问题1解决步骤：

1. 数据预处理

异常值检测

分析变量之间的关系

统计故障数量

1. 建模

建模思路：1.这是一个多分类问题，处理多分类想到三个处理方法：

1. 直接进行多分类建模，优点是能够直接训练数据，得到标签，简洁方便；缺点是训练效果不好，训练速度慢，且无法应用多个模型建模，不方便对新数据进行预测。尝试后淘汰
2. 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。通过模型训练完该数据后，预测时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。优点是在0，1训练预测得到的结果非常好，并且训练速度快；缺点是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。尝试后淘汰
3. 将每个故障分离出来训练，转换成多个2分类问题，为每个故障都建立一个单独的模型。优点是训练速度相对较快，模型得到的结果也十分理想；缺点是步骤相对繁琐，特别是处理数据方面，但最后鉴于总体效率和效果还是选择该方法。

处理数据思路：因为有10条生产线的数据，并且每条生产线有9个故障，所以要把十条生产线的故障都进行采样，并且将相同故障数据放到同一个表中。所以最后分为了9个故障表，每个故障表都有十条生产线的数据。

采样思路：由于数据量实在是过于庞大，每条生产线的数据都有大于700w条的内容，如果采用上采样的方法，一般的用户电脑性能远远不够。再通过对下采样方法的尝试后发现，数据量虽然大大减少，但是训练出来的模型效果也十分好，所以直接采用下采样的方式。

详细建模流程：这是一个分类问题，采用的也都是经典的分类模型。

前后采用过随机森林，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，最后通过对单个模型的性能测试以及参数的调试，最终决定采用随机森林，逻辑回归，XGBoost，梯度提升四个模型。

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。最后再将模型应用到所有处理好的数据集中，并且将每个故障模型都保存下来准备应用到问题2的数据中。