报告结果数量预测

数据来源：

数据说明

## 数据处理：

首先对数据进行异常值处理

|  |  |  |
| --- | --- | --- |
| Date | word | Number of reported results |
| 2022/12/16 | rprobe | 22853 |
| 2022/12/11 | naïve | 21947 |
| 2022/11/30 | study | 2569 |
| 2022/11/26 | clen | 26381 |
| 2022/10/05 | marxh | 30935 |

|  |  |  |
| --- | --- | --- |
| Date | word | Number of reported results |
| 2022/12/16 | probe | 22853 |
| 2022/12/11 | naive | 21947 |
| 2022/11/30 | study | 24419 |
| 2022/11/26 | clean | 26381 |
| 2022/10/05 | marsh | 30935 |

**justification：**

·在2022/11/30，Number of reported results仅为2569，而Number in hard mode为2405，我们用当周Number of reported results值的均值替换当日的该值，改为为24419

·由于该游戏的答案都是有词意的，且都是5个字母组成，我们修改了没有词义和拼写错误的单词，并将他们改为较为常见的单词加入数据集中。

## 数据分析：

本文首先探究星期一至星期日对wordle报告结果数量是否有差异性，结果数量是否呈均匀分布。本文利用卡方检验中的拟合性检验判断是否为均匀分布。

服从自由度为6的卡方检验（fo为实际数，fe为理论数）

Xij表示第i周星期j的报告数量,。（i=1，2……50,j=1,2……7）

Foj=求和（i=1-50）xij

Fe=sum\_foj/7



=23172.9062>卡方0.05（6）=12.5916，故拒绝原假设，星期一至星期日对Number of reported results具有差异性。工作日周一至周五Number of reported results显著大于周末，同时趋势从周日开始递增到周三随后递减至周日。

## 预测模型建立

首先我们求出一周内每天的reports数量的差异变化,构建一个 weekly wordle hot 分布，H\_j（j=1，2……7）表示一周中wordle游戏报告结果的概率分布Hj= 白板上写着字

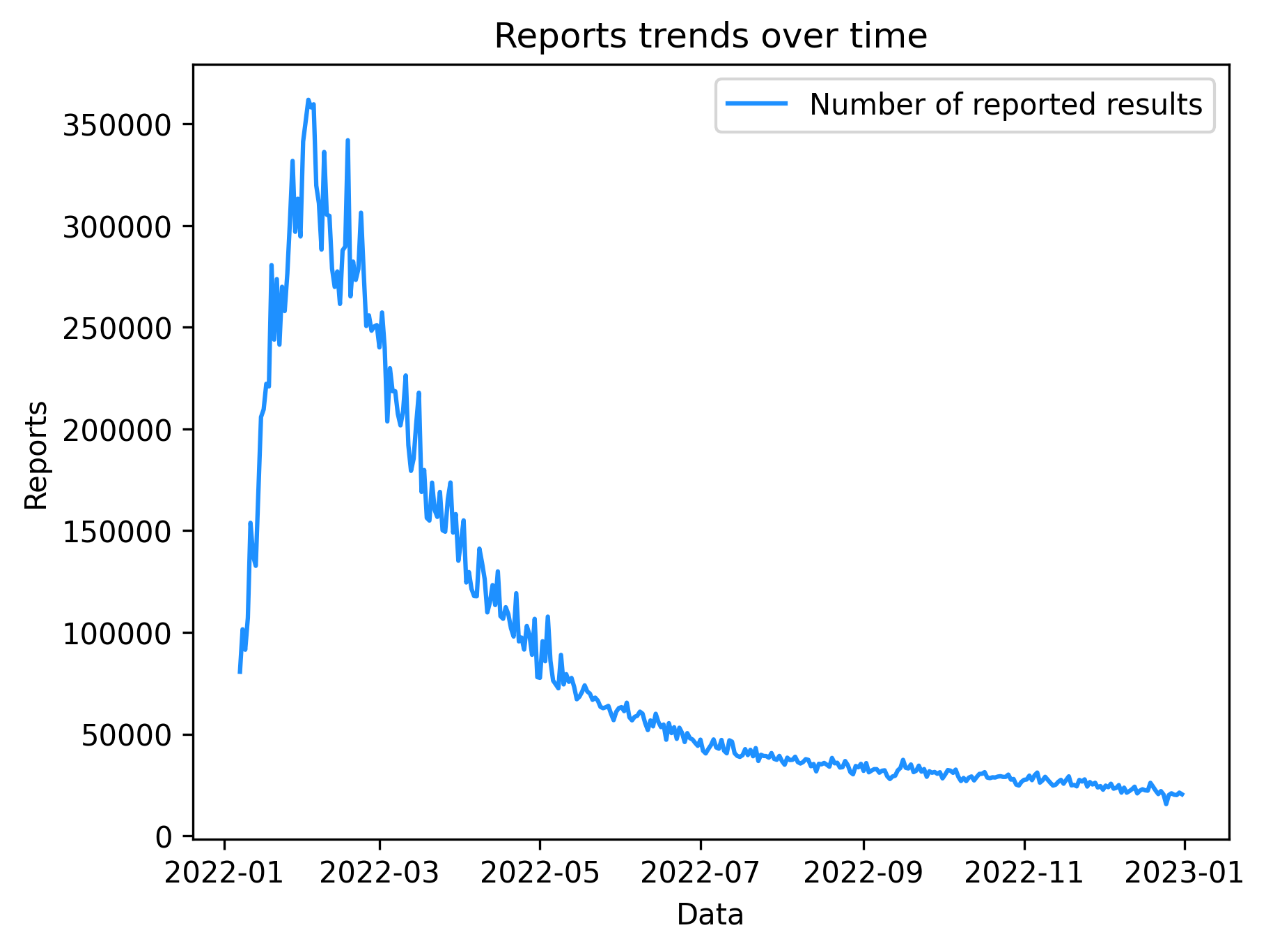
描述已自动生成



以上为我们求出的wordle hot 分布:w\_hot

然后再预测周间差异，利用50周的周平均值预测第59周的报告结果平均数，再根据周内差异，周一至周日分布差异不同，预测2023年3月1日星期三的报告结果数量。

y\_day=y\_week\*w\_{hot}^T

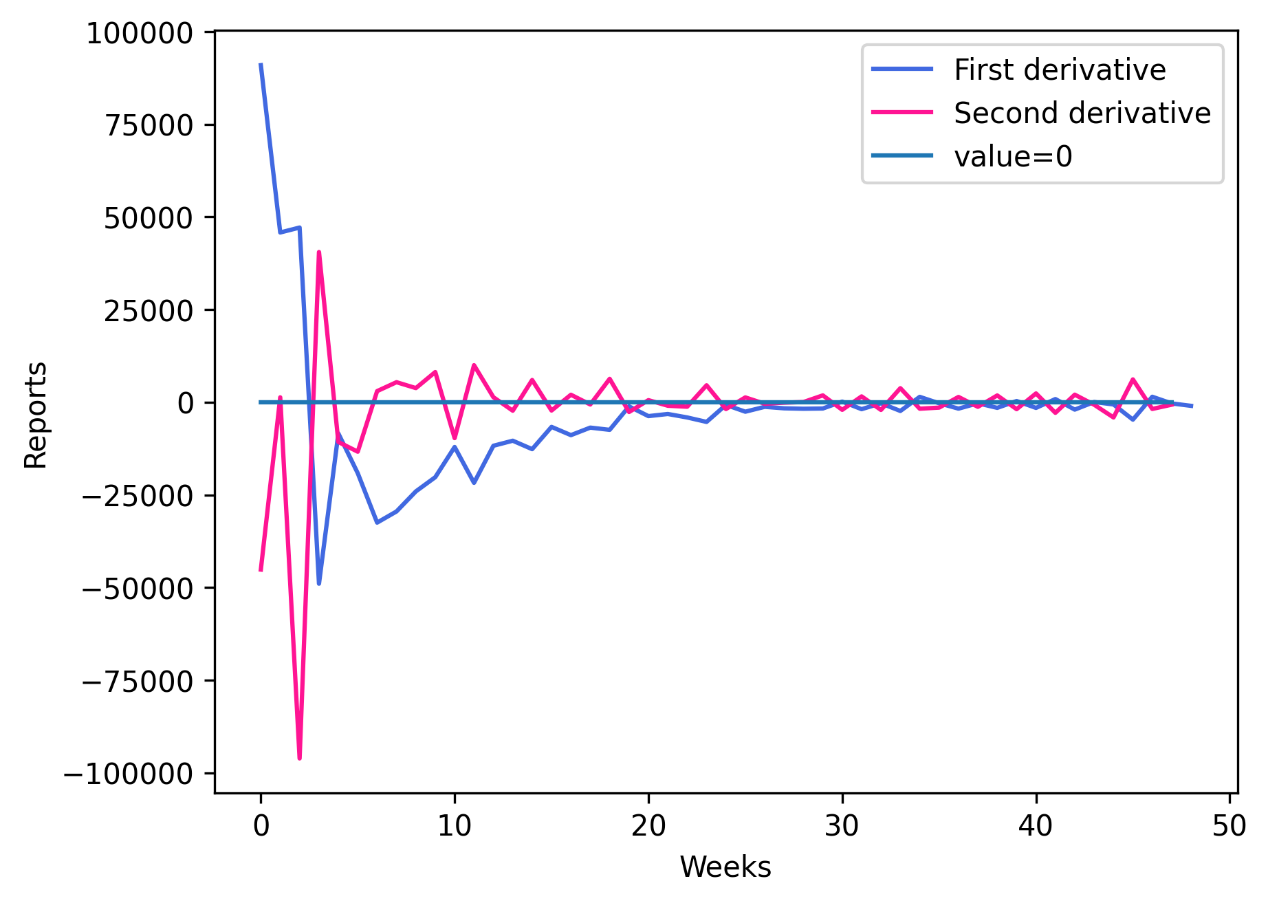


根据折线图不难发现，wordle 的Number of reported results最初快速增加达到顶点，随后逐渐缓慢下降，游戏热度衰退。

我们设定2022/1/10-2022/1/16为第一周，故2023/3/1为第59周，从图中可以判断，该游戏经过前期快速爆红，报告结果激增，达到最大值后保持下降趋势，下降速度先快后慢，第20周后，数据点更加密集，函数趋势稳定。

我们对上图的报告数量随时间的趋势图像做两次求导，可以发现，在20周后，二次求导的函数值基本浮动在0附近，可以认为函数下降趋势趋于稳定

为了减少噪声干扰，我们取**20到50周**的数据预测第59周值。



## 模型求解：

指数拟合：

我们利用Matlab中的curve fitting进行指数拟合

在置信区间95%的情况下，假定函数方程为a\*x^b+c，求得

a = (1.43e+06, 2.183e+06) b = (-1.189, -1.065)，我们取平均值最后确定a=1.807e+06 b=-1.127 c=0

模型评估指标：

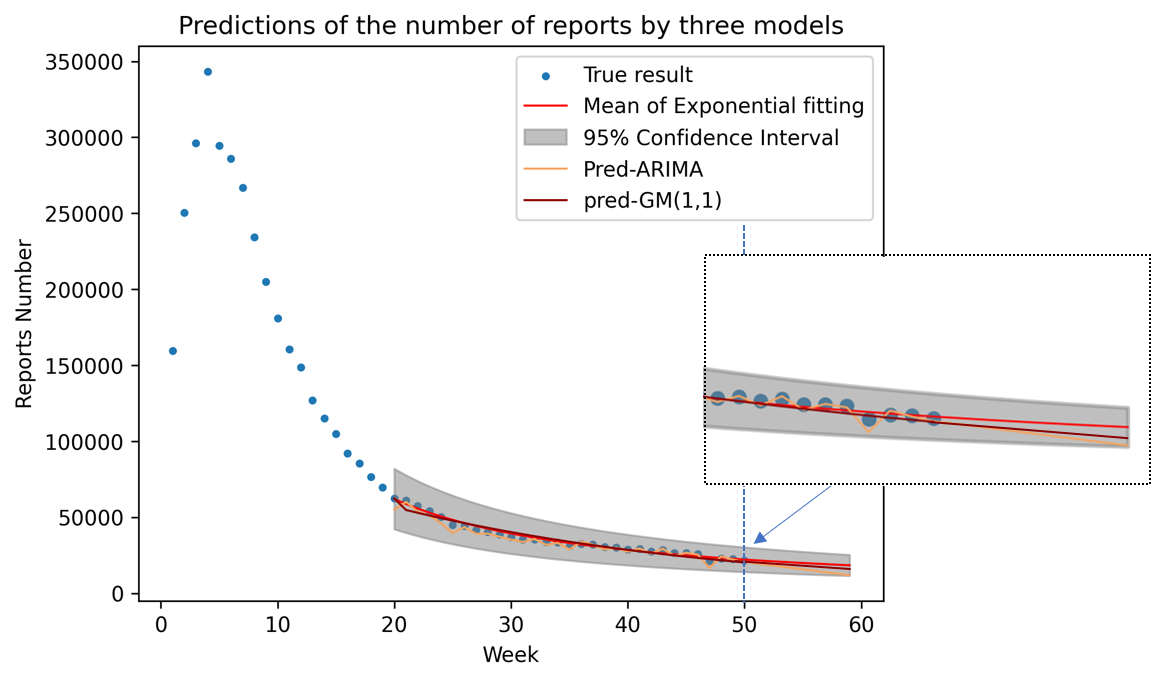
|  |  |  |
| --- | --- | --- |
| SSE | R-square | RMSE |
| 8.147e+07 | 0.9794 | 1675 |

模型拟合效果图

图表, 散点图

描述已自动生成

同时我们采取灰色预测模型和arima（0，2，0）时间序列预测模型对第59周进行预测



从图中我们可以很清楚的看到，本题指数拟合效果较好，数据点在拟合曲线附近。最终预测第59周的预测区间为:

【11485，25311】

根据wordle hot分布和第59周周平均值18248，

Xij=周平均值\*7\*Hj

X59,3=18490



最终我们根据三个模型的预测结果，确定使用指数拟合来预测第59周数据

1. 指数拟合的拟合效果和模型准确性上明显好于另外两者。
2. 从第一周到第50周绘制的散点图以及一阶导数和二阶导数中可以很清晰的看出每周报告结果数量呈下降趋势且下降的趋势变慢趋于稳定。而GM(1,1)和ARIMA（0，2，0）模型预测，未能体现下降的趋势变慢。
3. 结合相关研究和论文，发现一款游戏经过爆火到逐渐热度下降，最后游戏玩家会逐步下降且下降速度变慢最终趋于一个最小值。GM(1,1)和ARIMA模型预测下，最终会为负值。