论文写作

Q1

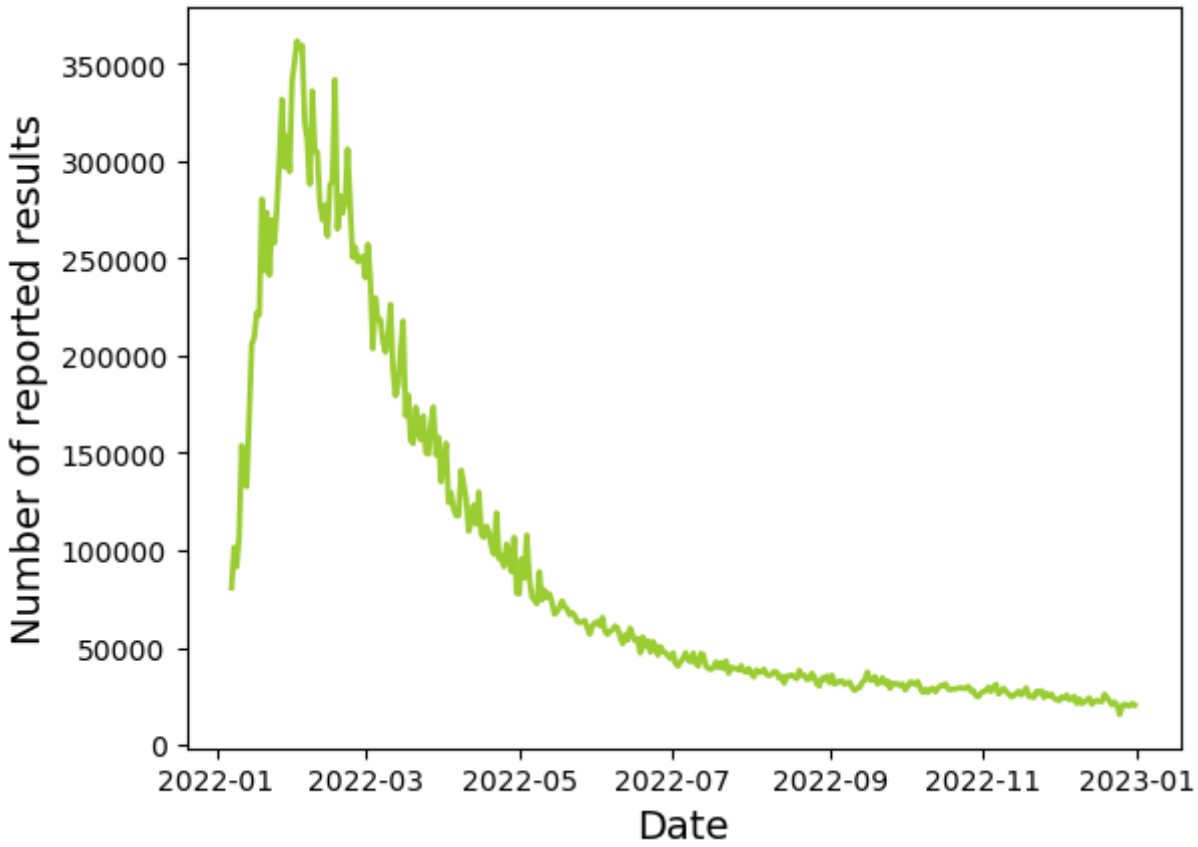
The number of reported results vary daily. Develop a model to explain this variation and use your model to create a prediction interval for the number of reported results on March 1, 2023.

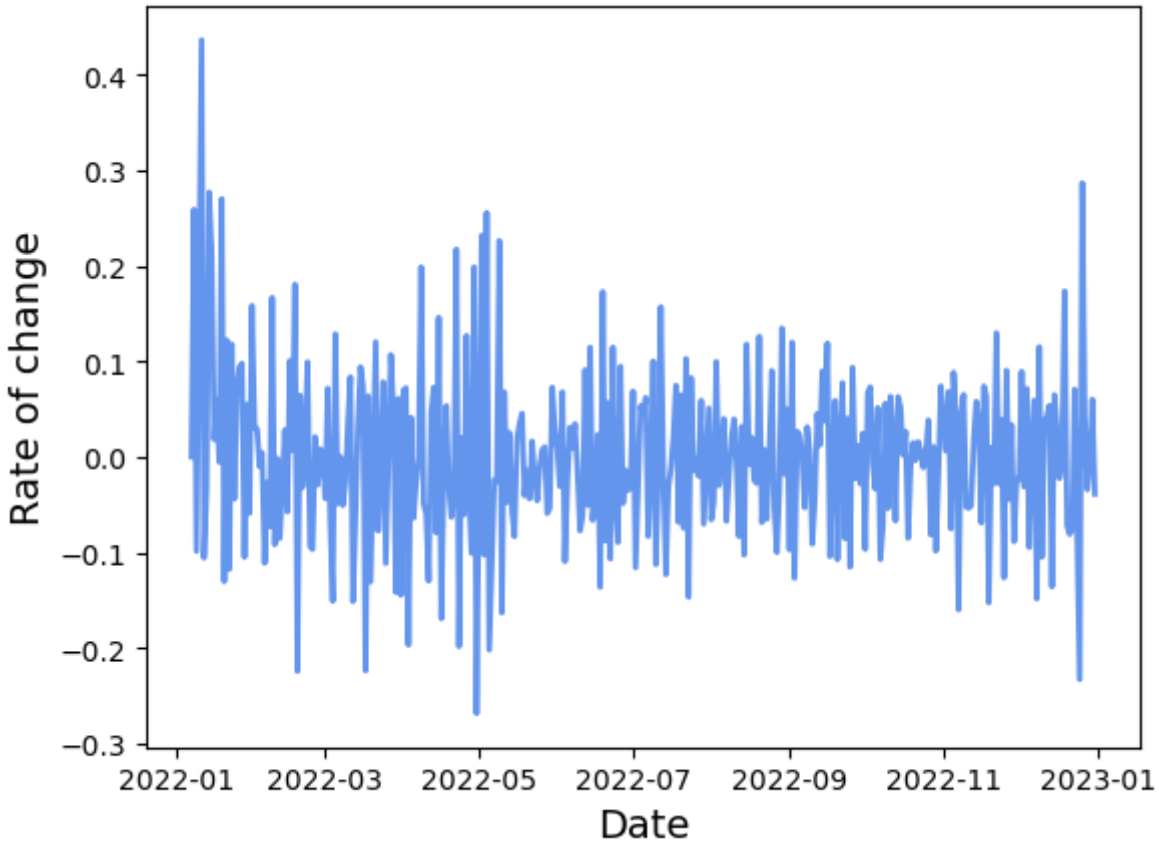
报告结果的数量每天都在变化。开发一个模型来解释这种变化，并使用您的模型为2023年3月1日报告的结果数量创建一个预测区间。

问题分析

1. 开发模型用于解释变化 (回归拟合)
2. 运用模型进行预测分析 (预测)

第一问是一个时间序列分析的问题，我们先对数据进行了预处理，将离群值找出并利用临近均值进行了数据修正，并作出原数据的时序图与一阶差分数据的时序图（并排放两个图）：





由于该数据没有明显的周期特征，一年的数据量不存在周期性。观察一阶差分图预测数据的一阶差分为一个平稳数据，所以我们准备使用传统的ARIMA模型来进行解释与预测。同时，我们也采用回归树（Desicion Tree Regressor）模型进行对比预测，以便验证模型的合适性。

模型原理

ARIMA模型是一种时间序列预测模型，可以用来对未来时间序列的走势进行预测。ARIMA代表自回归（AR）、差分（I）和移动平均（MA），这三个术语对应了该模型的三个主要组成部分。

自回归（AR）指的是使用过去的观测值来预测未来的观测值。该模型假设未来的值是过去的值的线性组合，其中的权重由自回归系数（AR系数）控制。移动平均（MA）指的是使用过去的预测误差来预测未来的观测值。该模型假设未来的值是过去的预测误差的线性组合，其中的权重由移动平均系数（MA系数）控制。差分（I）指的是对时间序列进行差分处理，以使其变得平稳（即均值和方差不随时间变化）。通过对时间序列进行一次或多次差分，ARIMA模型可以去除趋势和季节性因素，从而使时间序列变得平稳。平稳的时间序列可以更容易地建立模型并进行预测。

回归树是一种基于决策树的机器学习算法，用于对连续型数据进行回归分析。回归树的基本思想是将数据集递归地划分成多个区域，每个区域用一个常数来表示该区域内所有数据点的输出值。回归树递归地将数据集划分成多个区域，直到每个区域内只剩下一个数据点或达到预定的停止条件。

模型假设

1. 每日的报告数量仅与时间有关，也就是说，每日单词的难度系数，选择困难模式的人数和尝试数量的百分比以及其他未提及的指标对每日报告数量无关。
2. 数据的一阶差分为一个平稳的时间序列
3. 仅已知2022年1月7日到2022年12月31日的数据来做预测，不考虑美赛导致游戏关注度上升等偶然因素对游戏热度的影响。

主要模型

ARIMA模型

1. 用ADF单位根检验一阶差分时间序列的平稳性

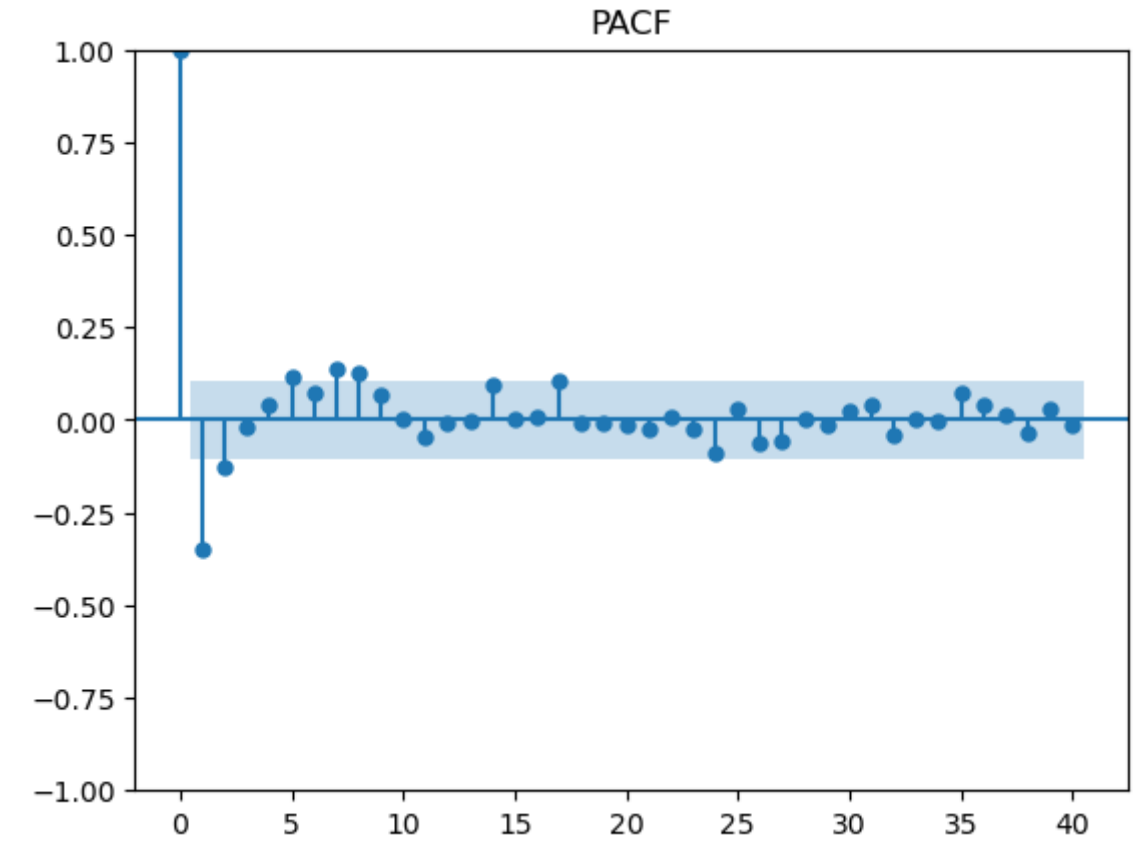
参数	数值
ADF值	-7.1359413592169885
p值	3.4210335686568374e-10
1%置信度	-3.4496162602188187
5%置信度	-2.870028369720798
10%置信度	-2.5712922615505627

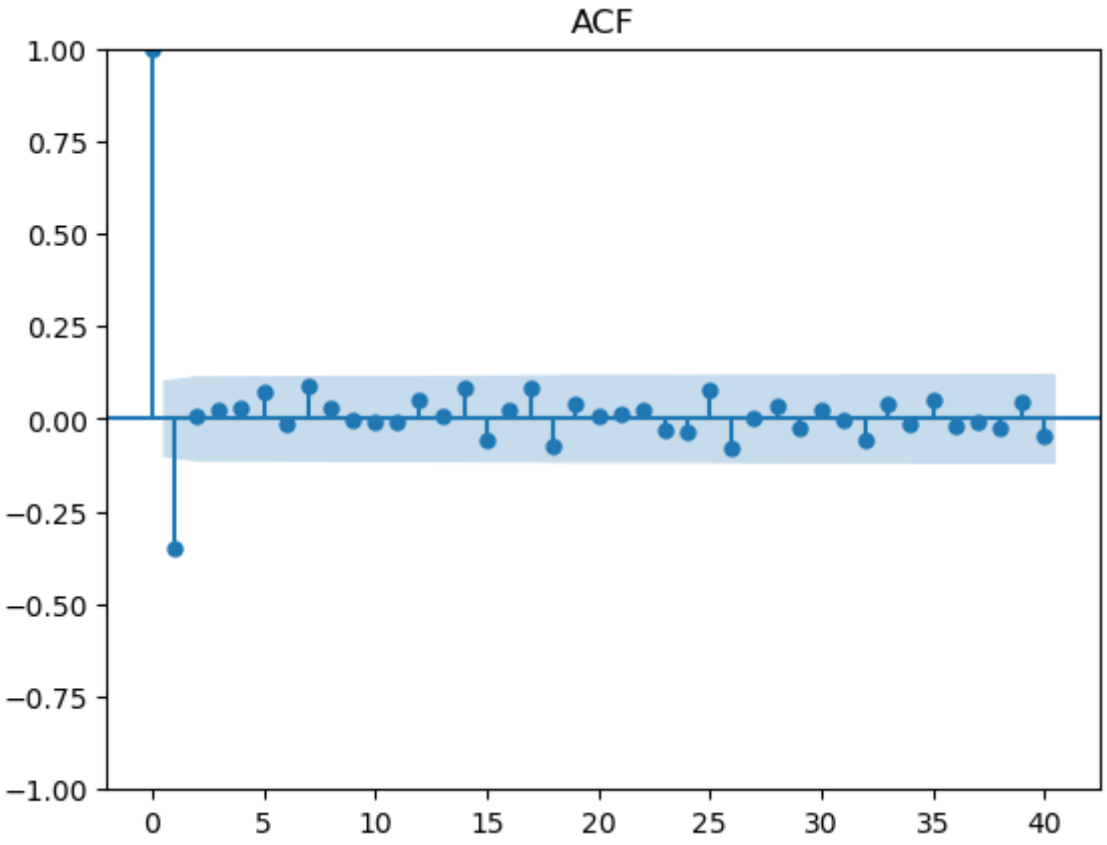
   有大于99%的概率拒绝原假设，因此该序列为一个平稳的时间序列

2. 用Ljung-Box检验验证序列是否为白噪声

   每一个P值都小于0.05或等于0，说明该数据不是白噪声数据，数据有价值，可以继续分析。

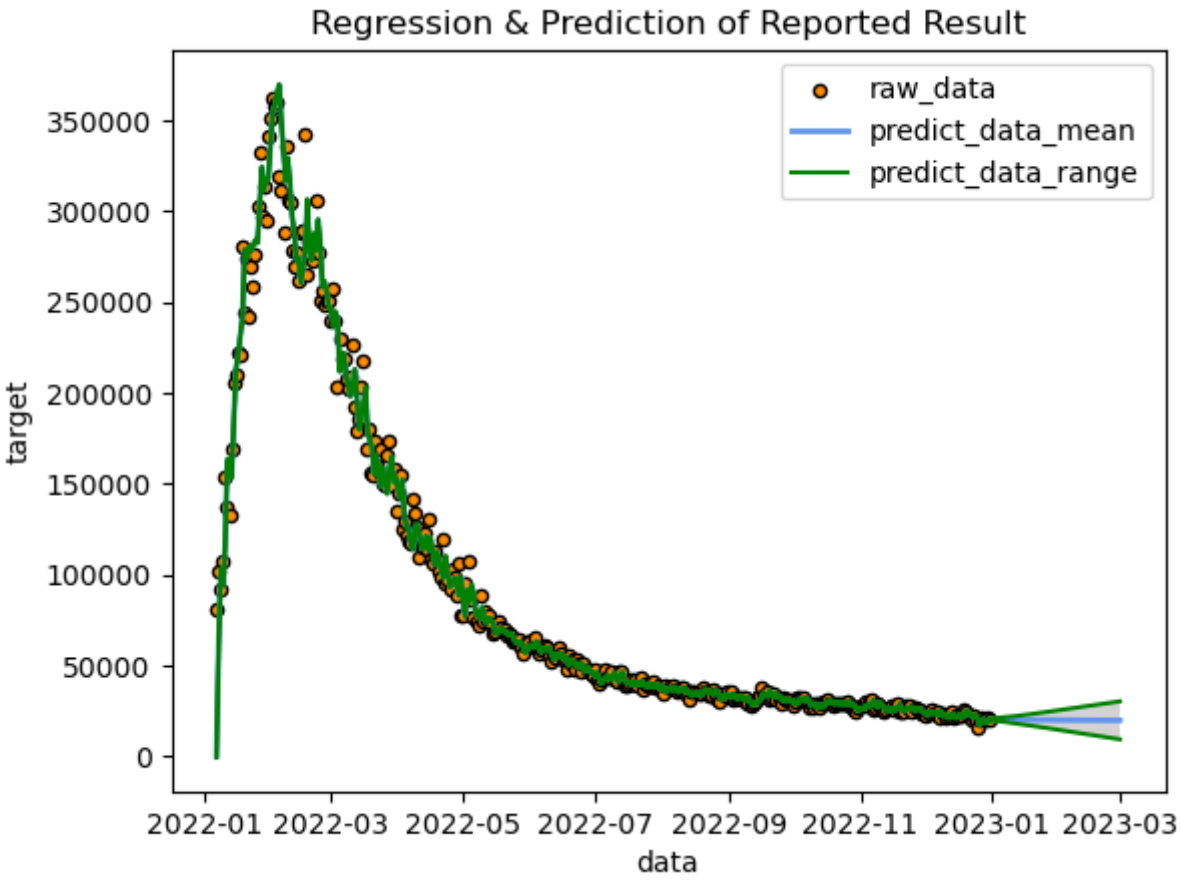
3. 画pacf图和acf图



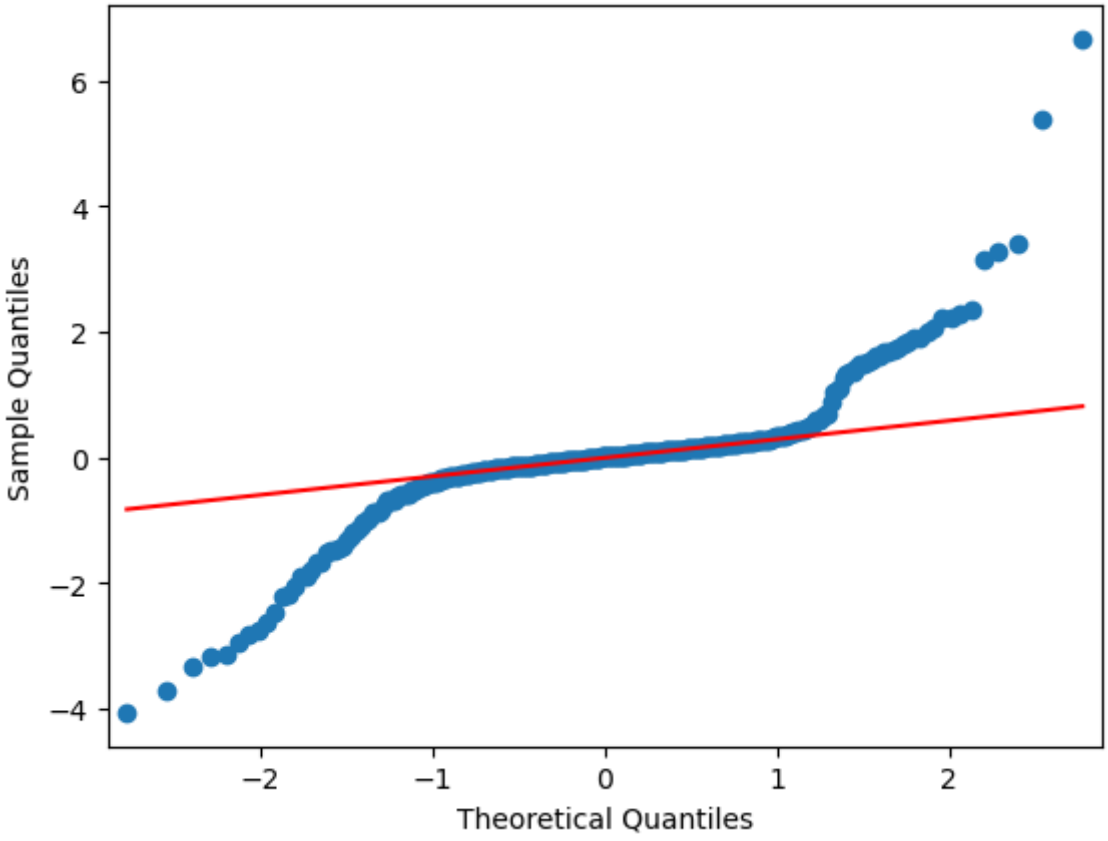


4. 模型调参，通过热力图确定AR()与MA()模型参数

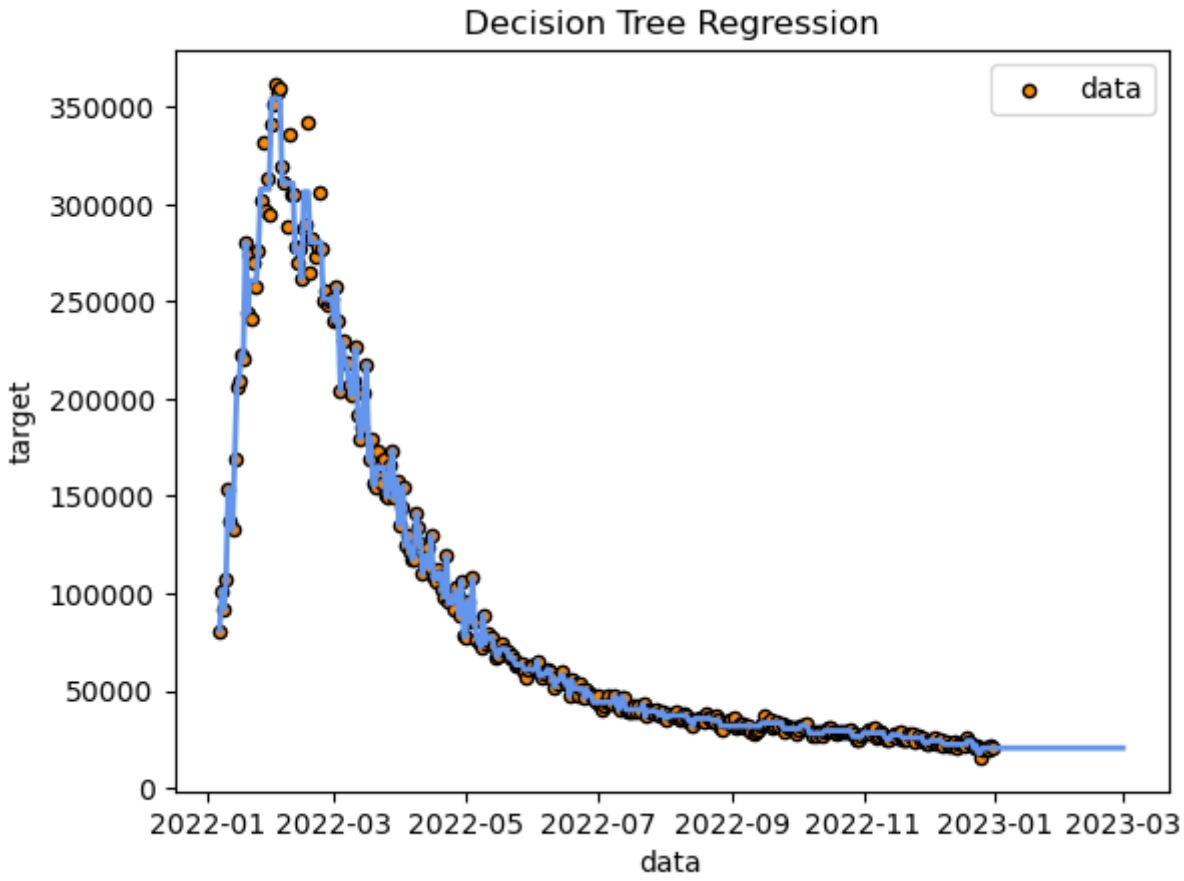
5. 综合情况考虑决定使用ARIMA(4,1,2)模型对时间序列拟合



6. 残差QQ图



7. 随机森林法进行进一步验证



8. 得出结论

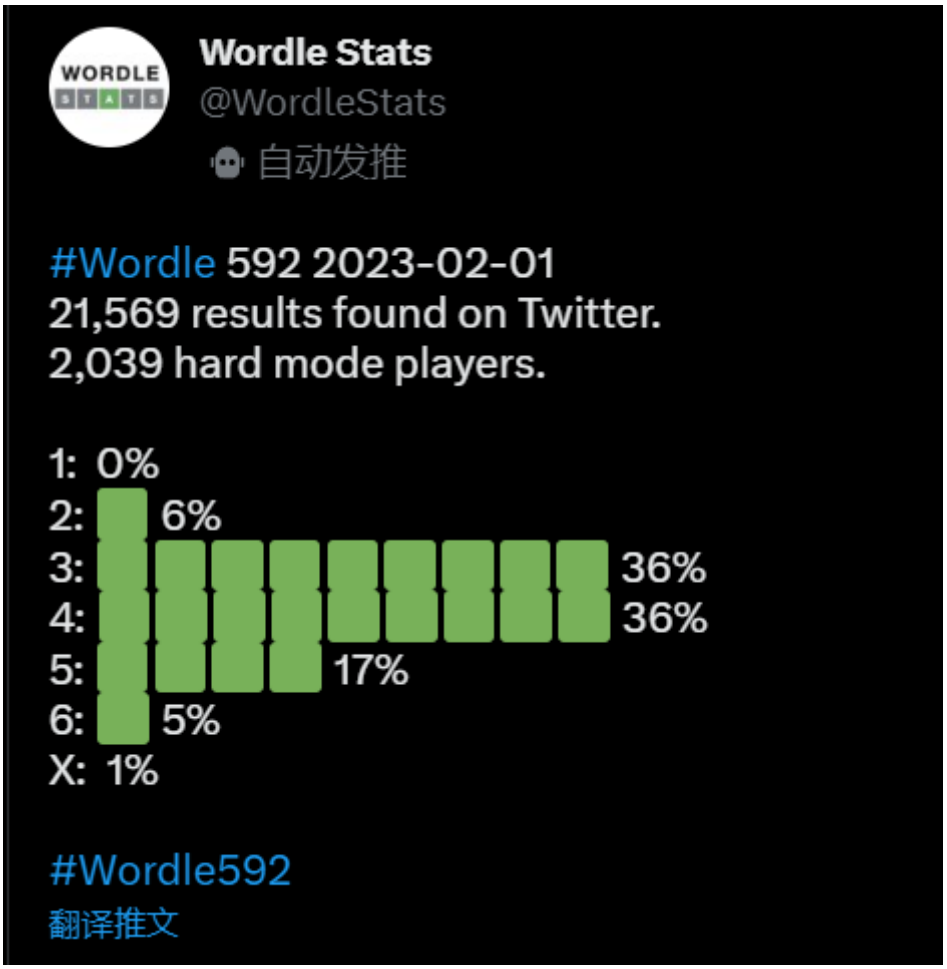
- 描述数据特征

  已完成

- 预测将来时间的报告人数（2.1）

  由ARIMA模型：

  - 预测区间的平均值为：19847.08 = 19847
  - 预测区间上限：25321.29 = 25321
  - 预测区间下限：14372.87 = 14373



基本在预测区间中心，模型拟合效果很好

- 预测将来时间的报告人数（3.1）

  由ARIMA模型：

  - 预测区间的平均值为：19697.53 = 19698
  - 预测区间上限：30148.69 = 30149
  - 预测区间下限：9246.36 = 9246

  由回归树模型（训练准确度0.98259）

  - 预测值：20439.16 = 20439