产品的关注度。

请对所给数据进行以下方面的分析，要求最终的分析将不仅仅有益于商家，更有益于宝妈们为宝贝选择适合自己的奶粉。

（1） 以评价量为因变量，分析其它变量和评价量之间的关系

（2） 以评价量为因变量，研究影响评价量的重要因素。

二、 问题分析

2.1 对于问题一的分析

该问题要求以评价量为因变量，分析其它变量和评价量之间的关系。本题基于某电商平台提供的 846 条关于婴幼儿奶粉的销售信息，采用多元线性回归分析法建立与评价量（因变量）与其余指标（自变量）之间的关系相关的模型，并用以为商家供货和宝妈们为宝贝们选择适合的奶粉。

在销售信息所给出的 11 条指标中，除评价量（因变量）外，商品毛重和团购价为定量指标，它们提供了客观的度量标准，帮助了解和评估事物的状态和表现，指导决策和管理。结合其余 8 条定性指标来获得更全面和准确的理解。

对于定量数据，直接进行多元线性回归分析，若拟合度大小 R^2 满足需求则依次进行多重共线性检验、异方差性检验最终得到一个有关自变量（评价量）与定量指标（产品毛重、团购价）之间关系的方程，根据所得方程，进行解释分析其它变量和评价量之间的关系。

对于定性数据则需在多元线性回归分析操作之前进行虚拟处理，将文本类、非数值类数据进行转换，再依次进行操作。最终结合对于定性数据所得的分析结果对整体模型进行分析。

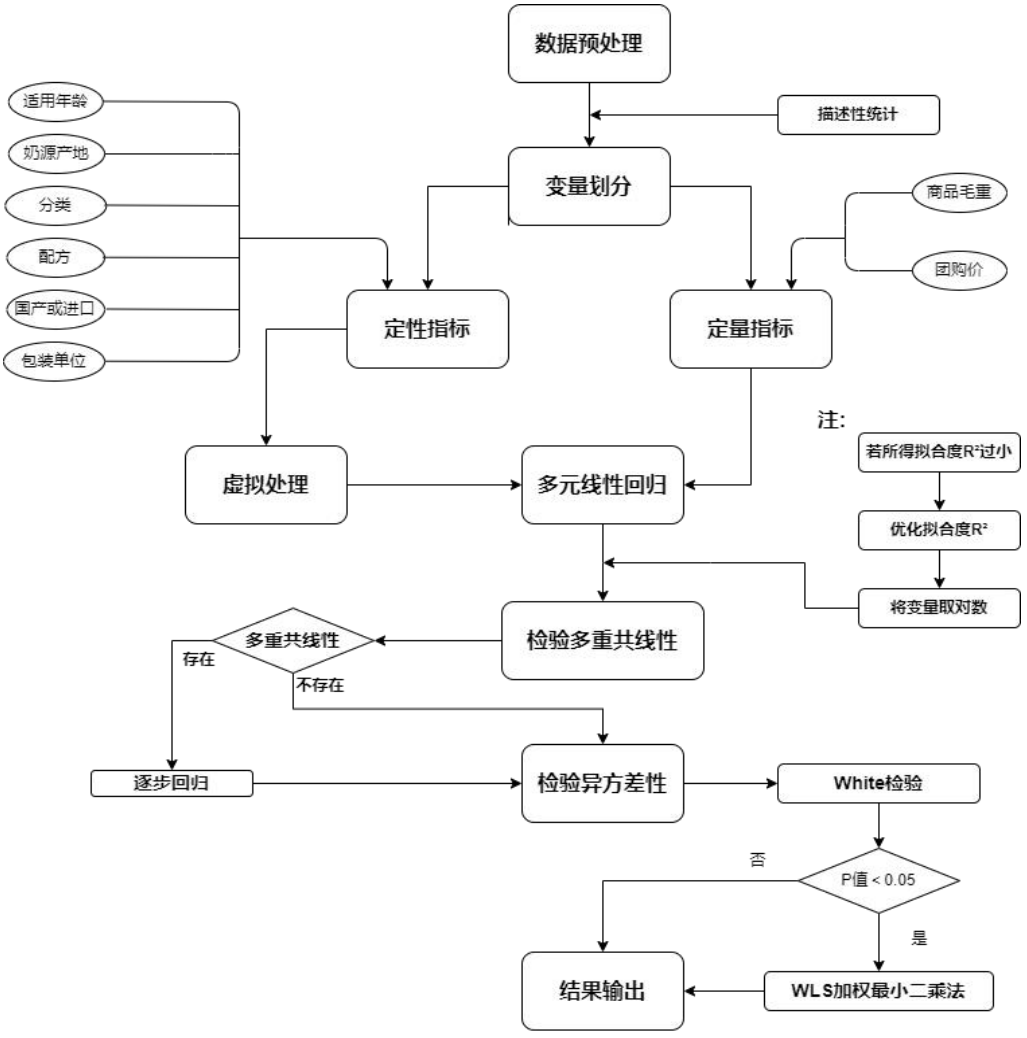


图 1 多元线性回归流程图

2.2 对于问题二的分析

该问题要求以评价量为因变量，研究影响评价量的重要因素。本题基于某电商平台提供的 846 条关于婴幼儿奶粉的销售信息，本题通过建立随机森林模型确定在以评价量（自变量），其余指标（自变量）为基础中的重要因素以及相应预测。

同问题一划分变量，将所得到的定量指标与虚拟转化好的定性指标导入数据集，将数据集划分成训练集（X-train）和测试集（Y-train），构建随机森林模型，然后通过基尼指数（Gini）进行特征重要性评分，再通过 K 折验证交叉法确定阈值，将所得到的阈值设置为评判阈值，最终确定重要因素。

本题通过随机森林模型确定本题所含关系里面的重要因素，并且可用于商家进行预测，以达到更好的销售前景。

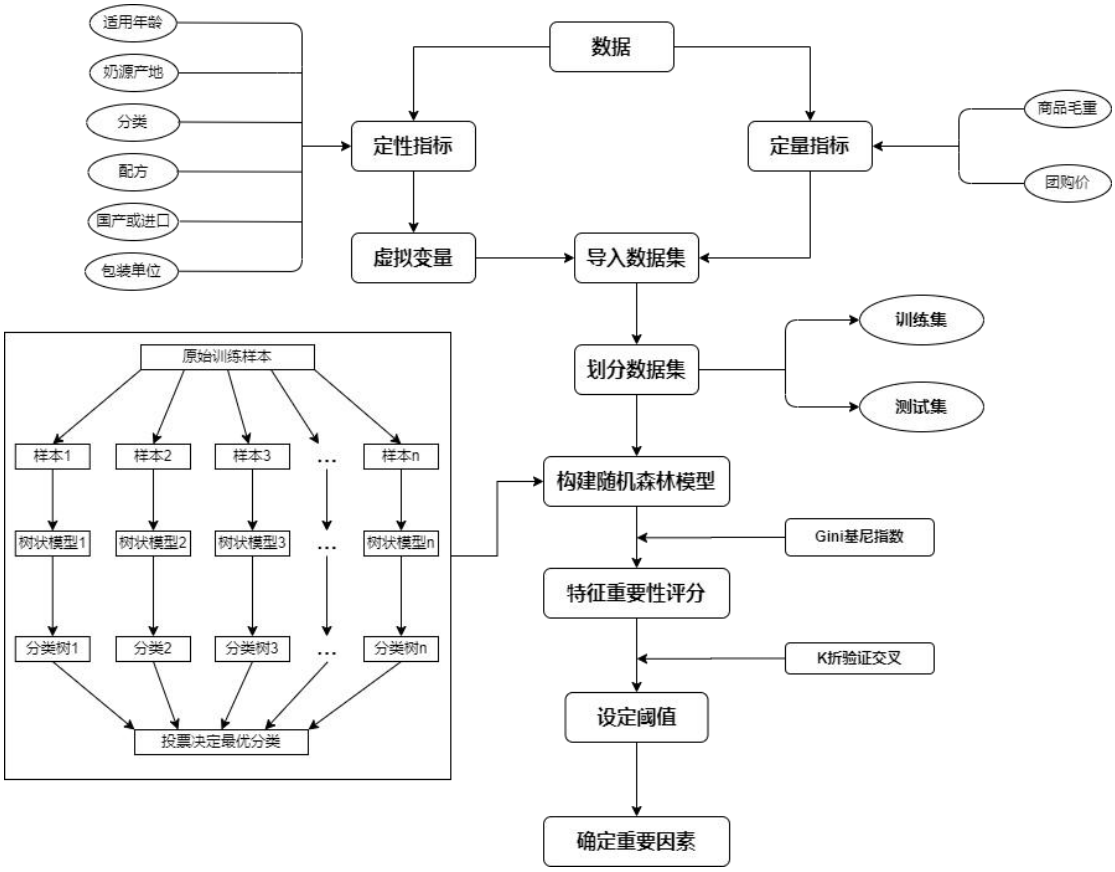


图 2 随机森林模型流程图

三、 基本假设与符号说明

3.1 基本假设

1、假设商品名称在本题中与评价量不存在线性关系。

2、假设将定性数据虚拟转化成数值型数据能进行处理并反映问题。

3、假设收集到的数据是准确、完整和可靠的，并且遵循了相关的数据收集准则和标准。

4、假设每个观测之间是相互独立的，即每个观测值不受其他观测值的影响。

5、假设每个婴幼儿的成长周期一致，并且营养需求一致。

3.2 符号说明

符号	说明
x_n	自变量指标（奶源产地、团购价、商品毛重、包装、配方、段位、国产与进口、分类）
Y	因变量（评价量）
CAC	国际食品法典委员会
VIF	方差膨胀因子
r	皮尔逊相关系数
n	样本数量
β	相关系数
WLS	加权最小二乘法
GI	基尼指数
$\hat{p}_m$	概率估计值
VIM	评分因子
E	评估指标

四、 数据预处理

4.1 变量说明

表 1 变量说明

变量类型		变量名称	说明
因变量		评价量	侧面反映顾客对产品的关注度
自变量	**定量指标**	商品毛重	数据值位于 0.12-8.64 之间
		团购价	数据值位于 9.9-2598 之间
	定性指标	商品名称	含有 84 个商品名称
		奶源产地	含有 9 个奶源产地（包含其它）
		国产或进口	仅有国产与进口 2 个数值
		适用年龄	含有 5 个年龄分段

	定性指标	包装单位	含有 4 种包装单位
		配方	含有 3 种配方
		分类	含有 2 种分类
		段位	含有 4 类段位

4.2 数据处理

4.2.1 段位数据处理

表 2 段位划分标准表

段位	1 段	2 段	3 段	4 段
中国标准	0-6 个月	6-12 个月	12-36 个月	36-72 个月
CAC 标准	0-12 个月	12-36 个月	\	\
其它	0-3 个月	3-6 个月	6-12 个月	12 个月以上

本题中按照中国标准进行数据处理，将不符合标准的样本信息剔除，并在后续操作中将段位数据与适用年龄视为同一指标。

4.2.2 商品名称数据处理

商品名称数量繁杂，且不存在其余指标均相同，最终评价量只受商品名称影响的情况，所以在本题中商品名称对于评价量的影响不计入模型建立，即剔除商品名称指标

4.2.3 定性数据虚拟化

将奶源产地、国产与进口、适用年龄、包装单位、配方以及分类，这六类定性指标，通过独热编码（one-hot encoding）转化成多维二进制向量。

4.3 描述性统计

对某电商平台提供的846条关于婴幼儿奶粉的销售信息中所含的定量数据进行描述性统计。

表 3 描述性统计表

变量名称	数据量	平均值	标准差	最小值	最大值
评价量	819	15718.72	72767.13	1	683009
团购价（元）	819	365.0736	372.7826	9.9	2598
商品毛重（kg）	819	1.058974	0.7700795	0.12	8.64

五、 问题一的模型建立与求解

5.1 模型的建立

以评价量为因变量，分析其它变量和评价量之间的关系，建立多元线性回归

模型，得到相关方程，通过对方程解释得到自变量与评价量之间的关系

5．1.1 数据预处理

（1）定量指标

商品毛重与团购价进行统计，得到相关数据频数直方图。由图可以看出商品毛重绝大多数存在的范围区间在 0-2kg 之间，团购价绝大多数设置在 0-500 价位之间

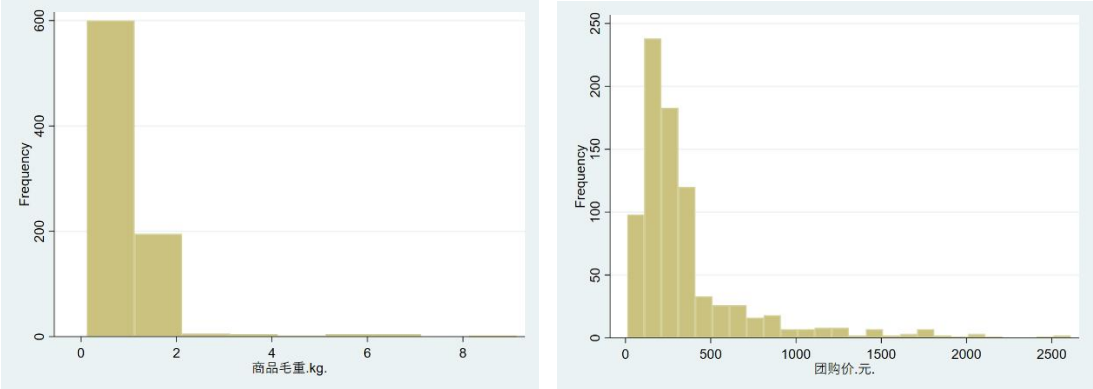


图 3 商品毛重频数直方图（左） 团购价频数直方图（右）

对商品毛重、团购价和评价量取对数，以提高稳定性。取对数可以提升数据的稳定性。特别是当数据的分布偏斜或存在异方差性（方差不恒定）时，取对数可以降低数据的变异性，使得数据更加平稳和符合线性回归的假设条件。将商品毛重和团购价分别对评价量进行回归分析，作出相关的回归直线散点图。由图可以得出，商品毛重过大或过小都只含有少量的评价量，当商品毛重在接近 1kg 范围里存在多数评价量。对于团购价，随着团购价的增长，评价量明显呈现负增长趋势，即团购价与评价量之间存在负相关性。

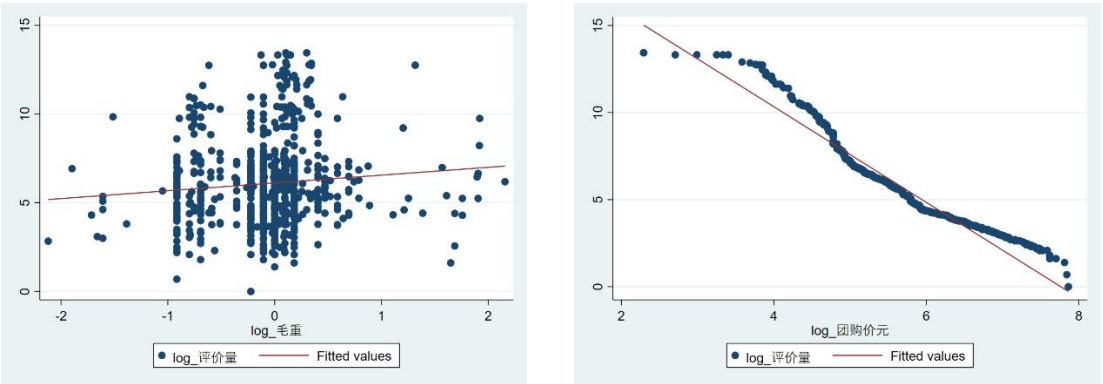
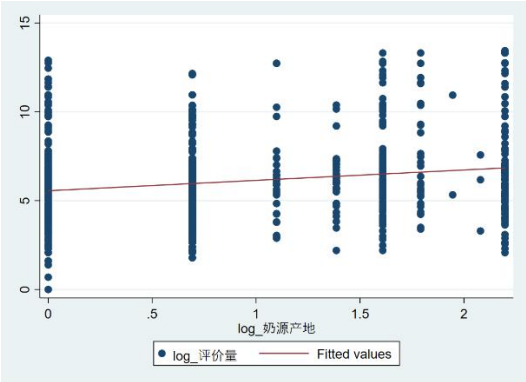


图 4 评价量和商品毛重的回归直线散点图 （左）
评价量和团购价的回归直线散点图（右）

（2）定性指标

将奶源产地、国产与进口、适用年龄、包装单位、配方以及分类，这六类定性指标，通过独热编码（one-hot encoding）转化成数值型数据，并相关统计作

图。



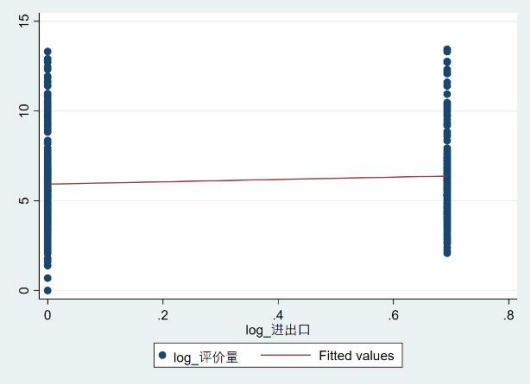
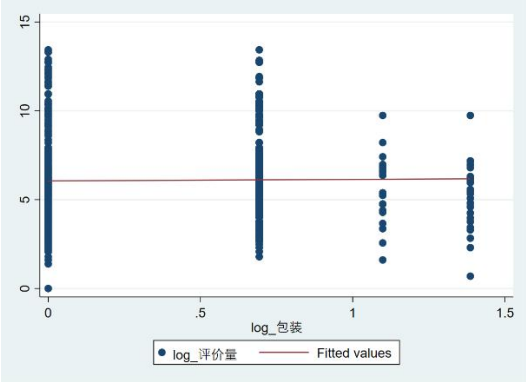


图 5 奶源产地（左） 国产与进口（右）



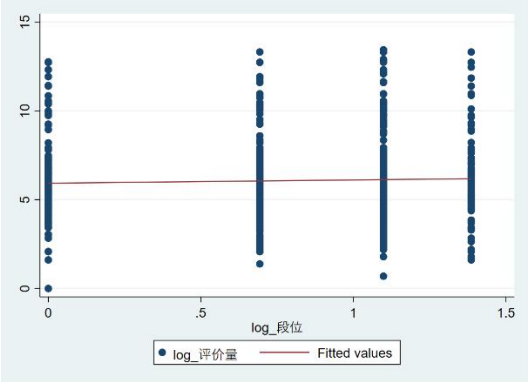


图 6 包装（左） 段位（右）

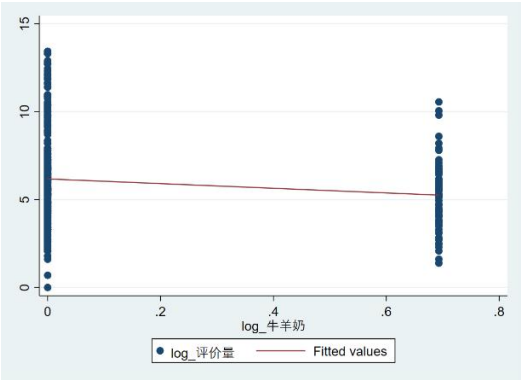


图 7 分类（左） 配方（右）

利用控制变量法，在分别对单一自变量同因变量，进行绘图分析，由图可知散点越集中的区域，其评价量越高。由于对定性指标的虚拟处理，图标出现明显的分段，每个分段代表一个定性指标的一个元素。包装为桶装、奶源产地为荷兰、分类为牛奶、配方为常规配方的婴幼儿奶粉评价量高。其中有两个因素所体现的评价量，没有明显区别，分别是国产或进口、段位。

5.2 模型的求解

5.2.1 多元线性回归

在方差分析表中，可以根据回归项的 F 值和 P 值来评估回归模型的显著

性，即自变量是否整体上显著影响因变量。如果回归项的 F 值较大且对应的 p 值较小，说明回归模型对解释因变量的变异性是显著的。反之，如果 P 值较大，说明回归模型不能显著解释因变量的变异性。

表 4 方差分析表（一）

方差来源	平方和	自由度	均方误差
回归	3.1487e+11	8	3.9359e+10
残差	4.0165e+12	810	4.9586e+09
总计	4.3314e+12	818	5.2951e+09

表 5 方差分析表（二）

数据量	F 值	P 值	R^2 值	均方根误差	判定系数
819	7.94	0.000	0.0635	70417	0.0727

所得拟合度指标 Adj R-squared=0.0635 远小于 0.9000，需要对其拟合度进行优化，将自变量和因变量进行同时取对数操作，重新进行回归，同时再进行多重共线性检验和异方差检验。

拟合度优化：

表 6 拟合度优化表

数据量	F 值	P 值	R^2 值	均方根误差	判定系数
819	1274.76	0.000	0.9257	0.63879	0.9264

对比表 6 与表 5 可知，R^2 值由 0.0635 提升至 0.9275 拟合度存在明显提升，且大于 0.9，需进行异方差存在检验。

5.2.2 多重共线性检验

（1）VIF 检验

VIF 用于判断自变量之间的相关性，特别是检测是否有相关性很高的变量导致模型的结果不稳定，对于线性回归模型中的每个自变量，VIF 可以通过计算该自变量与其他所有自变量之间的关系的平方来获得。VIF 越大，表示自变量与其他自变量之间的共线性越强，可能存在多重共线性问题。通常情况下，如果某个自变量的 VIF 大于一个阈值（通常为 5 或 10），则可以认为该自变量与其他自变量存在高度相关性，可能会影响模型的稳定性和可解释性。这时可以考虑采取一些处理方法，如删除冗余的自变量、合并相关的自变量或者进行正交化等，以解决多重共线性问题。

$$\mathrm{VIF_i}=\frac{1}{1-\mathrm{R}^2_{1-k/i}}$$

$\mathrm{R}^2_{1-k/i}$ 是将第 i 个自变量作为因变量，剩下的 $k-1$ 个自变量回归得到的拟合度。

$$\mathrm{VIF}=\max\left\{\mathrm{VIF_i}\right\}$$

若 VIF ＜ 10 则不存在严重的多重共线问题，反之则存在。

表 7 VIF 检验结果

变量名称	VIF	1/VIF
进出口	1.50	0.668178
奶源产地	1.32	0.755597
包装	1.17	0.853855
分类	1.08	0.925271
配方	1.08	0.928748
团购价	1.05	0.949537
商品毛重	1.05	0.956273
段位	1.03	0.972512
最大值	**1.50**	/

所含变量中，所有变量的 VIF 值都小于 10，，故变量之间不存在多重共线性。

（2）皮尔逊相关系数

皮尔逊相关系数（Pearson correlation coefficient），也称为皮尔逊积矩相关系数，是一种用于衡量两个连续变量之间线性关系强度和方向的统计量。它的取值范围在-1 到 1 之间。

通过计算两个变量之间的协方差和各自标准差的乘积来表示线性关系的强度

$$r = (\sum (^{(x-\bar{x})(y-\bar{y})}/_{(n*\sigma x*\sigma y)}))$$

其中，r 表示皮尔逊相关系数，X 和 Y 是两个变量的观测值，$\bar{X}$ 和 $\bar{Y}$ 分别是两个变量的均值，n 是样本数量，σX 和 σY 是两个变量的标准差。

表 8 皮尔逊系数相关矩阵表

	X2	X3	X1	X4	X5	X6	X7	X8
X2	1.00							
X3	-0.10	1.00						
X1	-0.19	0.14	1.00					
X4	-0.01	-0.07	-0.09	1.00				
X5	-0.01	-0.08	0.10	0.00	1.00			
X6	-0.02	0.06	-0.04	0.09	-0.10	1.00		
X7	-0.10	0.02	0.45	-0.31	0.23	-0.10	1.00	
X8	0.12	-0.04	-0.16	-0.13	-0.03	0.01	-0.15	1.00

相关系数的取值范围为 -1 到 1，数值越接近于 1 或 -1 表示变量之间的线性关系越强，而数值越接近于 0 则表示变量之间的线性关系越弱。

结合上面两个图表可以拒绝原假设，故变量之间存在多重共线性。

5.2.3 异方差检验

（1）White 检验

White 检验（White test）是一种针对线性回归模型中异方差性（heteroscedasticity）的检验方法，它是基于普通最小二乘法残差的方差与预测变量的关系。通过进行 White 检验，我们可以获得关于残差的异方差性信息，从而评估线性回归模型的合理性并采取相应的修正措施。

表 9 White 检验结果

源	卡方统计量	自由度	P 值
异方差性	324.32	2	0.0000
偏度	64.23	1	0.0000
峰度	22.57	1	0.0000
总体	411.12	4	0.0000

表中 p 值=0.000，强烈拒绝同方差的原假设，故存在异方差，利用加权最小二乘法（WLS）进行处理。

（2）WLS 加权最小二乘法

WLS 方法通过对观测值进行加权来解决异方差性的问题。具体而言，WLS 赋予误差项更大的权重，使得方差较小的观测值在估计过程中有更大的影响力，而方差较大的观测值则有较小的影响力。这样可以减小方差较大观测值对估计结果的扰动。WLS 方法可以提高回归模型的效果，并更准确地估计参数，常被用于处理异方差性问题。

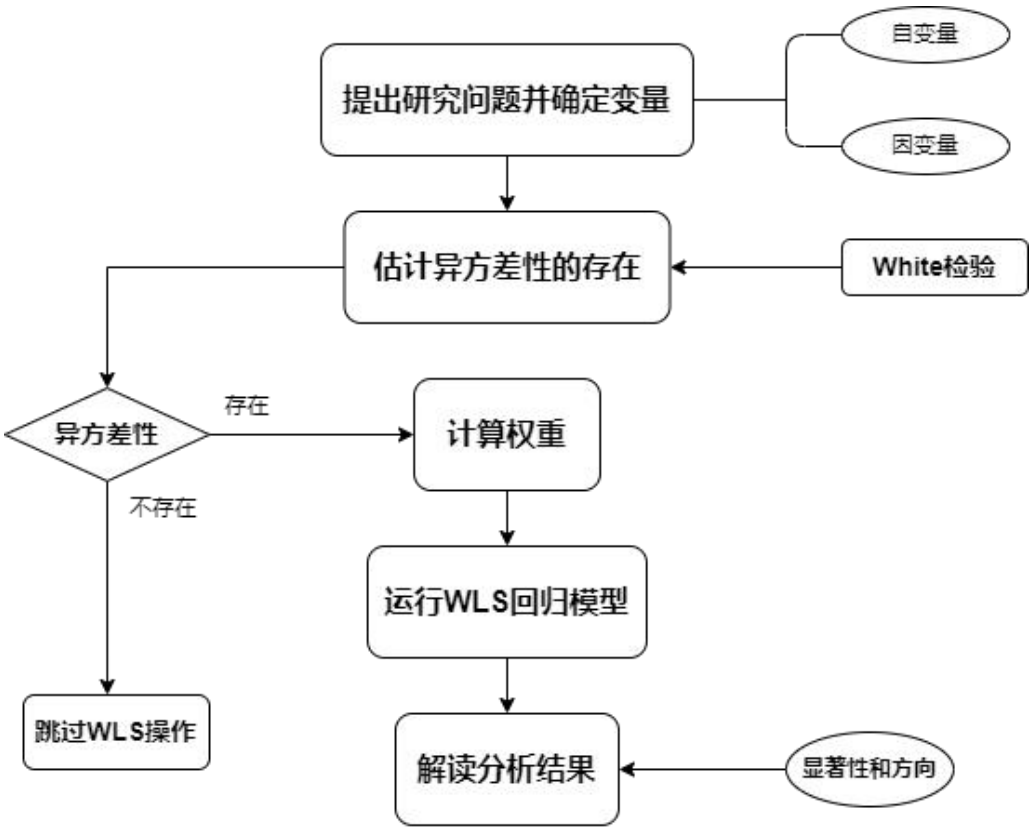


图 8 异方差检验流程图

表 10 经 WLS 后方差分析表（一）

方差来源	平方和	自由度	均方误差
回归	1033.33655	8	129.167869
残差	82.0243608	810	0.101264643
总计	1115.36992	818	1.3635219

表 11 经 WLS 后方差分析表（二）

数据量	F 值	P 值	R^2 值	均方根误差	判定系数

819	1275.54	0.000	0.9265	0.9257	0.31822

经过 WLS 加权最小二乘法处理后，均方误差和判定系数均得到明显优化。

表 12 经 WLS 回归后结果

评价量	系数	标准误差	T 值	P 值	Beta 值
奶源产地	0.1494301	0.020059	7.45	0.000	0.087304
团购价	-1.956599	0.0209993	-93.17	0.000	-0.931847
包装	-0.123997	0.0258693	-4.79	0.000	-0.0488832
段位	0.2328408	0.0262448	8.87	0.000	0.0897258
国产与进口	-0.140706	0.0507151	-2.77	0.006	-0.031943
分类	0.2014582	0.0425252	4.74	0.000	0.0473269
常数量	**16.30196**	**0.156912**	**103.86**	**0.000**	/

5.2.4 商品名称词云搜索

基于某电商平台提供的 846 条关于婴幼儿奶粉的销售信息，筛取并统计商品名称频次，通过词云搜索得到相关词云图。

图 9 词云分析图

由词云可得，宝妈们购买婴幼儿奶粉品牌更倾向于：惠氏、贝因美、飞鹤、君乐宝、合生元、美素、多美滋、澳优等品牌。

5.3 模型结果

原始方程：

$$\ln y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \cdots + \beta_k \ln x_k$$

回归后的方程：

$$\ln p = 16.30 + 0.15 \ln x_1 - 1.96 \ln x_2 - 0.12 \ln x_4 + 0.23 \ln x_6 - 0.14 \ln x_7 + 0.20 \ln x_8$$

图表中的数据可知拟合度R^2=0.926大于0.900,各自变量的p值均小于0.05，通过Beta值的绝对值可以看出自变量团购价对评价量的影响最大，团购价每上升1%评价量就对应下降1.96%，其次是段位和奶源产地

六、 问题二的模型建立与求解

6.1 模型的建立

以评价量为因变量，研究影响评价量的重要因素。本题通过建立随机森林模型，针对特征选择进行分析。随机森林能够评估特征的重要性，帮助我们选择最相关的特征。这对于数据分析和特征工程非常有用，可以帮助减少维度、降低复杂度，并找到对目标变量有最大影响力的特征。同时，随机森林也可以用于处理回归问题，即预测一个连续型的目标变量。本题要求以评价量为因变量，研究影响评价量的重要因素，即在评价量（因变量）和其余指标（自变量）之间，探究彼此存在的线性关系，再从中确定出重要影响因素。

5.1.1 数据预处理

同问题一，将数据划分为定量指标与定性指标， 定性指标转化为虚拟变量，再导入数据集。将数据集X和目标变量拆分为训练集X-train和测试集Y-train，将数据集中的20%分割为测试集，而剩下的80%将用作训练集。设置随机种子，以确保每次运行代码是都得到相同的拆分结果，以保证模型准确性。

6.2 模型的求解

6.2.1 随机森林模型

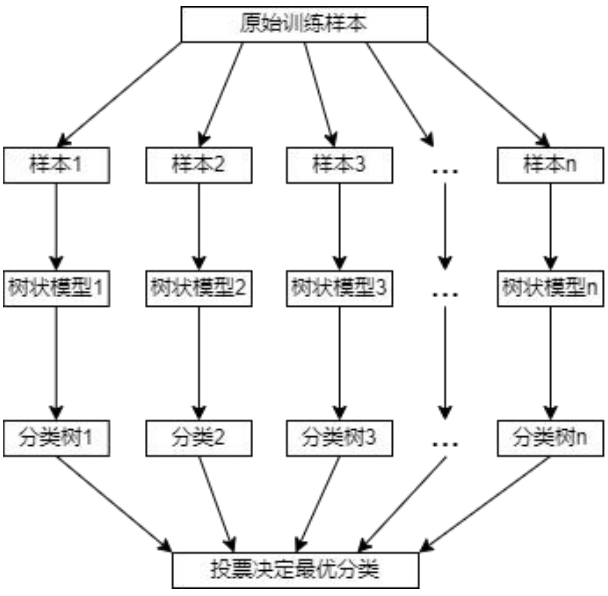


图 10 随机森林模型图

随机森林模型最终得到的结果取决于具体的应用场景和任务类型。本题中，随机森林模型可以提供以下几个结果：

（1）特征重要性：随机森林模型可以计算每个特征对最终预测结果的贡献程度，从而得到特征的重要性排名。这可以帮助我们了解哪些特征在模型中起到关键作用，从而进行特征选择或进行更深入的特征工程。

（2）预测结果：随机森林模型可以根据输入的特征向量进行预测，并给出一个或多个预测结果。预测结果的形式取决于具体的问题。

6.2.2 特征重要性评分

基尼指数：

基于基尼指数的变量重要性评分方法适用于分类问题，并且可以帮助我们理解每个特征对于预测结果的重要性程度。较高的变量重要性评分表示该特征对于预测结果的贡献更大。

Gini 指数的计算公式为

$$GI_{\mathrm{m}}=\sum_{k=1}^{K}\hat{p}_{mk}(1-\hat{p}_{mk})$$

$\hat{p}_m$ 为样本在节点 m 属于任意一类的概率估计值

变量 X_j 在节点 m 的重要性，即节点 m 分枝前后 Gini 指数变化量为:

$$VIM_{jm}^{(Gini)}=GI_m-GI_l-GI_r$$

GI_l 和 GI_r 分别表示由节点 m 分裂的两新节点的 Gini 指数

$$VIM_{ij}^{(Gini)}=\sum_{m=1}^{M}VIM_{jm}^{(Gini)}$$

变量 X_j 在第 i 棵树出现 M 次,则变量 X_j 在第 i 棵树的重要性为：

$$VIM_{ij}^{(Gini)}=\frac{1}{n}\sum_{i=1}^{n}VIM_{ij}^{(Gini)}$$

变量 X_j 在 RF 中 Gini 重要性定义为：

$$VIM_{j}^{(Gini)}=\frac{1}{n}\sum_{i=1}^{n}VIM_{ij}^{(Gini)}$$

其中，n 为 RF 中分类数的数量

表 13 特征值重要性评分表

排序	特征	重要性
1	团购价	9.964749e-01
2	段位_3 段	1.059679e-03
3	国产或进口_进口	9.352691e-04

4	段位_4 段	5.062788e-04
5	商品毛重	4.925371e-04
6	包装单位_盒装	2.158268e-04
7	奶源产地_澳洲/新西兰	83695163e-05
8	段位_2 段	7.914981e-05
9	奶源产地_爱尔兰	7.064934e-05
10	奶源产地_其它	4.023103e-05
11	奶源产地_荷兰	3.698725e-05
12	分类_羊奶粉	7.486454e-07
13	配方_有机奶粉	3.026807e-07
14	奶源产地_德国	2.808571e-07
15	奶源产地_法国	7.422742e-08
16	奶源产地_美国	6.369034e-08
17	包装单位_箱装	4.658896e-08
18	包装单位_袋装	4.191817e-08
19	奶源产地_英国	3.871964e-10
20	配方_特殊配方奶粉	1.747342e-11-

6.2.3 设定阈值

K 折验证交叉法：

在随机森林模型中，通过调整阈值可以改变模型预测结果的灵敏度和特异度。步骤如下：

（1）将数据集划分为训练集和测试集。

（2）对训练集进行 5 折交叉验证，将数据集划分为 5 个子集。

（3）对于每个阈值，循环进行以下步骤：

a、对于每个子集，训练随机森林模型并进行预测。

b、根据阈值将预测结果转化为分类标签。

c、计算每个子集的评估指标。

d、计算 5 个子集的平均交叉验证得分。

（4）选择得分最小的阈值作为最佳阈值。

需要注意的是，选择最佳阈值时，可以根据具体问题和需求来选择不同的评估指标

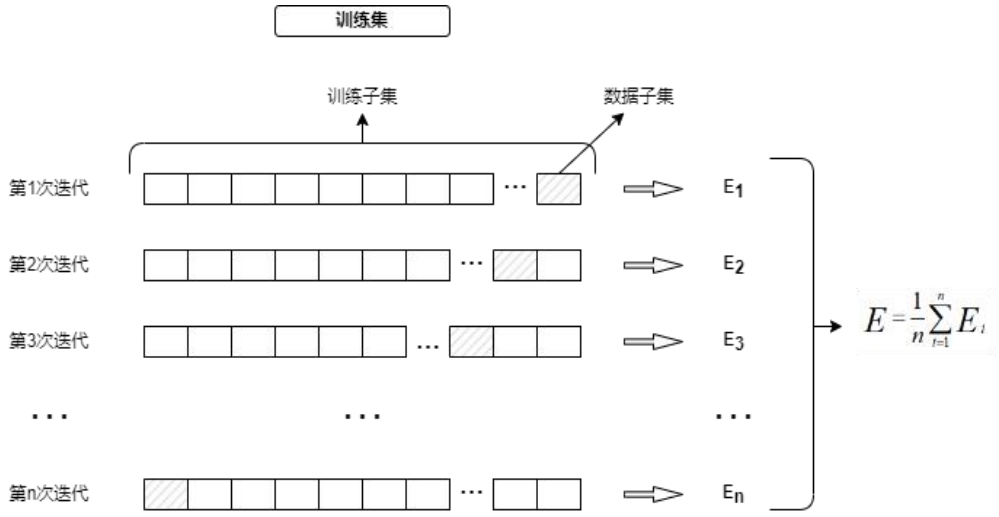


图 11 K 折验证交叉模型图

使用了 KFold 进行 5 折交叉验证，使用随机森林模型拟合数据，并计算每个阈值下的平均交叉验证得分。最后，找到得分最小的阈值，这被认为是最佳阈值。

最终所得最佳阈值：0.05，有且仅有团购价重要性大于该最佳阈值。

6.2.4 随机森林模型预测

随机森林模型将划分好的训练集进行多次数据处理，训练模拟得出模型，再对测试集进行测试，将测试结果与样本进行对比，最终得出预测结果，判定该模型是否能用于预测。在本题中，所得出的随机森林模型对于结果的预测具有相当优秀的可靠性。

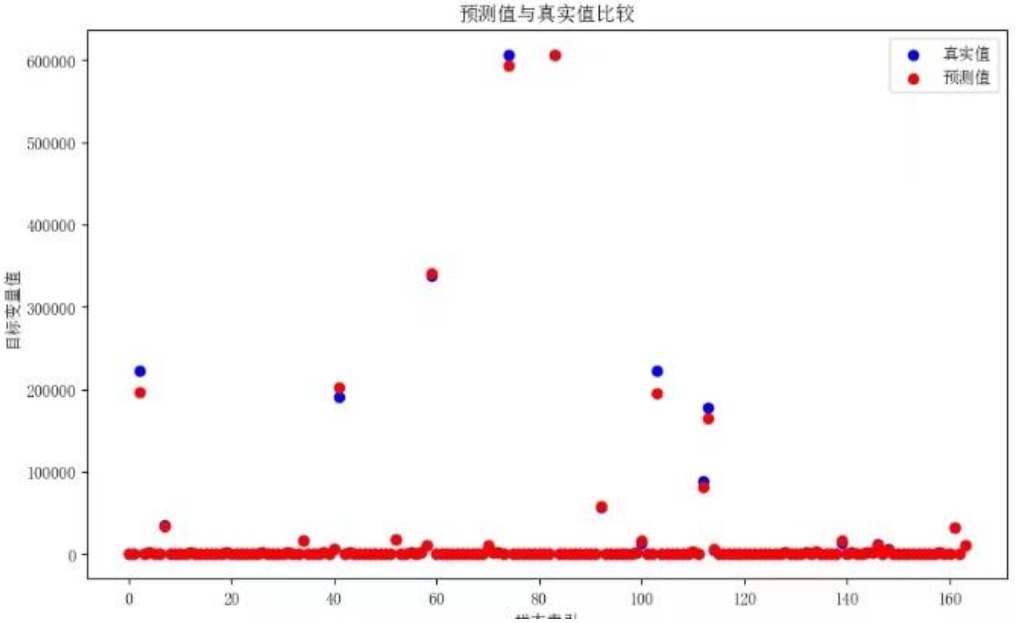


图 12 随机森林模型预测图

6.3 模型的结果

6.3.1 随机森林的结果分析

多元线性回归在数据量较大的样本处理中容易受共线性干扰，而随机森林模型由于随机性的引入，不容易陷入过拟合，具有一定的抗噪声能力，精确度较高，因此本题运用随机森林模型进行主要分析，多元线性回归作为对照。

根据指标重要性可得（见图），影响评价量的重要因素为：**团购价**。

6.3.2 两种模型的比较

由多元线性回归所得结果为团购价为重要因素，其次是段位和奶源产地，而随机森林所得结果为团购价为重要因素，且仅含有此一项为重要因素。相比之下，多元线性回归是根据各自变量对于因变量的相关系数的绝对值大小来确定重要因素，有大小区分，而随机森林是将多个因素进行重要性评分，最终通过求解阈值，并设定阈值，从而得出重要因素。两者都可得出各自变量之间的重要程度，但相比之下多元线性回归对于精确度稍缺，而随机森林模型能在多个因素中准确确定重要因素。

七、 模型的评价与改进方向

八、 参考文献

九、附录

9.1 程序代码

9.1.1 多元线性回归模型

```
# 描述性统计
sum 评价量 团购价元 商品毛重 kg
sum group1 group2 group3 group4 group5 group6

twoway (scatter log_评价量 log_奶源产地 )(lfit log_评价量 log_奶源产地 )
graph export "C:\Users\30408\Desktop\数学建模\day03\奶源产地散点、.png", as(png) name("Graph")
file C:\Users\30408\Desktop\数学建模\day03\奶源产地散点、.png saved as PNG format

twoway (scatter log_评价量  log_包装 )(lfit log_评价量 log_包装 )
graph export "C:\Users\30408\Desktop\数学建模\day03\包装散点图.png", as(png) name("Graph")
file C:\Users\30408\Desktop\数学建模\day03\包装散点图.png saved as PNG format

twoway (scatter log_评价量 log_配方 )(lfit log_评价量 log_配方 )
graph export "C:\Users\30408\Desktop\数学建模\day03\配方散点图.png",
```

```
as(png) name("Graph")
file C:\Users\30408\Desktop\数学建模\day03\配方散点图.png saved as PNG
format

twoway (scatter log_评价量 log_段位 )(lfit log_评价量 log_段位 )
graph export "C:\Users\30408\Desktop\数学建模\day03\段位散点图.png",
as(png) name("Graph")
file C:\Users\30408\Desktop\数学建模\day03\段位散点图.png saved as PNG
format

twoway (scatter log_评价量 log_进出口 )(lfit log_评价量 log_进出口 )
graph export "C:\Users\30408\Desktop\数学建模\day03\进口散点.png",
as(png) name("Graph")
file C:\Users\30408\Desktop\数学建模\day03\进口散点.png saved as PNG
format

twoway (scatter log_评价量 log_牛羊奶 )(lfit log_评价量 log_牛羊奶 )
graph export "C:\Users\30408\Desktop\数学建模\day03\牛羊奶散点.png",
as(png) name("Graph")
file C:\Users\30408\Desktop\数学建模\day03\牛羊奶散点.png saved as PNG
format

scatter log_评价量 log_团购价元

twoway (scatter log_评价量 log_团购价元 )(lfit log_评价量 log_团购价
元 )
graph save "Graph" "C:\Users\30408\Desktop\数学建模\day03\取对数的团购
价元的散点图+回归直线.gph"
file C:\Users\30408\Desktop\数学建模\day03\取对数的团购价元的散点图+
回归直线.gph saved

twoway (scatter log_评价量 log_毛重 )(lfit log_评价量 log_毛重 )
graph export "C:\Users\30408\Desktop\数学建模\day03\取对数的毛重散点直
线 图.png", as(png) name("Graph")
file C:\Users\30408\Desktop\数学建模\day03\取对数的毛重散点直线
图.png saved as PNG format

twoway (scatter log_评价量 log_团购价元 )(lfit log_评价量 log_团购价
元 )
graph export "C:\Users\30408\Desktop\数学建模\day03\取对数的团购价元的
散点直线图.png", as(png) name("Graph")
file C:\Users\30408\Desktop\数学建模\day03\取对数的团购价元的散点直线
图.png saved as PNG format
```

```
# VIF 检验
estat vif

# White 检验
estat imtest, white

# WLS 回归
cor log_评价量 log_团购价元 log_毛重 log_奶源产地 log_包装 log_配方
log_段位 log_进出口 log_牛羊奶
reg log_评价量 log_奶源产地 log_团购价元 log_毛重 log_包装 log_配方
log_段位 log_进出口 log_牛羊奶

predict yhat,xb
predict res,resid

gen res2=res^2
scatter res2 log_评价量
graph export "C:\Users\30408\Desktop\数学建模\day03\残差图.png",
as(png) name("Graph")
file C:\Users\30408\Desktop\数学建模\day03\残差图.png saved as PNG
format

reg res2 yhat
scatter res log_评价量
estat hettest,iid rhs
quietly reg log_评价量 log_奶源产地 log_团购价元 log_毛重 log_包装
log_配方 log_段位 log_进出口 log_牛羊奶
predict e1,residual

gen e2=e1^2
gen lne2=log(e2)
predict lne2f(option xb assumed; fitted values)

gen e2f=exp(lne2f)
reg log_评价量 log_奶源产地 log_团购价元 log_毛重 log_包装 log_配方
log_段位 log_进出口 log_牛羊奶 [aw=1/e2f]

# 回归分析
reg log_评价量 log_奶源产地 log_团购价元 log_毛重 log_包装 log_配方
log_段位 log_进出口 log_牛羊奶 [aw=1/e2f],b (sum of wgt is
24.3929860486817)
reg log_评价量 log_奶源产地 log_团购价元 log_毛重 log_包装 log_配方
log_段位 log_进出口 log_牛羊奶 [aw=1/e2f],r (sum of wgt is
24.3929860486817)
```

```
reg 评价量 团购价元 商品毛重 kg group1 group2 group3 group4 group5
group6
reg log_评价量 log_奶源产地 log_团购价元 log_包装 log_段位 log_进出口
log_牛羊奶
reg log_评价量 log_奶源产地 log_团购价元 log_包装 log_段位 log_进出口
log_牛羊奶 [aw=1/e2f],b (sum of wgt is 24.3929860486817)
```

9.1.2 随机森林模型

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import KFold, train_test_split
from sklearn.metrics import mean_squared_error

# 读取数据集并进行虚拟变量转换
data = pd.read_excel('E:\数学建模\day03\筛选后数据 3.xlsx')
qualitative_vars = ['奶源产地', '国产或进口', '包装单位', '配方', '分
类', '段位']

for var in qualitative_vars:
    dummy_cols = pd.get_dummies(data[var], prefix=var, drop_first=True)
    data = pd.concat([data, dummy_cols], axis=1)
    data.drop(var, axis=1, inplace=True)

# 构建随机森林模型，并获取特征重要性
X = data.drop('评价量', axis=1)
y = data['评价量']
rf = RandomForestRegressor(n_estimators=100, random_state=42)
rf.fit(X, y)
importance = rf.feature_importances_
feature_importance = pd.DataFrame({'特征': X.columns, '重要性':
importance})
feature_importance = feature_importance.sort_values('重要性',
ascending=False)
print(feature_importance)

# K 折交叉验证和模型训练
kfold = KFold(n_splits=5, shuffle=True, random_state=42)
rf = RandomForestRegressor(n_estimators=100, random_state=42)
scores = []

for train_index, val_index in kfold.split(X_selected):
```

```
    X_train, X_val = X_selected.iloc[train_index],
X_selected.iloc[val_index]
    y_train, y_val = y.iloc[train_index], y.iloc[val_index]

    rf.fit(X_train, y_train)
    y_pred = rf.predict(X_val)
    score = np.mean((y_pred - y_val) ** 2)
    scores.append(score)

avg_score = np.mean(scores)
print("平均均方误差:", avg_score)

# 特征选择
threshold = 0.05
selected_features = feature_importance[feature_importance['重要性'] >
threshold]['特征']
X_selected = X[selected_features]
print(X_selected.shape)

# 拆分数据集为训练集和测试集，进行预测和评估
X_train, X_test, y_train, y_test = train_test_split(X_selected, y,
test_size=0.2, random_state=42)
rf.fit(X_train, y_train)
y_pred = rf.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
print("均方误差:", mse)

# 可视化
plt.rcParams['font.family'] = 'SimSun'
plt.figure(figsize=(10, 6))
plt.scatter(range(len(y_test)), y_test, color='b', label='真实值')
plt.scatter(ran
```