## 问题1:故障自动识别模型的构建

**1. 数据预处理**

* 缺失值检测与数据类型查看：数据质量都相对完好，没有发现缺失值，所有的数据也都是int类型，不需要做特殊的处理。
* 统计各个变量以及查看分布情况：初步统计发现状态都是0与1的布尔类型值，而故障值虽然有变化，但所有的故障也都可以看成一个0 1分布，说明这是一个分类问题，并且出现故障的数量很少，同时也是一个类别不均衡数据。其他特征列都是一个累加的的数列。
* **异常值检测：** 首先绘制热力图，快速了解数据的总体分布情况，以及查看各个特征之间的相关性。再绘制箱形图查看是否有异常值，发现数据质量非常完好，几乎没有异常值的出现。

**箱形图**（也称为**箱型图**或**箱须图**）提供了连续变量分布的快照，尤其擅长于比较数据集中各组别的分布情况：

* 1. **数据集的总体情况观察：箱形图可以帮助你快速了解数据的总体分布情况。它以一种简化的格式显示大量的信息，包括数据的集中趋势、偏度和散布。**
  2. **比较不同组别间的数据分布：箱形图在比较不同组别间的数据分布时发挥了作用。其紧凑的设计提供了一个整洁的数据摘要，通过盒状和须状标记的定位，可以轻松地比较各组的分布特性。**
  3. **用途：**

**集中趋势：通过中位线和箱子的整体垂直位置，可以评估数据的集中趋势。**

**方差：盒子和箱须的长度代表数据的方差，越长表示方差越大。**

**偏度：通过中位数线在箱体和箱须中的位置，可以判断分布是否呈偏态。**

* 统计查看标签列的数据分布情况：通过绘制柱状图，可以清楚的看到每个标签的数据多少以及分布情况，为后面建模采取的采样措施选择方法奠定基础。

**2. 建模**

* 因为是多分类问题，面对多分类我们采取了三种建模策略（尝试后淘汰）：
  1. 将每个标签组合作为一个类别，转化为多分类问题：

优点：

能够直接训练数据，得到标签，简洁方便；

缺点：

多分类建模可采用的模型少，并且多分类组合类别数量多，模型复杂度高，训练模型速度慢，预测速度也慢。

尝试过程：采用的是决策树模型进行训练，绘制出混淆矩阵与查看评分指标时发现效果一般，特别是对于个别故障例如4003，故障的召回率虽然高，但是准确率相对低，只有0.76。

* 1. 先转换为二分类再聚类（尝试后淘汰）：

优点：

在0，1训练预测得到的结果非常好，并且因为前者是二分类问题，后者又是聚类问题，两个方向的模型训练速度都比较快；

缺点：

在聚类时效果不理想，并且难以对聚类的故障进行划分与原先故障对应。

尝试过程：将多分类先转化为二分类问题，将所有故障都归到一个列中，并将所有出现的故障的值都变为1，没出现故障则为0。将所有数据都整理完后，通过模型训练完该数据后，测试集发现评分指标接进1，说明效果很好。这样在预测后可以得到所有出现故障1的样本，再通过聚类的方法，将所有的故障重新转换为多个故障类。但是通过单独拿出一部分的故障1类样本进行聚类，前后使用了KMeans模型、层次聚类模型发现效果均不理想，与故障1类的样本偏差很大，说明效果并不好。

* 1. 将多标签问题转化为多个二分类问题，每个标签对应一个二分类器（最终采纳的方法）。

优点：

训练速度相对较快，模型得到的结果也十分理想；

缺点：

步骤相对繁琐，特别是处理数据方面。

* 采样处理：

由于是类别不均衡问题，所以肯定设计建模前的数据采样方法，目前主要有两种方式，上采样和下采样。

上采样：又称插值采样，是增加数据点数量的处理方法，用于增加数据集的大小或提升信号的采样率。

下采样：又称减采样，是一种减少数据点数量的处理方法。主要目的是减少数据集大小或降低信号的采样率。在执行下采样时，通常会舍弃一些数据点或通过某种方法（如取平均、采取每个第n个数据点等）缩减数据集的规模。

由于有十条生产线的数据，每条生产线有大于700w条的数据，，尽管上采样能够更好的保留多类样本的特性，并且增加少类样本达到均衡，但是由于过大的数据量，将会导致模型复杂度倍增，对计算机的性能要求也大幅提升，由于设备条件的欠缺，所以舍弃上采样的方法。而对于下采样，尽管会损失一些多类样本的特性，但是用下采样的方法在该数据集的训练和测试发现，效果依然很好，所以最终采用下采样的方式。

* 模型选择：

目前把该问题当作一个二分类问题处理，所以模型也选择一些分类模型参考。参考过的模型有随机森林，逻辑回归，支持向量机，XGBoost，梯度提升以LSTM神经网络等一系列模型，通过对单个模型的调参与测试后发现：

支持向量机在故障1001的拟合速度很快，但是却在故障2001的拟合速度很慢，甚至感觉难以收敛，训练时间超过其他模型的几倍，所以舍弃支持向量机。

LSTM是一种强大的循环神经网络，能够利用之前的时间序列数据对后面的时候序列数据进行预测，但是训练时发现训练时间太长，并且LSTM的数据处理也与一般的机器学习模型数据处理方式不同，不利于对于最后的集成模型，所以也舍弃。

**随机森林**是一种集成学习方法，集成多棵决策树来提高预测性能。在随机森林的测试中，学习速度快，调参简单，并且测试效果优秀，所以采用该模型。

XGBoost是一种优化的分布式梯度提升库，广泛的应用于数据挖掘、推荐系统等领域。尽管在其算法的超参数较多，并且需要多个尝试，但是发现测试结果好，并且运行速度较快，所以也采用了该模型。

**梯度提升树 (GBT)** 是一种基于梯度下降的增量学习方法，通过迭代地构建多个决策树来逐步改善模型的性能。该模型的超参数也较多，需要进行不断的尝试，训练速度略慢于其他模型，但训练效果也相当不错，所以采用该模型。

逻辑回归是是一种广泛应用于分类问题的监督学习算法，该模型经典十分经典，测试该模型效果时发现也相当不错，并且训练速度快，所以采用该模型。

* 模型建立：

最终决定采用随机森林，逻辑回归，XGBoost，梯度提升四个模型。并将四个模型放到一起组成多数投票的集成组合模型，从而避免单个模型的偏差。通过四个模型的训练，先将所有的处理好的数据划分训练集与测试集，进行一次训练和测试，进一步查看组合模型的性能，结果发现十分可观。有个小细节需要注意，最后需要再将模型应用到所有处理好的数据集中进行训练，然后将每个故障模型都保存下来准备应用到问题2的数据中。

**主要实验结果呈现：**

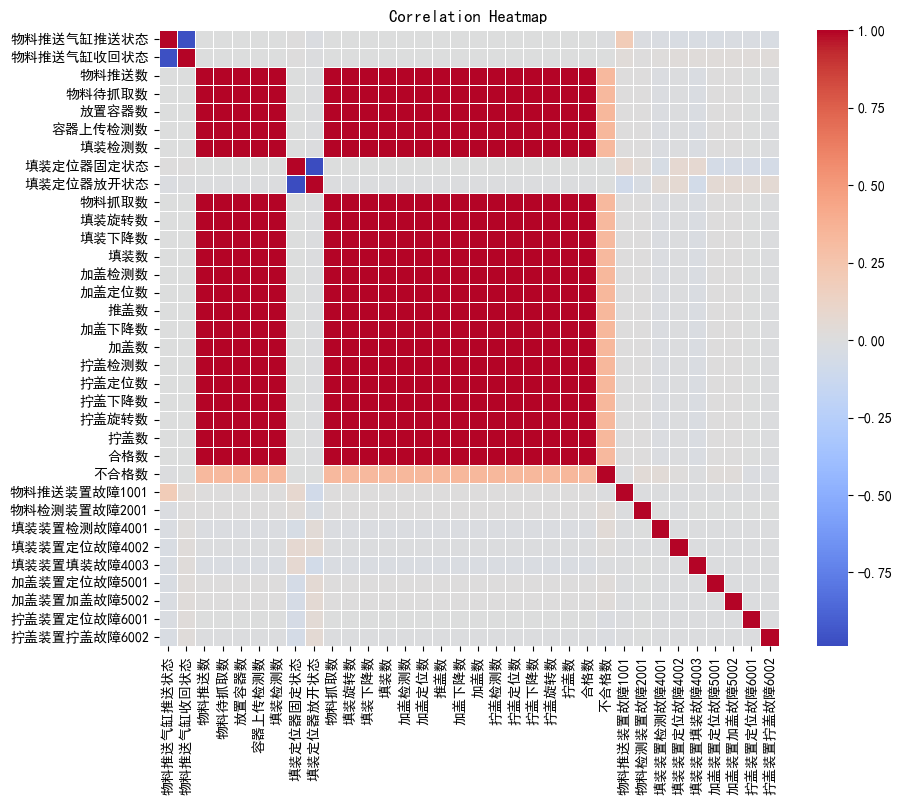
**数据预处理：**

缺失值数量统计：

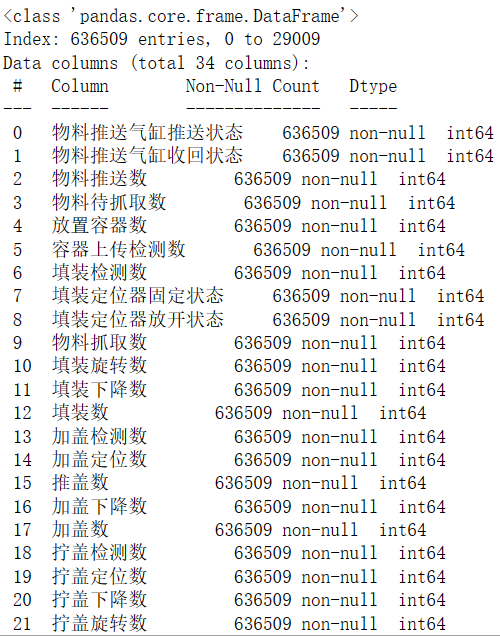
图形用户界面, 文本, 应用程序

描述已自动生成

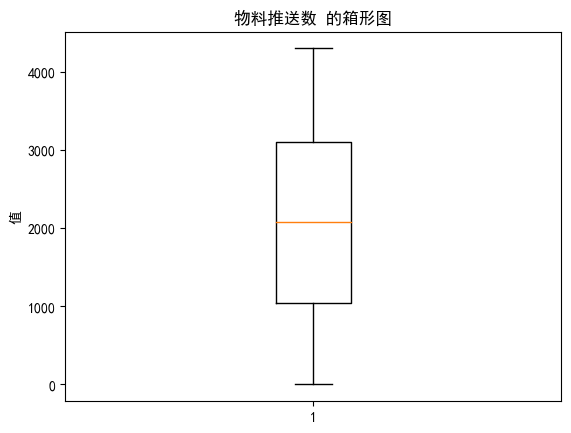
热力图：

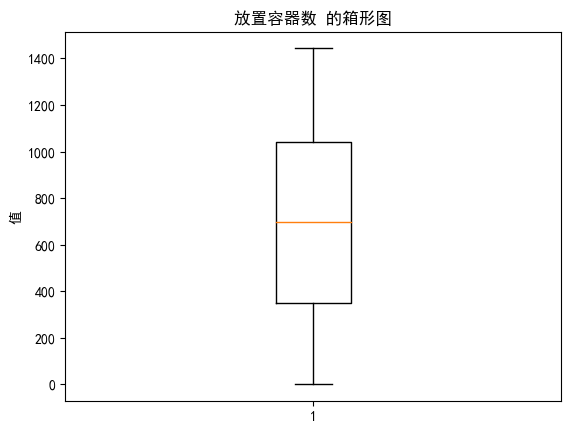


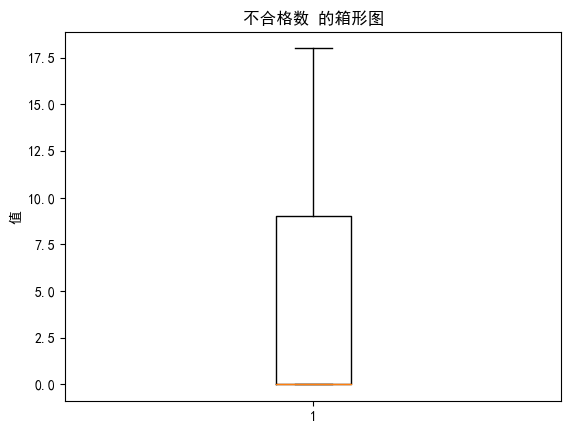
数据类型查看：



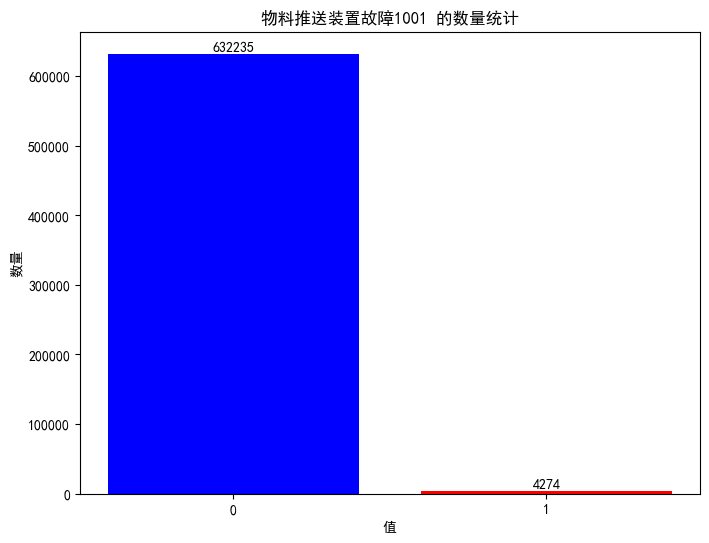
部份特征箱形图：

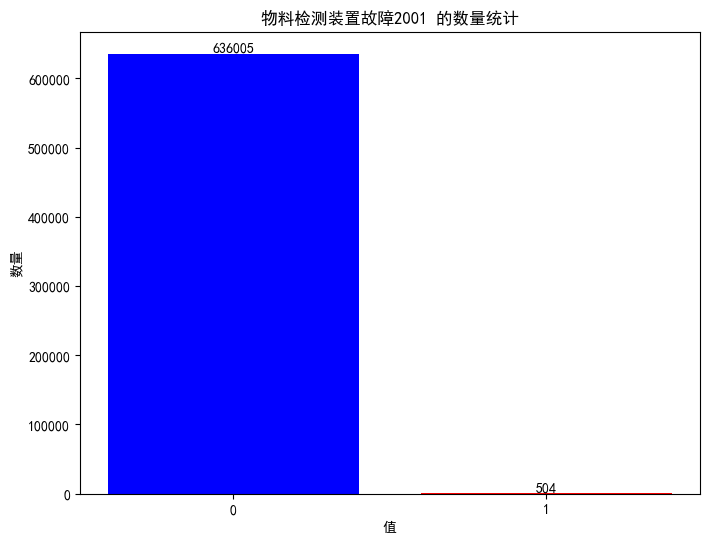


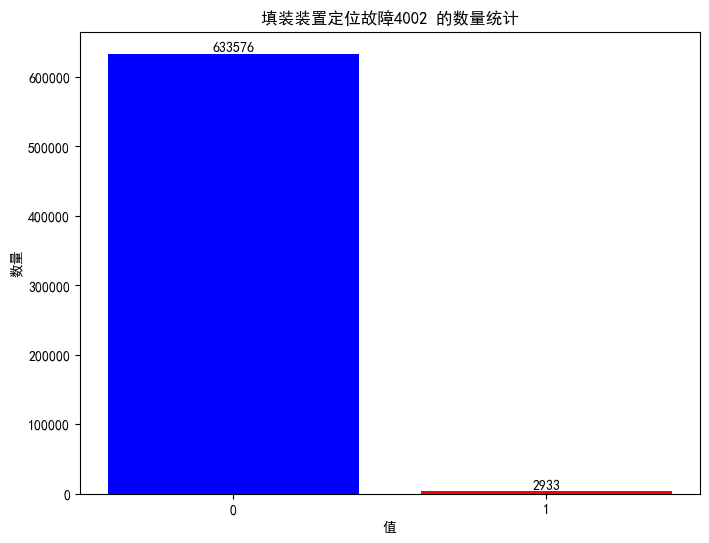




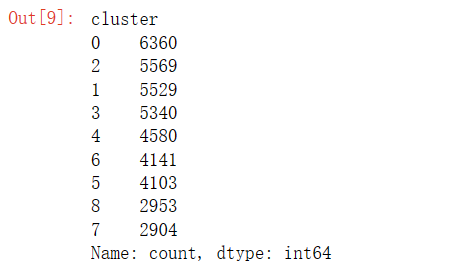
部份故障数量统计：







聚类模型结果：



下采样：

手机屏幕的截图

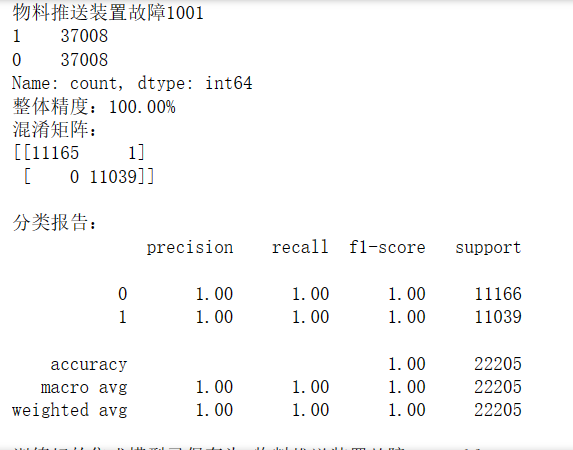
描述已自动生成

一对多逻辑回归模型：

表格

描述已自动生成

集成模型部份结果：

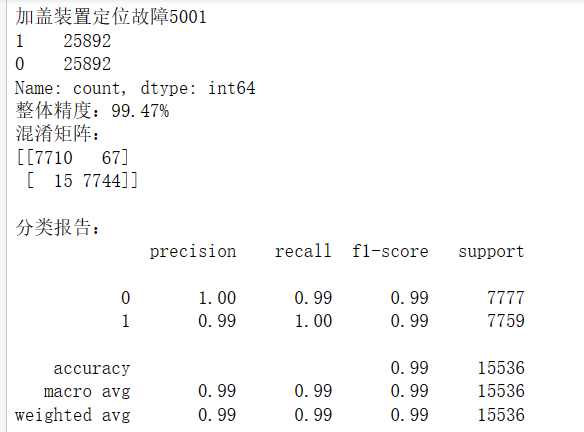


手机屏幕截图

描述已自动生成

手机屏幕截图

描述已自动生成



## 问题2:故障自动报警的实现及结果分析

## 问题3:产品产量和合格率影响因素分析

## 问题4:最佳操作人员排班方案及相关结果分析