论文写作

Q2

对于一个给定的未来解决方案词，在未来的日期，开发一个模型，使您可以预测报告结果的分布。换句话说，预测未来日期(1,2,3,4,5,6,X)的相关百分比。你的模型和预测有哪些不确定性?请给出一个具体的例子，说明你对2023年3月1日“EERIE”一词的预测。你对模型的预测有多大信心?

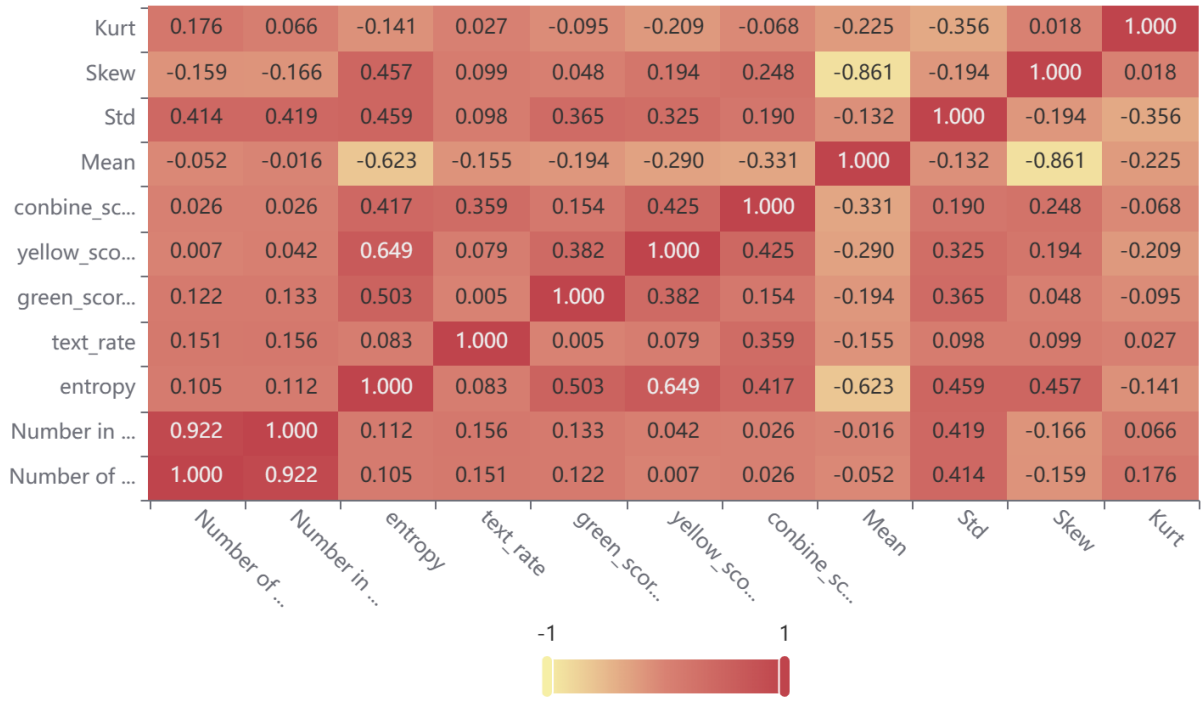
问题分析

1. 开发模型对未来日期 未来方案词（具体表述看英文问题）预测报告结果的分布
2. 模型预测的不确定性
3. 模型预测的信心

第二问我们挖掘了更多数据的属性1 2 3 4（加上四个属性）

并且我们发现，题目给定的数据集分布，也就是未来日期(1,2,3,4,5,6,X)的相关百分比，是一个定序数据，均有一定的相关性，所以我们使用xxx方法（问++），将相关百分比解释为均值，标准差，偏度，与峰度四个量。这四个值的大致分布如下（2*2的图像，找++要四个值的分布）

我们通过相关性分析，发现**偏度与峰度**与单词的属性基本没有相关性（给数据（表格形式），++也做了这个数据，给出热力图如下，SPSS pro可以去做），并且**均值**与**标准差**相关性很低可以作为独立的两个指标去分析。



对于时间方向的考虑，由于第一问我们使用ARIMA模型给出了时间与报告人数的关系，所以在这一问，我们用报告人数代替时间去影响报告结果的分布。

于是，通过随机森林回归模型分进行回归，我们用五个维度的指标（1234+人数）去映射**均值与标准差**这两个指标，将**均值与标准差**拟合分布函数（正态or偏态，问++）去预测报告结果的分布。

模型原理

随机森林回归是一种集成学习技术，用于解决回归问题。其主要原理是在输入数据的随机子集上建立多个决策树，并对这些树的输出进行聚合，以得出最终的预测结果。

1. 随机采样：在随机森林回归中，训练数据被随机采样并进行替换，从而创建多个自举样本。这意味着随机森林中的每个树都是在不同的数据子集上进行训练的。
2. 随机特征选择：对于每棵决策树，会随机选择一部分特征来对节点进行分割。这有助于减少树之间的相关性，并防止过拟合，从而确保没有单个特征支配了模型。
3. 决策树构建：随机森林中的决策树是使用递归过程构建的，基于选定的特征来分割数据。每棵树的生长深度通常是一个最大深度或当叶节点中的实例数低于某个阈值时停止生长。
4. 预测结果的聚合：一旦所有的决策树都构建完成，它们的预测结果将被聚合以得出最终的预测结果。对于回归问题，需要对森林中所有树的预测结果进行平均处理。
5. 超参数调整：随机森林回归具有几个超参数，可以通过调整来改善性能，例如树的数量、树的最大深度以及在每个节点上选择的特征数量等。这些超参数可以使用交叉验证等技术进行调整，以找到最佳的参数设置。

上面用chatGPT搜的，把步骤部分做到流程表中就行了

随机能够处理很高维度的数据，并且不用做特征选择(因为特征子集是随机选择的)，所以对于我们多个特征，随机森林模型的处理效果很好。并且为了更好的观察不确定性，随机森林能够给出哪些特征比较重要，改变重要特征更容易让模型受到扰动，以确定模型的灵敏度与不确定性。

模型假设

1. 假设时间参数只会影响每日的报告人数，对题目难度，参加游戏人群素质和累积的游戏经验无关。
2. 假设报告分布的标准差与峰度和单词的属性相互独立。
3. 假设偶然性造成的噪声不会掩盖均值与偏度的特征

主要模型

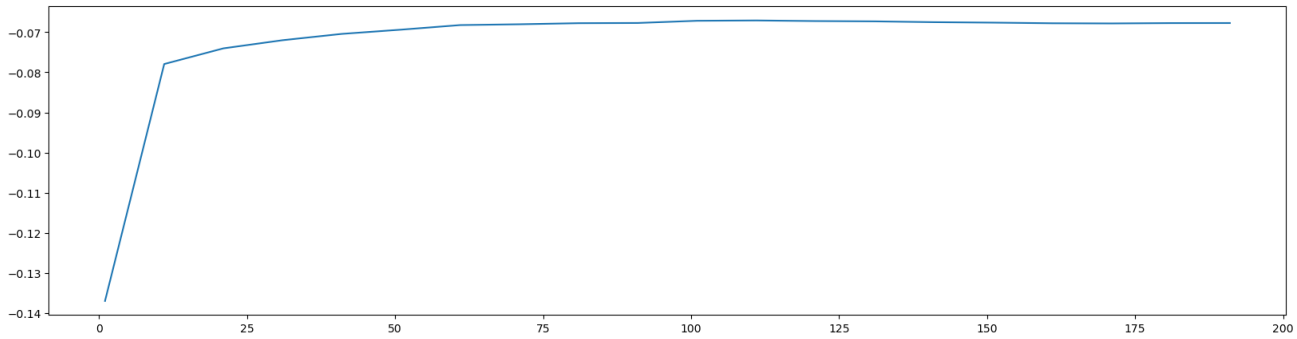
随机森林回归模型

先对**均值**分析

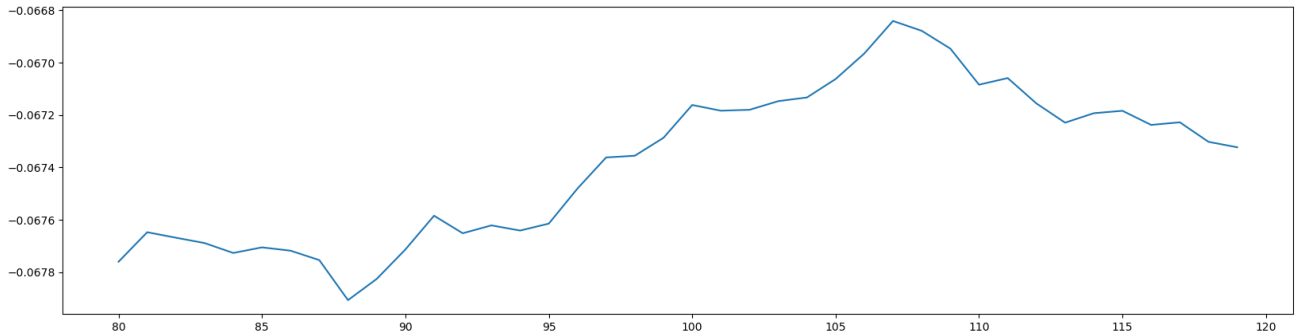
1. 随机森林调参

   通过网格搜索，学习曲线等确定随机森林的最佳参数

   大纵深学习曲线



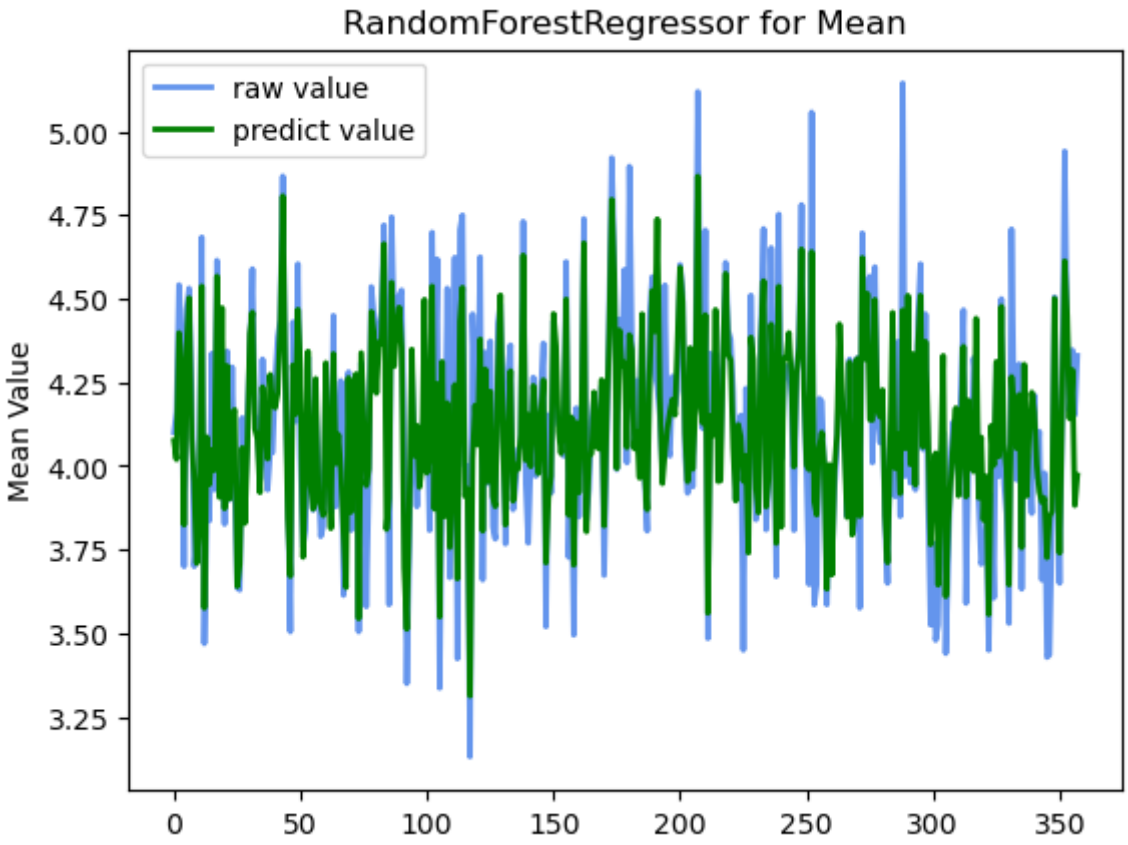
   小范围学习曲线



参数	参数值
`n_estimators0	107
`max_depth`	11
`min_samples_leaf`	None
`min_samples_split`	None

2. 随机森林回归

   回归图像



对于训练集的回归相关系数（R2）： 0.905

对于测试集的回归相关系数（R2）： 0.398

3. 特征的重要程度

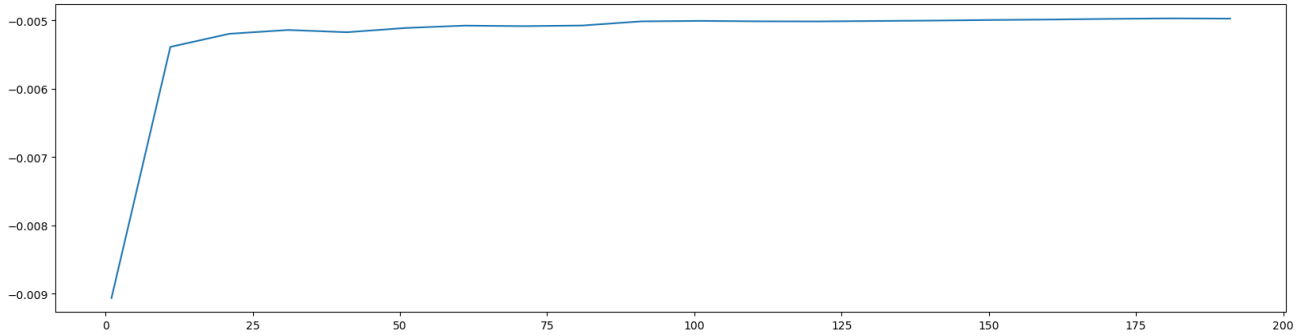
特征名称	重要程度 (%)
Number of reported results	7.52
entropy	61.67
text_rate	9.09
green_score	6.66
yellow_score	12.17
conbine_score	2.90

再对**标准差**分析

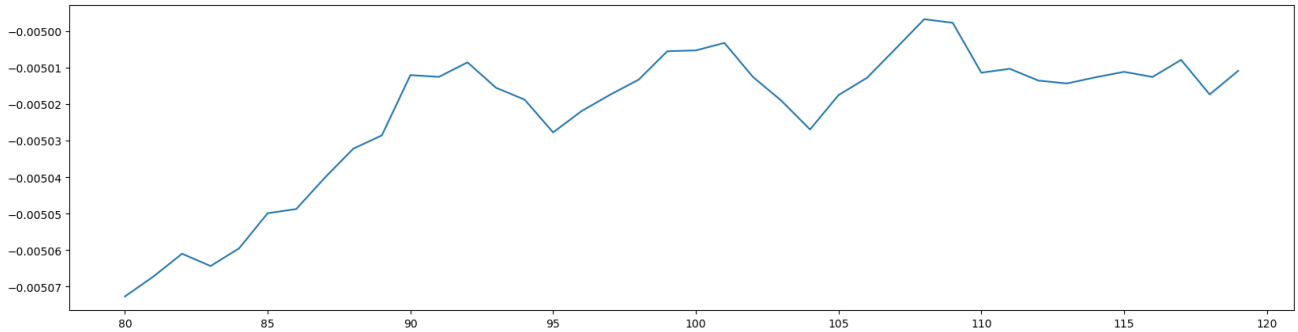
1. 随机森林调参

通过网格搜索，学习曲线等确定随机森林的最佳参数

大纵深学习曲线2



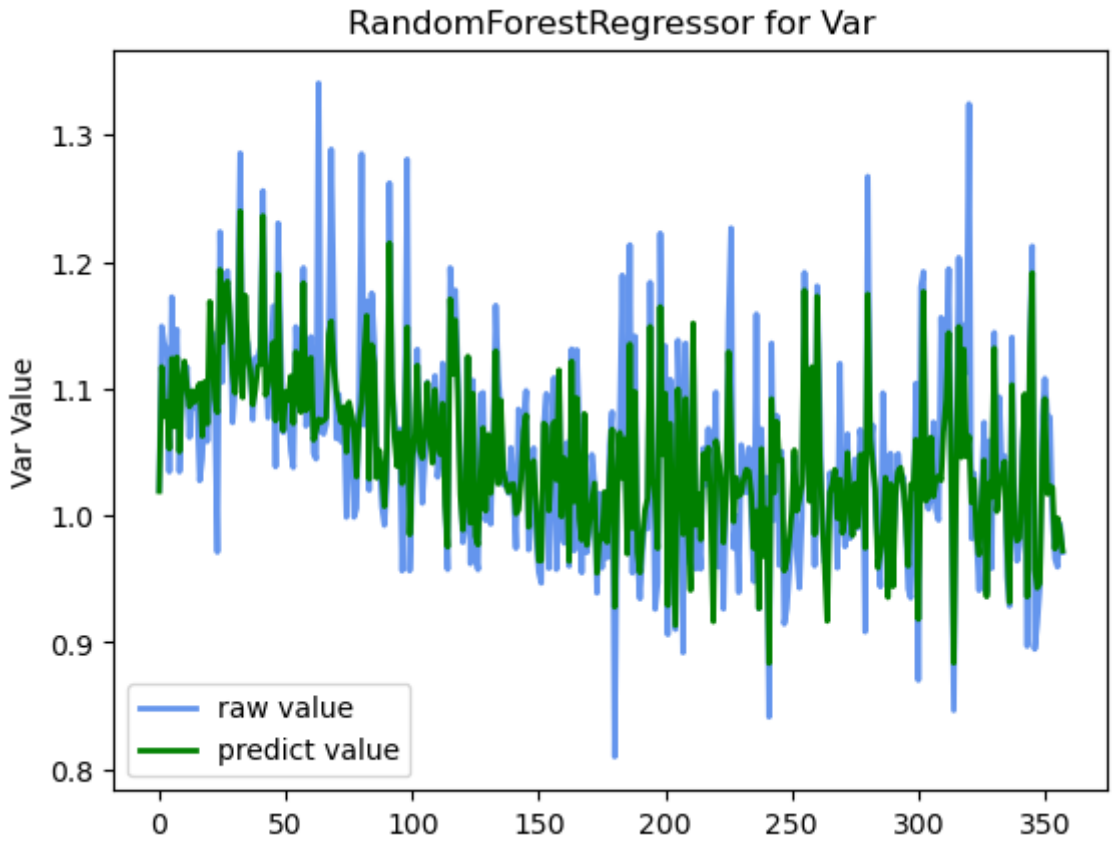
小范围学习曲线2



参数	参数值
n_estimators	108
max_depth	None
min_samples_leaf	None
min_samples_split	None

2. 随机森林回归

回归图像



对于训练集的回归相关系数（R2）： 0.908

对于测试集的回归相关系数（R2）： 0.357

3. 特征的重要程度

特征名称	重要程度 (%)
Number of reported results	31.37
entropy	24.65
text_rate	8.74
green_score	16.82
yellow_score	10.16
conbine_score	8.24

预测分析

EERIE单词属性及预测人数

特征名称	值
Number of reported results	20439(Q1回归树分析预测值)
entropy	9454.13
text_rate	9.856e-07
green_score	0.4573
yellow_score	0.4553
conbine_score	2.78

预测值:

Mean: 4.32138891

Std: 1.0459