猜词的洞见：Wordle背后的数据探索与分析

Insights from guessing words: the data exploration and analysis behind Wordle

基于ARIMA模型和MIMO G

基于情绪分析模型和ARIMA模型的市场数据跟踪测量

summary

Wordle 是风靡全球的猜词游戏。在这个游戏中，你可以在六次机会中尝试成功猜出每天唯一的单词，并将自己的骄傲成绩分享到社交平台。在猜词游戏中，最让人好奇的无疑是今天的单词究竟是什么，这种好奇心以及这个游戏所带来的社交属性让玩家们为它疯狂。但是游玩过程中，显然单词的难易程度会极大影响当天玩家的尝试次数，并可能对后续游戏的流量造成影响，所以我们希望在对于每日单词的分析中去发现一些有趣的规律，并向纽约时报的Wordle编辑提出我们的建议。

数据预处理

基于对 Wordle 的规则和社交网络特点的理解，我们从空间和时间两个维度分析了需要的数据特征。我们参考了权威的英语语素构词学以及全面的 Wordle 攻略，提出具有单词拓扑-玩家水平和融合社交网络活跃度的异构时空特征，并遵循数据标准化原则对特征量化处理，得到定量和定类（以 IF 开头）属性。特征工程是本研究的重中之重，贯穿后续的相关性分析、预测和分类任务。随后，我们对现有数据进行统计性描述，包括数据的集中趋势与离散趋势，发掘数据集的特性。并且清除了异常数据，完成了数据清洗工作。

时序分析-时序预测

在数据预处理后，我们通过构建ARIMA模型对每天游玩Wordle的人数进行了时序分析，模型的拟合优度表现非常优秀。我们除了对现有数据进行拟合外，还进一步分析得出了未来数据的预测区间。然后针对词语属性是否对困难模式选择率有影响这一问题，我们选择使用Spearman相关性分析来解决，分析结果证明两者之间没有显著相关性，这进一步指明了我们提取的单词特征与玩家困难模式选择率的正交性，为我们后续的属性选择提供了数据支撑。

MIMO XGBoost: 同步训练-异步使用的联合模型

为了预测 1 try-7 or more tries (X) 的分布情况，我们提出了一种 MIMO XGBoost 模型，可在预测时考虑他们间的关联性并引入频率和为 1 的约束。具体而言，我们时间同步地训练了 7 个 XGBoost 回归模型，根据单词相关的多个时空属性预测 1 try-7 or more tries (X) 的分布情况，即 MIMO。并且设计了分布式的损失函数，确保 7 个模型朝一个共同目标优化参数，即满足频率和为 1 的约束。联合训练完成后，每个模型均可独立使用，即异步使用。实验结果表明，我们提出的 MIMO XGBoost 相较于独立训练 7 次 XGBoost 模型，使 1 try-7 or more tries (X) 的频率更接近 1，同时在测试集上得到了更好的精度和拟合优度。我们预测了单词 EERIE 在 2023.3.1 的玩家尝试次数分布。我们也给出了各属性在难度分类中的重要性比例，其中The Number of duplicate words N\_dw， Vowel beginning or not IF\_ve, Workday or not IF\_w and The number of infrequently used alphabets N\_iw 的重要性分别占到16.7%, 12.0%, 11.6% and 10.9%。值得注意的是，我们的 MIMO XGBoost 为后续分类任务提供了基础。

基于 Empirical Rule 的分类模型

我们定义了平均尝试次数 ATN 作为对于玩家群体而言特定单词难度的表征，并且对平均值尝试次数 ATN 进行了 Shapiro-Wilk 正态性检验，结果表明 ATN 可接受为正态分布。随后，我们根据 Empirical Rule 给 ATN 划分出三个区间，分别对应 easy, normal, hard 三种单词难度。自此，单词难度分类任务转化为对于 ATN 的预测任务，这使得我们可以沿用上述 MIMO XGBoost 模型预测特地单词在特定日期的 ATN，同时保持与上述预测任务一致的高分类精度。我们的分类模型表明，EERIE 的难度为 hard。

洞见数据

## 1. Introduction

### 1.1 Background

“The purpose of computation is insight, not numbers.”

-Richard Hamming

Wordle是目前世界上最火的猜词游戏之一，在这个游戏中，你可以在六次机会中尝试成功猜出每天唯一的单词，并将自己的骄傲成绩分享到社交平台。在猜词游戏中，最让人好奇的无疑是今天的单词究竟是什么，这种好奇心以及这个游戏所带来的社交属性让玩家们为它疯狂。但是游玩过程中，显然单词的难易程度会极大影响当天玩家的尝试次数，并可能对后续游戏的流量造成影响，所以我们希望在对于每日单词的分析中去发现一些有趣的规律，并向纽约时报的Wordle编辑提出我们的建议。

### 1.2 Problems Restatement

问题陈述+流程图

* 纽约时报要求我们分析被提供数据，得到有价值的信息，具体包括：
  + 解释 reported results 随时间变化的原因并给出特定日期的预测。
  + 判断单词是否会 affect the percentage of scores reported that were played in Hard Mode。
  + 预测特定单词在特定日期的 distribution of the reported results
  + 创建分类模型，为单词的难度分类。
  + Identify the attributes of a given word that are associated with each classification.
  + 评价我们的模型。
  + 发掘数据集的特性。
* 为了实现上述目标，我们提出了一系列子任务，且它们间存在先后制约，具体而言：
  + 对有关数据进行统计性描述，包括数据的集中趋势与离散趋势，发掘数据集的特性。
  + 分析 reported results 的时间序列图，结合社交网络传播特性，确定适合的预测模型。
  + 针对 Wordle 游戏的规则，提取单词的特征，并确定各特征的度量。
  + 分析单词的相关特征与困难模式选择率之间的相关性，以确定单词对玩家选择的 mode 的影响。
  + 针对社交网络的传播特性，基于日期提取特征并确定其度量。
  + 基于日期和单词的相关特征，预测报告结果的分布情况。
  + 针对 Wordle 游戏的规则，确定单词难度的度量，并为单词的难度分类。

### 1.3 （Literature Review）

## 2. Assumptions & Nomenclature

### 2.1 Assumptions

我们在模型中做了几个假设。之后，我们可能会放松这些假设，以优化我们的模型，使其更适用于复杂的现实环境。

* 人们在推特上面上传的记录是准确的。
* 所提供的数据不包括来自竞争对手或恶意客户的不正确数据。
* 每天玩家的数量记录等于统计的数量总和，即，在任何提供的数据集中都没有遗漏或排除的记录。
* 玩家上传的记录是游玩当天发布的。
* 每个玩家最多一天上传一次记录。

### 2.2 Nomenclature

：输出值

：预测输出值

V\_d:差值

W\_i:weight

PS：part of speech词性

：word frequency词频

ATN：Average number of attempts平均尝试次数

：Difficulty Level难度等级

：Percentage in hard mode

：Percentage of the i tries

：The Number of duplicate words重复字母个数

：The number of vowels元音个数

：The number of consonants辅音个数

：The number of infrequently used alphabets不常用单词字母个数

：The number of syllables音节数

：Repetition is continuous or not重复是否连续

：Vowel beginning or not是否元音开头

：Vowel ending or not是否元音结尾

：Vowel is continuous or not元音是否连续

：Polysemous or not是否是多义词

：Workday or not是否是工作日

## 3. Data Preprocessing

### 3.1 Feature Extraction & Quantification

已知的数据中只包含日期，单词，reported results 数量和玩家尝试次数分布等属性，为了完成玩家尝试次数的分布的预测，我们需要更多属性以构建我们的模型。

针对 Wordle 的规则和社交网络的特点，我们从空间和时间两个维度提取数据的特征：

空间维度上，Wordle 是一个面向字母的英语单词解谜游戏，这决定了单词的拓扑极大地影响着人类玩家尝试次数的分布。通过分析社交网站上的 Wordle 攻略和英语单词的语素构词学，我们提出使用如下特征度量单词的拓扑： 文本

描述已自动生成

图为mummy单词的量化结果（图）。

另一方面，困难模式也是尝试次数的重要影响因素，我们引入困难模式选择率：

H\_r = num of rr in hard mode / num of rr

时间维度上，社交网络的活跃度可能会影响游戏的参加人数以及玩家水平结构，进而可能会影响尝试次数的分布。我们选取纽约证券交易所的交易日数据对日期标注，引入 Workday or not 属性作为社交网络活跃度的表征。

综上，我们提出具有单词拓扑-玩家水平和融合社交网络活跃度的异构时空特征，并遵循数据标准化原则对特征量化处理，得到定量和定类（以 IF 开头）属性。

### 3.2 Data standardization

标准化是必要的一步。这里我们采用Min-Max Normalization方法，根据原始数据的最大值Max和最小值Min对数据进行处理。处理后的数据符合区间从0到1的分布。转换函数为

在标准化步骤之后，得到了统一区间的标准数据(mummy单词的结果见表3)。

表2：未标准化前的数据

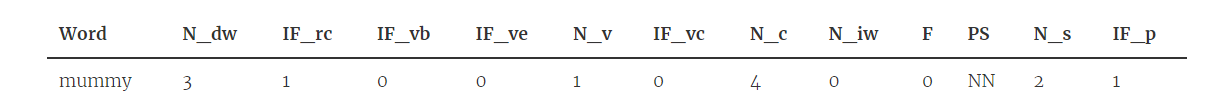
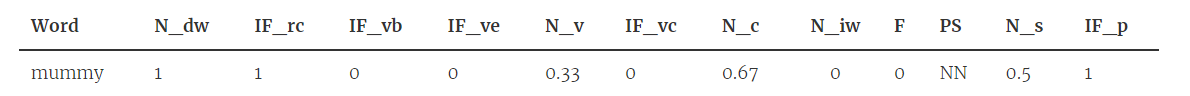


表3：标准化后的数据



### 3.3 Data Cleaning

通过对提供的数据进行观察，我们发现在数据集中有一些异常值。在Word属性中有两个单词只有4个字母，有一个单词有6个字母，明显不符合猜测字母数为5个的单词的游戏规则。另外在数值上也有表现异常的单词，例如，在1try-7tyies属性中，尝试次数的总百分比远大于100。类似的，在Number of reported results属性中，两个单词的统计数量远低于平均的数量级。考虑到这些单词对于我们的建模可能有负面影响，我们忽视了这些异常单词。

最后基于上述过程，我们汇总出了一个文件（word\_normalization）包含了所有出现过的单词，并将单词的属性标准化的。

## 4. Q1的模型名字

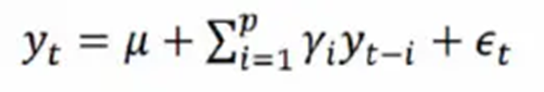
在这一节中，我们利用Autoregressive Integrated Moving Average(ARIMA)预测模型作为对于Wordle未来在线人数的数据模拟度量，基于历史时间上Wordle的人数表现情况，同时通过滞后移动平均线来获取平滑的时间序列，从而对未来某个时间段的在线人数做出一个大致估计。然后我们对单词属性进行了特征提取，利用相似性矩阵来分析单词的任何属性是否影响到参与困难模式的人数占总人数的百分比。

## Model Review

### 1.Auto Regression(AR)模型

自回归模型描述的是当前值与历史值之间的关系，用变量自身的历史时间数据对自身进行预测。

其中，一般情况下，p阶自回归过程的公式定义：



一般的P阶自回归模型 AR：

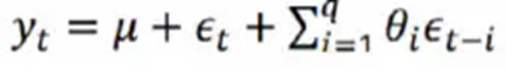


从公式中可以看出，当前值是通过历史值来预测的，p是自回归模型中的一个阶数，表示用几期的历史值来预测当前值。

### 2. Moving Average(MA)模型

移动平均模型关注的是自回归模型中的误差项的累加。

其中，q阶自MA模型的公式定义：



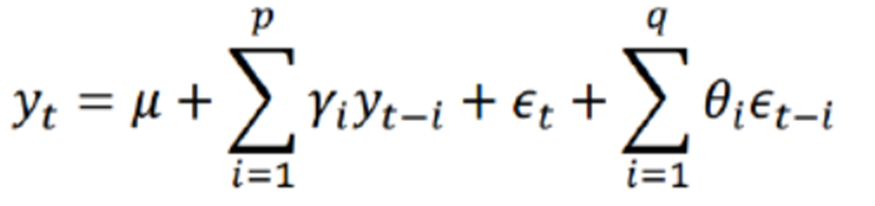
在AR模型中，如果ut不是一个白噪声，通常认为它是一个q阶的移动平均：



移动平均法能有效地消除预测中的随机波动

## 3. 自回归移动平均(ARMA)模型

自回归移动平均模型由两部分组成：自回归部分和移动平均部分，回归方程表示为：



从回归方程可知，自回归移动平均模型综合了AR和MA两个模型的优势，在ARMA模型中，自回归过程负责量化当前数据与前期数据之间的关系，移动平均过程负责解决随机变动项的求解问题

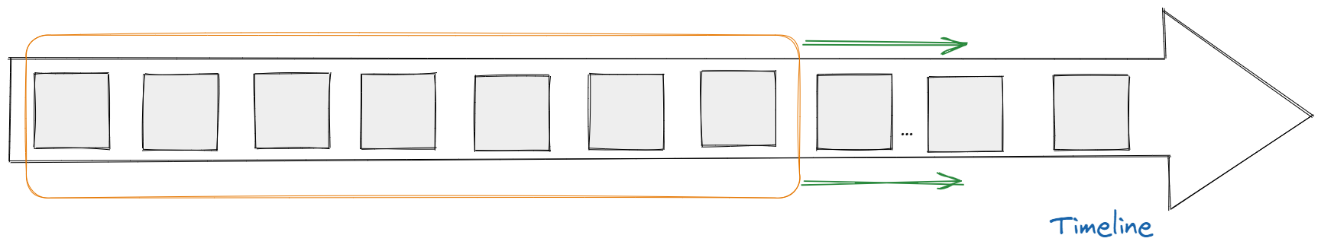
## 4.差分自回归移动平均(ARIMA)模型-ARIMA(p，d，q)

自回归模型（AR）、移动平均模型（MA）和差分法（I）结合，我们就得到了差分自回归移动平均模型 ARIMA（p、d、q），其中 d 是需要对数据进行差分的阶数，ARIMA是经过差分后的ARMA模型

## Model Analysis

步长选择：

考虑到ARIMA模型对于预测太长的结果具有较高的不确定性，我们加入滑动窗口来探究未来玩家的数量，步长设为7，使得预测的时间单位从天变为了周，可以大幅度减少预测次数。



1.ARIMA模型要求序列满足平稳性，查看ADF检验结果，根据分析t值，分析其是否可以显著性地拒绝序列不平稳的假设(P<0.05)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ADF检验表 | | | | | | | |
| 变量 | 差分阶数 | t | P | AIC | 临界值 | | |
| 1% | 5% | 10% |
| Average Number | 0 | -3.276 | 0.016\*\* | 651.242 | -3.616 | -2.941 | -2.609 |
| 1 | -4.002 | 0.001\*\*\* | 627.622 | -3.616 | -2.941 | -2.609 |
| 2 | -3.867 | 0.002\*\*\* | 620.775 | -3.621 | -2.944 | -2.61 |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | | | | | | | |

通过查表以及对差分序列图像的对比，我们得出最好的差分值为1阶，在差分为1阶时，显著性P值为0.001\*\*\*，水平上呈现显著性，拒绝原假设，该序列为平稳的时间序列。



最佳差分图

2.查看差分前后数据对比图，判断是否平稳，同时对时间序列进行偏（自相关分析），根据截尾情况估算其p、q值，如下图所示

（两图合并）



**最终差分数据自相关图(ACF)**

图表

描述已自动生成

**最终差分数据偏自相关图(PACF)**

我们最终确定ARIMA模型为ARIMA(2,1,0)，模型参数表如下，表格中展示了本次模型检验结果，包括样本数、自由度、Q统计量和信息准则模型的拟合优度，根据检验表，模型的拟合优度R²为0.977，模型表现优秀。

|  |  |  |
| --- | --- | --- |
| ARIMA模型（1,1,0）检验表 | | |
| 项 | 符号 | 值 |
|  | Df Residuals | 45 |
| 样本数量 | N | 48 |
| Q统计量 | Q6(P值) | 3.317(0.069\*) |
| Q12(P值) | 13.838(0.031\*\*) |
| Q18(P值) | 18.141(0.111) |
| Q24(P值) | 18.315(0.435) |
| Q30(P值) | 18.384(0.784) |
| 信息准则 | AIC | 1041.317 |
| BIC | 1046.867 |
| 拟合优度 | R² | 0.977 |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | | |

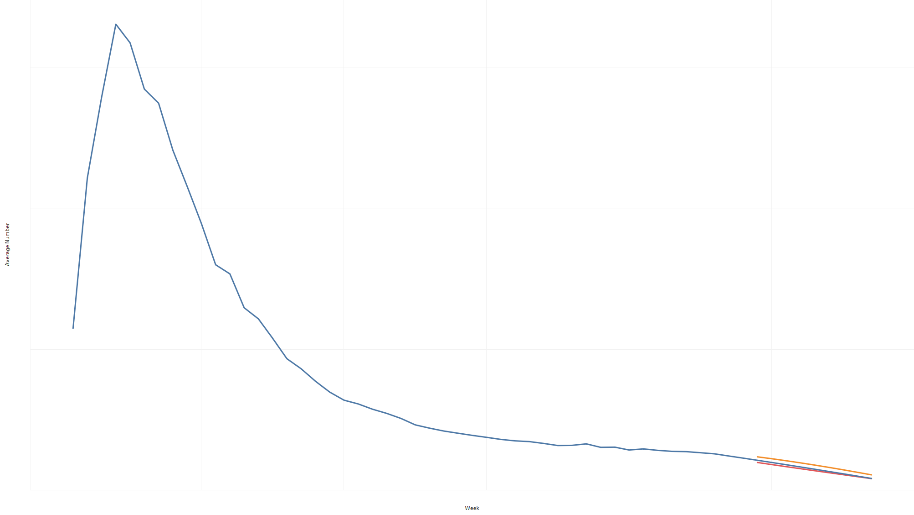
最终我们得出了对于模拟平均人数的时间序列图和预测区间图，如下图所示：

（两个图和一个）

图表, 折线图

描述已自动生成

*（放大左下角的图）*



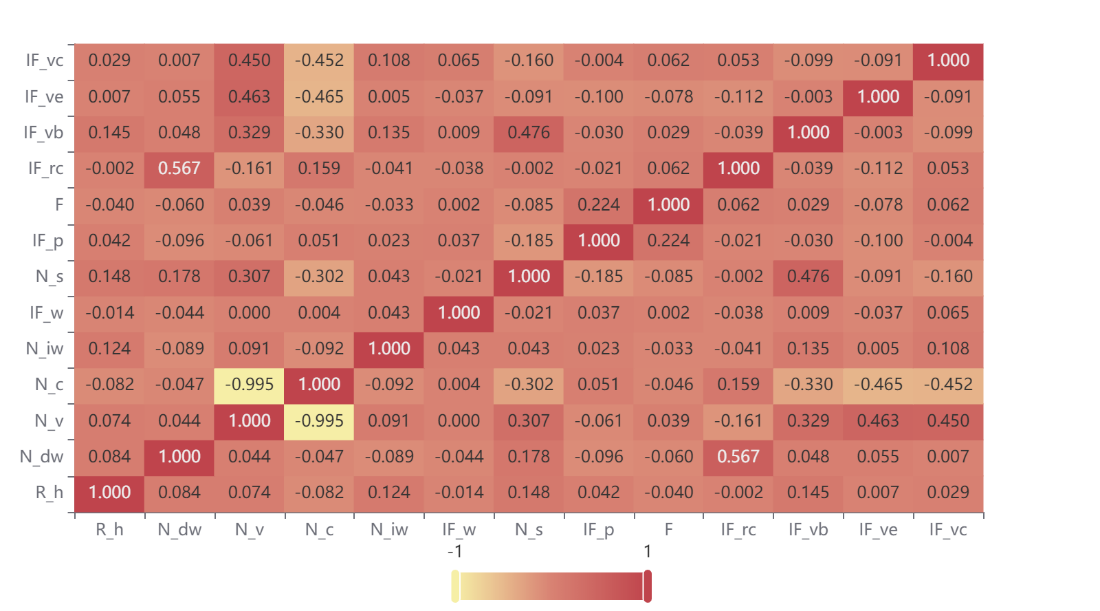
预测区间的得出是通过预测每周的最小玩家数量和最大玩家数量，将其数量分别作为区间的左右临界值得出一个更加可靠的数值。

## 相似性矩阵

猜想：因为在Wordle游戏中，玩家直到游戏结束前，都无法知道单词具体是哪一个，所以这不应该影响到玩家对于困难模式的选择，基于这一猜想我们进行了如下分析。

在分析词语属性和参与困难人数百分比的相关性上，我们选择使用Spearman相关性分析。因为Pearson相关性分析主要用于分析满足正态分布的两定量变量之间的关系，若两变量中包含等级变量，或变量不符合正态分布，或变量分布类型未知时，采用Spearman相关性分析是更合适的分析方式，Spearman相关性分析的基本思想是：分别对两个变量*X*、*Y*做秩变换(rank transformation)，用秩次*RX*和*RY*表示，取其平均秩次为每个数据的秩次，其中Spearman相关系数*rs*的计算公式为：

分析如下图：



从图中可以很清晰的看出，参与困难人数百分比和词语的各属性之间具有极弱相关不相关的关系，所以可以认为词语的属性并不能影响游戏当天选择参与困难模式的人数。

## 5. 优化XGBoost Model

该模型不仅要使用3.1节中提取出的特征值，而且时间维度也是一个重要的影响因素。我们在根据社交网络的活跃度在时间上的变化，总结出了 属性。综上所述，我们优化了XGBoost回归预测模型，并且基于3.1节的结论，以单词的各个属性与时间属性作为输入值，对（1，2，3，4，5，6，X)的分布情况进行预测，并且对模型的准确度进行了量化测试。

### 5.1 XGBoost Model Analysis

### 5.1.1 Decision Tree Ensembles

树集成模型由一组分类和回归树 (CART) 组成,从数学上讲，我们可以将模型写成以下形式：

其中，K是树的数量，f是F的函数空间，F是可能的CART集合。上述模型的目标函数由下式给出：

其中，第一项是损失函数，第二项是正则化参数。

### 5.1.2 Tree Boosting

对于当前树模型，采用的学习方式是定义目标函数并对其进行优化，上面我们已经定义了目标函数式，考虑到学习树结构的难度，所以我们另外采取了加法策略对其优化，每次添加一棵新树：

然后选取可以优化目标函数的树，并考虑使用均方误差 (MSE) 作为损失函数：

将损失函数进行泰勒展开至二阶并去掉所有的常量之后可得：

在这个定义下，目标函数的值只取决于和. 这就是 XGBoost 支持自定义损失函数的方式

### 5.1.3 The Structure Score

重新制定树模型后，我们可以将目标值写为：

若树的结构部分 q 已知，可使用目标函数寻找最优 Wj，并得到最优目标函数值。其本质可归为二次函数的最小值求解问题。解得：

## 5.2模型优化

### 5.2.1 XGBoost回归预测模型的不足

我们这里使用的XGBoost模型是一个多输入单输出模型，只能一次性预测一个fig:。如果只使用该模型，不同的fig:就**没有了关联性**，可能会导致**预测的总百分**比加和远**大于或小于100**,下标为单词black的预测：

XGBoost在单词**black**上的应用

|  | 1 try | 2 tries | 3 tries | 4 tries | 5 tries | 6 tries | 7 or more tries (X) | total |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **percentage** | 0 | 8.9793 | 27.9642 | 32.1828 | 15.0472 | 5.0703 | 1.0029 | 90.2467 |

使用简单的XGBoost回归预测模型对单词进行预测，发现预测值的总和距100偏差较大，换句话说，之间的关联性缺失。

### 5.2.2 优化XGBoost模型

在实际中fig:之间的关系如下：

因为四舍五入的关系，允许最后数据的误差值的大小为1。

进而可以计算出预测值和100的difference value（），计算公式如下：

对于这个模型来说，每个预测值都会对产生影响，他们分别的权重为:

综合预测值权重和difference value，我们给出了每个预测值对于偏差值的影响值()，相关公式如下：

将公式几（上一个公式）作为代价函数的补充加入XGBoost回归模型的目标函数（公式几？），就可以得到优化后的XGBoost回归模型，优化后的模型可以符合各个之间的关联性，公式如下：

将损失函数进行泰勒展开至二阶并去掉所有的常量之后任然可得：

目标函数的值只取决于和. 符合XGBoost 支持自定义损失函数的方式

## 3.模型评估

### 3.1模型测试

使用word\_normalization文件中的70%数据作为训练集，剩下的部分作为测试集，对测试集的Percentage of the i tries（fig:) 进行计算，得到的各个tries的预测值，**在表i中左边为预测值，右边为真实值**

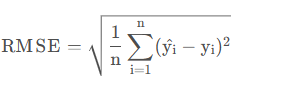
3 tries 测试样本与真实样本的部分表格

|  | predicted | true value |
| --- | --- | --- |
| **begin** | 26.021 | 26 |
| **being** | 21.982 | 22 |
| **berth** | 23.769 | 24 |
| **black** | 30.904 | 31 |
| **booze** | 6.988 | 7 |

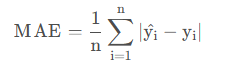
从上表可以看出，所有的预测结果和真实结果相差很小，最多只有**0.24**的差距。

### 3.2 量化评估

进行我们利用几个常见统计参数对于得到的数据结论进行量化检验，Root Mean Square Error（）是我们选择使用的第一个衡量标准，它可以有效检验预测值的准确度，相关公式如下：



另外我们选取了Mean Absolute Error（MAE）进行准确度衡量，MAE可以反映出预测值误差的实际情况，相关公式如下：



最后选取常规的拟合优度fig:作为模型拟合的整体评估，下表为测试集与训练集的检验结果统计（)：

EERIE的评估结果

|  | RMSE | MAE | R^2 |
| --- | --- | --- | --- |
| **training set** | 2.907 | 0.807 | 0.95 |
| **test set** | 2.208 | 1.094 | 0.94 |

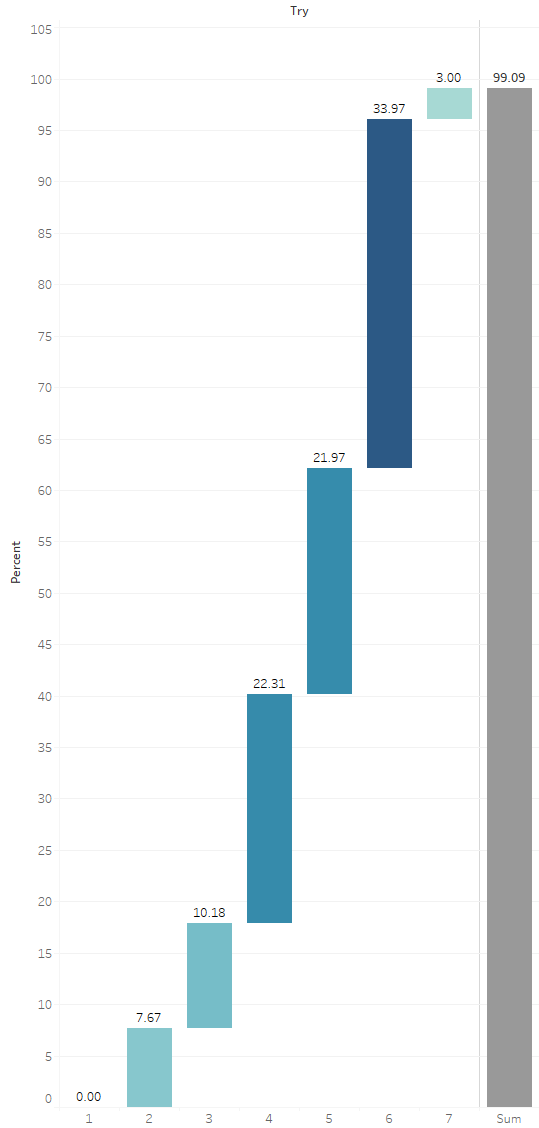
训练集的评估指标分别为RMSE=2.907，MAE=0.807，拟合优度fig:，可以认为我们优化的模型具有良好的拟合性。

测试集的评估指标分别为RMSE=2.208，MAE=1.094，拟合优度fig:，说明该模型的预测性十分准确。

### 3.3 预测EERIE的结果

使用上述模型，我们得出了2023年3月1日EERIE一词的预测结果（由于四舍五入的原因，total的结果可能并不符合100%）

EERIE的预测结果图



基于3.2节的评估结论，我们有很强的信心对于EERIE单词fig:分布的预测

### 3.4 uncertainty

尽管我们的预测模型具有很好的拟合度和预测准确性，但是该模型是基于2.1节的假设训练出来的模型，具有理论性，真实情况会更加复杂，有很多不受控的影响因素可能会对我们的预测结果产生冲击。举例来说，在推特上发布成绩这一行为具有很强的主观性，如果EERIE这一词对于很多人来说太过于复杂，可能会使人们的成绩降低，不好的成绩可能会导致他们不会在推特上上传自己的成绩。但是如果忽略这部分主观影响，在理论情况下，我对于我们的预测模型有很强的信心。

## 6. Q3的模型名字

在5.2节中，我们使用了预测模型来衡量单词与日期对（1，2，3，4，5，6，X）的影响程度，基于特征提取的十多种属性。基于此，在本节中，我们进一步介绍了能直接反应单词困难度指标——Average number of attempts——它直观地描述了人们猜出Wordle的一个单词所需要的平均次数。因此，我们对根据ATN的分布进行分析，并且通过ATN划分了不同的单词难度。最后应用5.1节的模型，可以对单词的难度进行分类。

原理

The normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is widely used in statistics and probability theory. Many natural phenomena, such as test scores, follow a normal distribution.

In addition, the central limit theorem states that the sum of many independent random variables, regardless of their distribution, tends to follow a normal distribution, provided that the sample size is large enough.

基于我们之前的发现（H\_r相对稳定）以及 Wordle 随机选择每天的单词，我们有理由猜测尝试次数的分布可接受为正态分布。因此，我们考虑根据根据正态分布的性质划为单词分难度。

为了验证上述猜想，我们需要对玩家尝试次数进行正态性检验。然而，数据集中只包含 1 try-7 or more tries (X) 的分布情况，难以直观地反应玩家对特定单词的尝试次数。

由此，我们提出平均尝试次数 ATN，以一种全局视角量化玩家的尝试次数：

ATN = ...

其中，7 or more tries (X) 的权重被设置为 10

基于此，我们的目标转化为对 ATN 进行正态性检验。

正态性检验

通常正态分布的检验方法有两种，一种是Shapiro-Wilk 检验，适用于小样本资料（样本量≤5000）；另一种是Kolmogorov–Smirnov检验，适用于大样本资料（样本量>5000）。由于数据集中包含 353 条记录（清洗后），我们选用 S-W 检验：

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

[Shapiro–Wilk test - Wikipedia](https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test)

### 详细结论

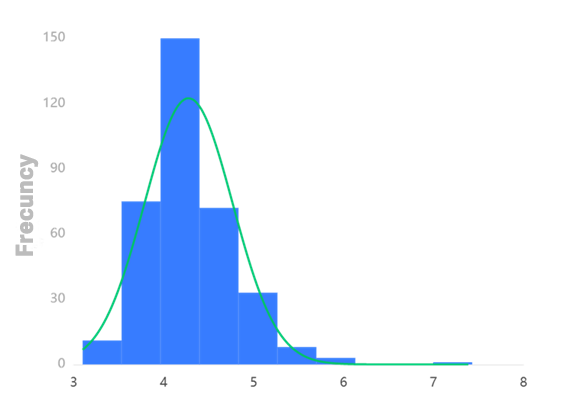
**输出结果1：总体描述结果**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变量名 | 样本量 | 中位数 | 平均值 | 标准差 | 偏度 | 峰度 | S-W检验 | K-S检验 |
| ATN | 353 | 4.21 | 4.277 | 0.489 | 1.359 | 5.171 | 0.931(0.000\*\*\*) | 0.077(0.027893892156622635) |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | | | | | | | | |

**图表说明：**

ATN采用S-W检验，显著性P值为0.000\*\*\*，水平呈现显著性，拒绝原假设，因此数据不满足严格正态分布。然而，通常现实研究情况下很难满足检验，ATN峰度（5.171）绝对值小于10并且偏度（1.359）绝对值小于3，应结合正态分布直方图、 PP图或者QQ图进行进一步分析。

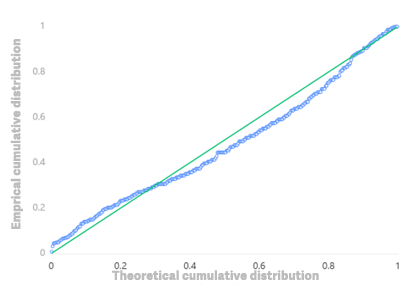
**输出结果2：正态性检验直方图**



**图表说明：**

上图展示了ATN数据的正态性检验直方图，若正态图基本上呈现出钟形（中间高，两端低），则说明数据虽然不是绝对正态，但基本可接受为正态分布。

**输出结果3：正态性检验P-P图**

****

**图表说明：**

上图是ATN计算观测的累计概率（P）与正态累计概率（P）的拟合情况。拟合程度越高越服从正态分布。

**输出结果4：正态性检验Q-Q图**

图表, 折线图

描述已自动生成

**图表说明：**

Q-Q图，全称“Quantile Quantile Plot”。用图形的方式比较观测值与预测值（假定正态下的分布）不同分位数的概率分布，从而检验是否吻合正态分布规律。并且将实际数据作为X轴，将假定正态时的数据分位数作为Y轴，作散点图，散点与直线重合度越高越服从正态分布，散点差异愈大越不服从正态分布。

综上，PP图和QQ表现出较好的你和情况，可接受ATN为正态分布。

难度的度量

The normal distribution is defined by two parameters: its mean μ and standard deviation σ. The mean determines the center of the distribution and the standard deviation determines the spread or width of the distribution.

通过上述正态性检验，我们可认定 ATN 的平均值 μ = 4.277，标准差 σ = 0.489 。我们根据 empirical rule 为 ATN 划分出三个区间：（改图）

图表, 直方图

描述已自动生成

[4.277-3\*0.489, 4.277-0.489)

[4.277-0.489, 4.277+0.489)

[4.277+0.489, 4.277+3\*0.489)

基于empirical rule，ATN 分布在以上三个区间内的概率为 99.73%。

ATN 68.27% 的数据落在 [4.277-0.489, 4.277+0.489) 内，我们认为该区间内的 ATN 反映了单词的中间难度—normal。且基于 Wordle 的机制，我们认为单词难度与 ATN 正相关。自此，我们自豪地宣布我们的难度分类模型：（改图）

if ATN 属于 xx: easy

elif ATN 属于 xx: normal

elif ATN 属于 xx: hard

难度的预测

得益于 x.x 节的工作，我们已经可以以较高的精度预测特单词在特定日期的玩家尝试次数分布（1 try-7 or more tries (X)），应用公式 X.X，我们可以得到特定单词的 ATN，进而由模型x.x得到其难度分类。

实例：预测2023.3.1日单词 EERIE 的难度分类（流程图）

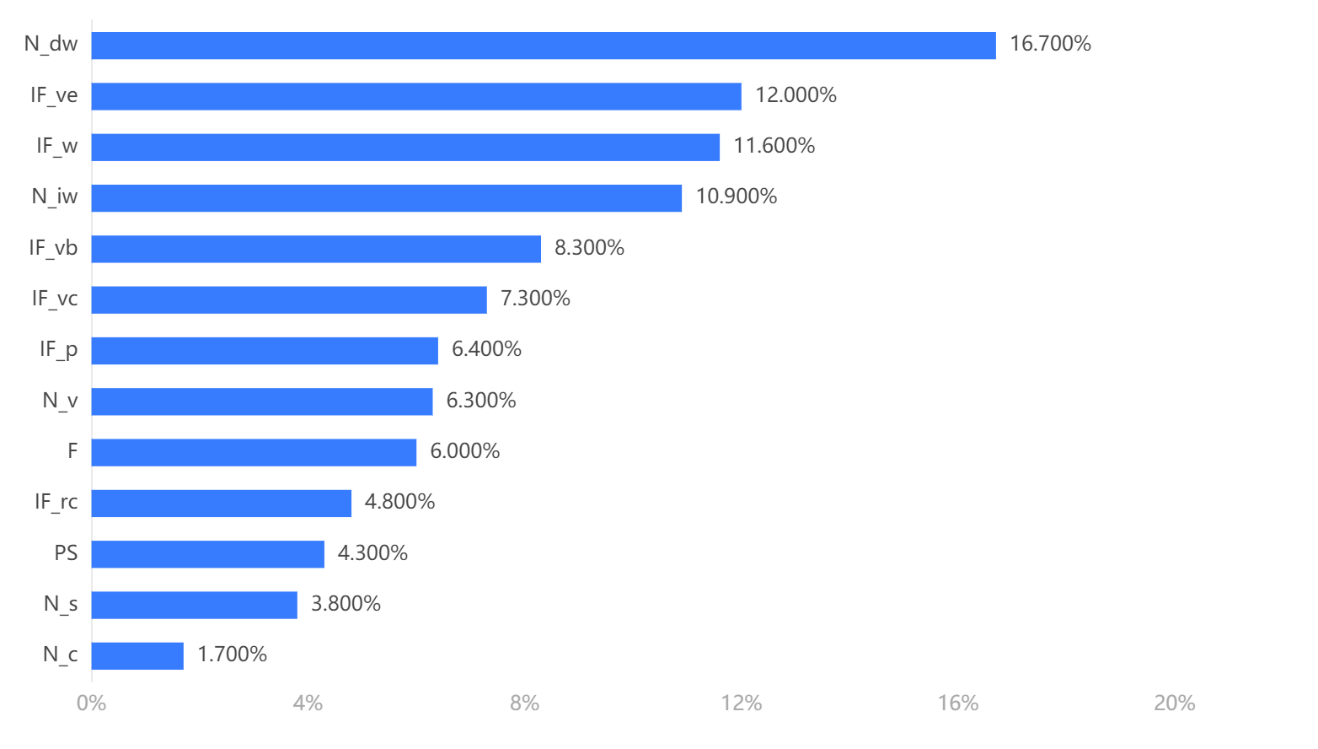
首先，通过我们的特征提取器，得到 EERIE 的 13 个属性：

第二步，将 13 个变量作为预测模型 x.x 的输入，得到 EERIE 的尝试次数分布：

第三步，根据公式 x.x 计算 EERIE 的 ATN = (7.67\*2+10.18\*3+22.31\*4+21.97\*5+33.97\*6+3\*10) \*0.01= 4.79

第四部，根据分类模型 x.x 得到 EERIE 的难度分类为hard。

确定重要属性

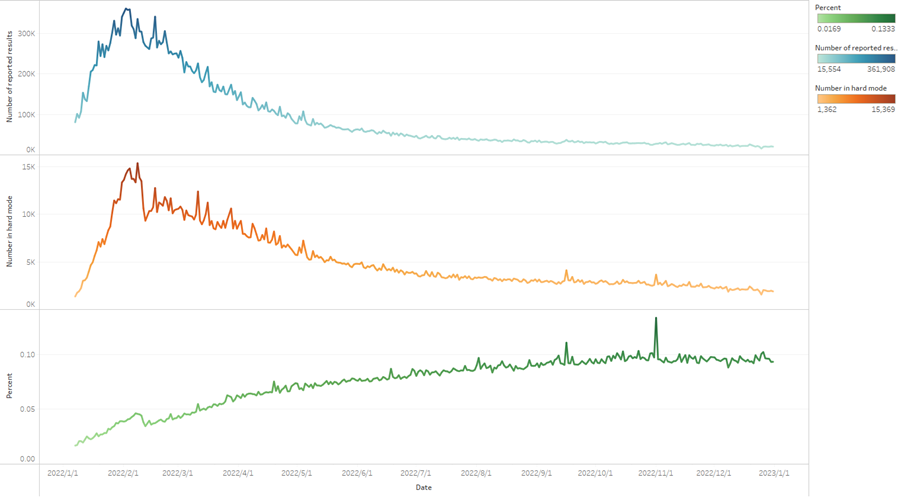


上柱形图展示了各属性在难度分类中的重要性比例，其中The Number of duplicate words N\_dw， Vowel beginning or not IF\_ve, Workday or not IF\_w and The number of infrequently used alphabets N\_iw 的重要性分别占到16.7%, 12.0%, 11.6% and 10.9%。这符合我们的一般认知，也和社交网站上的Wordle攻略有很高的相容性。

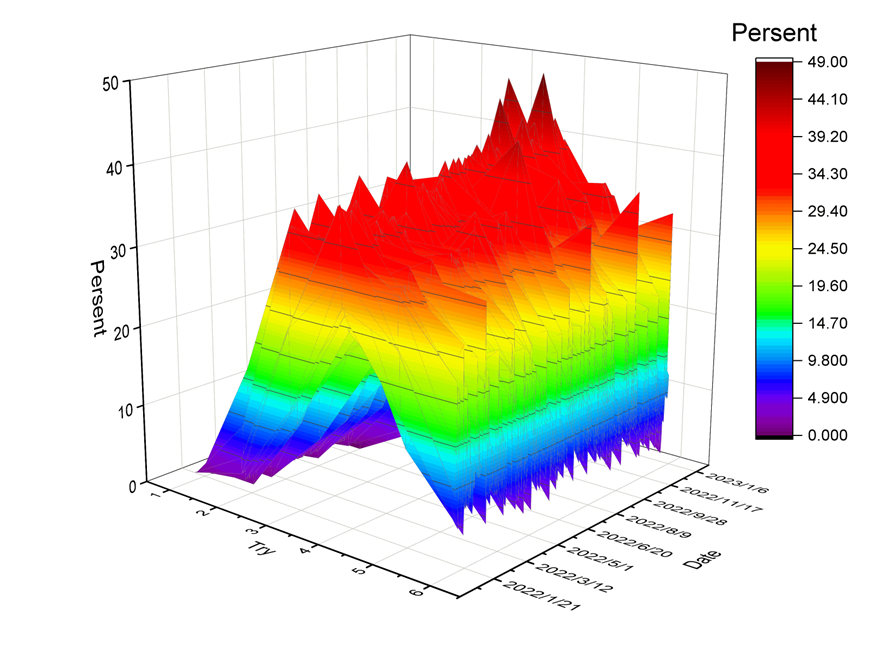
模型评估

由于分类模型 x.2 依赖于 ATN，而ATN依赖于预测模型 x.1。所以模型x的分类模型精度理论上取决于x.1。自处不再赘述，请参考x.x节的工作。

## 7. Data Insights

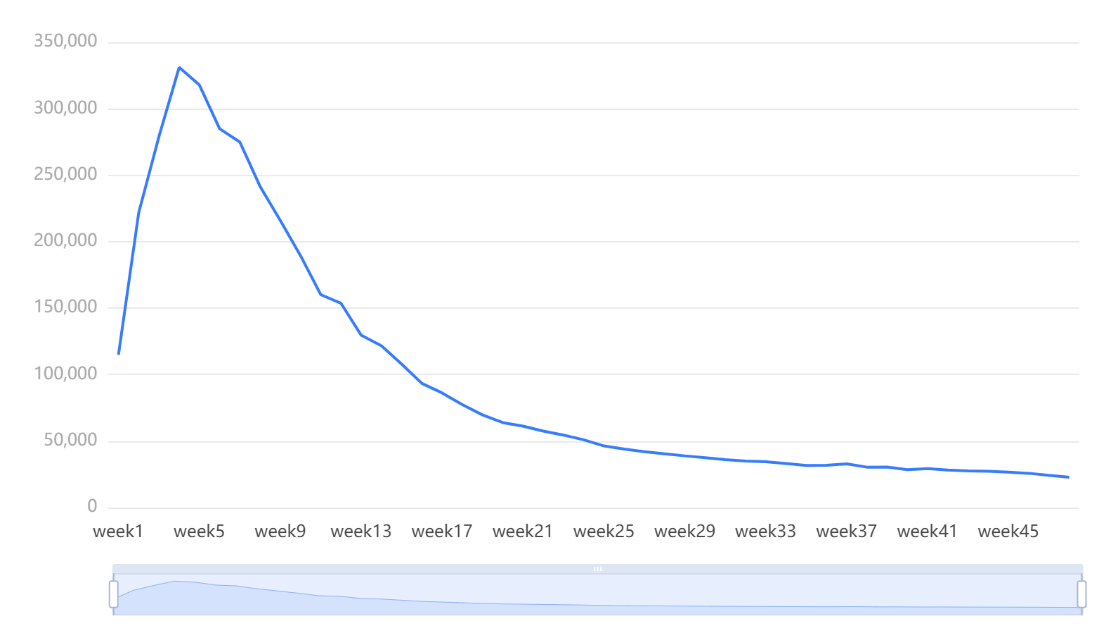


上面三张图分别是参与游戏人数、选择困难模式人数、选择困难模式人数占总人数的百分比随时间变化的数据图。从图中我们很容易看出，不管什么模式，每天游玩人数都呈现先上升后下降的趋势，这个游戏的热度相较之前有所下降。同时我们注意到，困难模式人数的百分比在不断小幅度上升，结合上面两幅图我们可以大致认为主要是由总人数下降所导致，但这一百分比也反映了每天仍有一批困难模式的忠实用户热衷于在猜词游戏中挑战自我。



上图反映的是尝试次数，日期和百分比之间相互数值关系，当日期固定在某一天观察尝试次数我们可以看出，其分布情况大致符合正态分布，百分比的每日最高点集中在2、3、4、5尝试次数之间，这也印证了我们在前面关于数据分布的猜想。所以1次与7次尝试次数的情况在排除个别玩家的异常情况下，可以认为占有极低的比例，这一特征与我们之前对于单词的预测得到了很好的互相印证。

## 8. SENSITIVITY ANALYSIS





上面两张图分别是ARIMA模型中的零阶差分图和一阶差分图，虽然两者在ADF检验表上的表现都很好，但通过图像我们还是可以看出一阶差分图所反映的是更为平稳的时间序列，这在我们随后的模型数据中得到了验证。

## 9. Strengths and Weaknesses

### 9.1 Strengths

- 充分利用信息:在设计我们的测量时，我们考虑了所提供数据中的单词的其他有用属性;

- 出色的模型表现:我们模型高拟合度和低误差值表明我们的模型具有出色的表现，并且我们的测量是准确的;

- 正确的方法选择:我们揭示了不用尝试次数之间的联系，这要归功于我们对方法的选择;

- 高鲁棒性:一般来说，我们的模型对参数的值变化并不敏感。

### 9.2 Weaknesses

- 特征提取的一定程度的随意性:我们在提取部分特征值时采用了随意性思维;

- 训练词集的局限性:我们在XGboost模型中用来训练的词集在一定程度上缺乏规模，可能会导致预测结果的偏差;

## 10. LETTER

Dear director,

​ 感谢您邀请我们为顾问!我们已了解您的具体要求，并已充分评估我们任务的可行性。在这里，我们向您展示我们的详细解决方案。我们根据Wordle上每天的单词，分析了影响单词难度的相关因素。我们发现，可以利用单词的许多属性，对单词困难度进行区分，供你跟踪。为了对困难度进行量化，提出了基于单词本身的分类模型，该模型引入了定义的数据度量——困难度。在计算单词难度时，我们考虑以下几个主要因素:

- word frequency

- Number of duplicate words

- The number of vowels

- Polysemous or not

- Vowel beginning or not

通过将这些因素结合在一起，我们可以对这个单词本身进行评估，并且我们通过优化XGboost模型，得到了单词对于尝试1次到7次的分布影响。换句话说，你可以直接通过单词的量化指标来按需调整游戏难度。

经验知识表明，游戏的运营与其玩家的数量密切相关。我们采用ARIMA来对该问题进行预测，事实证明，我们可以根据给出的玩家数量时间序列来精确地预测游戏在不久的将来的日玩家参与数量。

至于你所关心的，单词属性是否会影响玩家们对于困难模式的选择，我们通过计算困难模式玩家的选择比例，并且对其和单词属性进行了相关性分析，我们发现，单词的各个属性并没有表现出和困难模式的选择有强关联性。当然，由于数据量不够庞大，这个结论还需要更多数据支撑。

最后，我们通过对于玩家人数、困难模式人数、玩家总数的趋势观察，我们发现玩家的总人数虽然在开始上升迅速，但是随后的时间里面在不断下降，困难模式的选择率小幅度上升，说明流失的玩家大部分是普通玩家，他们并不是猜单词的专家。根据以上分析，我们向您提出了如下建议:

•在公司条件允许的情况下，我们建议你在玩法上提供新的思路，进而可以提高玩家们的热情;

•如果您的公司没有足够的精力(由于人手不足或预算有限)，我们建议您将有限的精力更多地投入到困难模式玩法的开发上（例如适度调高单词难度或者一天可以猜两个单词），因为您可以至少保证忠实用户的数量;

•无论如何，你的公司应该持续关注我们对Wordle的单词和用户分析，因为它们可以很好地反映玩家对于游戏本身的看法

If you want to know more details, please refer to our thesis. We will be glad to discuss with you on our solution details

Summary

In Wordle, the fun of guessing words and the potential social aspect of the game always attracts many players to the daily challenge. Therefore, it is crucial to increase the traffic to Wordle based on player data and word characteristics.

**Firstly**, based on an understanding of Wordle's rules and social network features, and by referring to authoritative English morpheme constructions and comprehensive Wordle cheat sheets, we propose heterogeneous spatio-temporal features with word topology-player level and fused social network activity, as well as quantifying the features following data normalization principles. **After** pre-processing the data, we constructed an **Autoregressive Integrated Moving Average(ARIMA) Model** to analyze the number of people playing Wordle each day in a time-series manner, and the model performed very well in terms of goodness of fit. In addition to fitting the existing data, further analysis was carried out to derive prediction intervals for future data. We then addressed the question of whether word attributes have an effect on the selection rate of difficult modes by using **Spearman's Correlation Analysis**, which proved that there was no significant correlation between the two, further indicating the orthogonality of the extracted word features and the selection rate of difficult modes.

**Then**, in order to predict the distribution of 1 try-7 or more tries (X), we propose a **Multiple-input and Multiple-output Extreme Gradient Boosting(MIMO XGBoost) Model** that takes into account their correlation and introduces a percentage sum of 1 constraint in the prediction. Specifically, we trained seven XGBoost regression models temporally synchronously to predict the distribution based on multiple spatio-temporal attributes related to words, i.e. MIMO, and designed a distributed loss function to ensure that the 7 models optimize the parameters towards a common goal. After joint training is completed, each model can be used asynchronously. Experimental results show that our proposed MIMO XGBoost Model brings the frequency closer to 1 than training the XGBoost models independently 7 times, while achieving better accuracy and goodness-of-fit on the test set. We predict the distribution of the number of player attempts for the word EERIE in 2023.3.1 and give the proportional importance of each attribute in difficulty classification, and critically, our MIMO XGBoost Model provides the basis for subsequent classification tasks.

**Next**, we defined the average number of attempts (ATN) as a characterization of the difficulty of a particular word for the player and performed a **Shapiro-Wilk Normality Test** on the ATN, which showed that the ATN was accepted as normally distributed. The ATN was then divided into three intervals according to Empirical Rule, corresponding to the three difficulty levels of easy, normal and hard. The task of classifying word difficulty is then transformed into a prediction task for ATN, which allows us to follow the MIMO XGBoost model described above in predicting the ATN of a given word, while maintaining a high classification accuracy consistent with the prediction task described above. Our classification model shows that EERIE has a hard difficulty.

**Finally**, we have collated and analyzed some interesting features of the dataset and provided the patterns from the above analysis and exploration as suggestions to the Puzzle Editor of the New York Times.

**Keywords:** the ARIMA Model; the MIMO XGBoost Model; Spearman's Correlation Analysis; Shapiro-Wilk Normality Test.

## 1. Introduction

### 1.1 Background

“The purpose of computation is insight, not numbers.”

-Richard Hamming

Wordle is one of the most popular word-guessing games in the world. In this game, you can successfully guess the only word of the day in six chances, and share your proud results on social platforms. In a word-guessing game, one of the most curious things is what the word of the day is, and that curiosity and the social nature of the game drive players crazy. But while playing, it's clear that the difficulty of a word can greatly affect the number of attempts a player makes that day, and potentially the traffic of subsequent games, so we hope to find some interesting patterns in our analysis of daily words and make our suggestions to the New York Times Wordle editors.

### 1.2 Problems Restatement

• The New York Times asked us to analyze the data provided to us and obtain valuable information, including:

o Explain the change in reported results over time and give a forecast for a specific date.

o Determine whether the word will affect the percentage of scores reported that were played in Hard Mode.

o Predict the distribution of a particular word on a particular day of the reported results

o Create a classification model to classify words by difficulty.

o Identify the attributes of a given word that are associated with each classification.

o Evaluate our model.

o Explore the characteristics of the dataset.

* In order to achieve the above goals, we propose a series of subtasks with precedence constraints. Specifically:
  + Make a statistical description of the relevant data, including the central tendency and discrete tendency of the data, and explore the characteristics of the data set.
  + Analyze the time series diagram of reported results and determine a suitable prediction model by combining the propagation characteristics of social networks.
  + According to the rules of the Wordle game, extract the features of words and determine the measures of each feature.
  + analyzes the correlation between the relevant features of the word and the difficult mode selection rate to determine the influence of the word on the mode chosen by the player.
  + Aiming at the propagation characteristics of social networks, features are extracted based on dates and their metrics are determined.
  + Predict the distribution of reported results based on relevant features for dates and words.
  + Determine a measure of word difficulty based on the rules of the Wordle game and classify the difficulty of words.

## 2. Assumptions & Nomenclature

### 2.1 Assumptions

We make several assumptions in our model. Later, we may relax these assumptions to optimize our model and make it more applicable to complex real-world environments.

The records people post on Twitter are accurate.

The data provided does not include incorrect data from competitors or malicious customers.

The number of players recorded per day is equal to the sum of the number of statistics, i.e., there are no omitted or excluded records in any of the provided datasets.

The record uploaded by the player was posted on the day of play.

Each player uploads records at most once a day.

### 2.2 Nomenclature

y\_i：Output value

\hat{y}\_i：Predicted output value

V\_d: Difference value

W\_i: Weight

PS： Part of speech词性

：Word frequency词频

ATN：Average number of attempts平均尝试次数

L\_d：Difficulty level难度等级

R\_h：Percentage in hard mode

R\_{ti}：Percentage of i tries

N\_{dw}：The number of duplicate words重复字母个数

N\_v：The number of vowels元音个数

N\_c：The number of consonants辅音个数

N\_{iw}：The number of infrequently used words不常用单词个数

N\_s：The number of syllables音节数

IF\_{rc}：Repetition is continuous or not重复是否连续

IF\_{vb}：Vowel beginning or not是否元音开头

IF\_{ve}：Vowel ending or not是否元音结尾

IF\_{vc}：Vowel is continuous or not元音是否连续

IF\_{p}：Polysemous or not是否是多义词

IF\_{w}：Workday or not是否是工作日

## 3. Data Preprocessing & Modeling Framework

### 3.1 Data Cleaning

By observing the provided data, we found that there are some outliers in the dataset. The Word attribute contains two words with only four letters and one word with six letters, which doesn't follow the rules of the game for guessing five-letter words. There are also numerically anomalous words, for example, in the 1 try-7 tyies attribute, the total percentage of attempts is much greater than 100. Similarly, in the Number of reported results property, the count of two words is well below the average by an order of magnitude. Considering that these words may hurt our modeling, we ignore these anomalous words.

Finally based on the above process, we summarize the file (word\_normalization) contains all the words that appeared, and the properties of the words standardization.

### 3.2 Feature Extraction & Quantification

We only have attributes such as date, word, number of reported results, and distribution of player attempts. In order to predict the distribution of player attempts. we need more attributes to build our model.

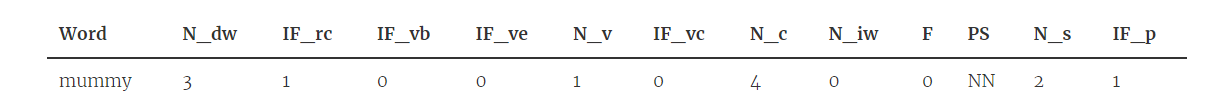
According to the rules of Wordle and the characteristics of social networks, we extract the features of data from two dimensions of space and time:

In the spatial dimension, Wordle is an alphabet-oriented English word puzzle game, which determines that the topology of words greatly affects the distribution of the number of attempts by human players. Based on the analysis of Wordle recipes on social networking sites and the morpheme formation of English words, we propose to use the following features to measure the topology of words:：

文本

描述已自动生成

图为mummy单词的量化结果（图）。



On the other hand, the difficult mode is also an important factor affecting the number of attempts. We introduce a difficult mode selection rate:

H\_r = num of rr in hard mode / num of rr

In the temporal dimension, social network activity may affect the number of participants and the level structure of the players, which may affect the distribution of the number of attempts. We select the trading day data of the New York Stock Exchange to label the date, and introduce the Workday or not attribute as the representation of social network activity.

In summary, we propose heterogeneous spatiotemporal features with word topology-player level and fused social network activity, and follow the principle of data standardization to quantify the features to obtain quantitative and categorical (beginning with IF) attributes.

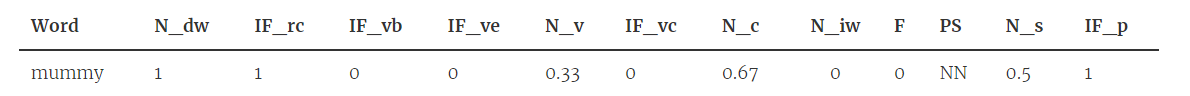
### 3.3 Data standardization

Standardization is a necessary step. Here we use the Min-Max Normalization method, which uses raw data based on its maximum value, Max, and minimum value, Min. The processed data fit the interval distribution from 0 to 1. The conversion function is

\frac{x-x\_{\min }}{x\_{\max }-x\_{\min }}

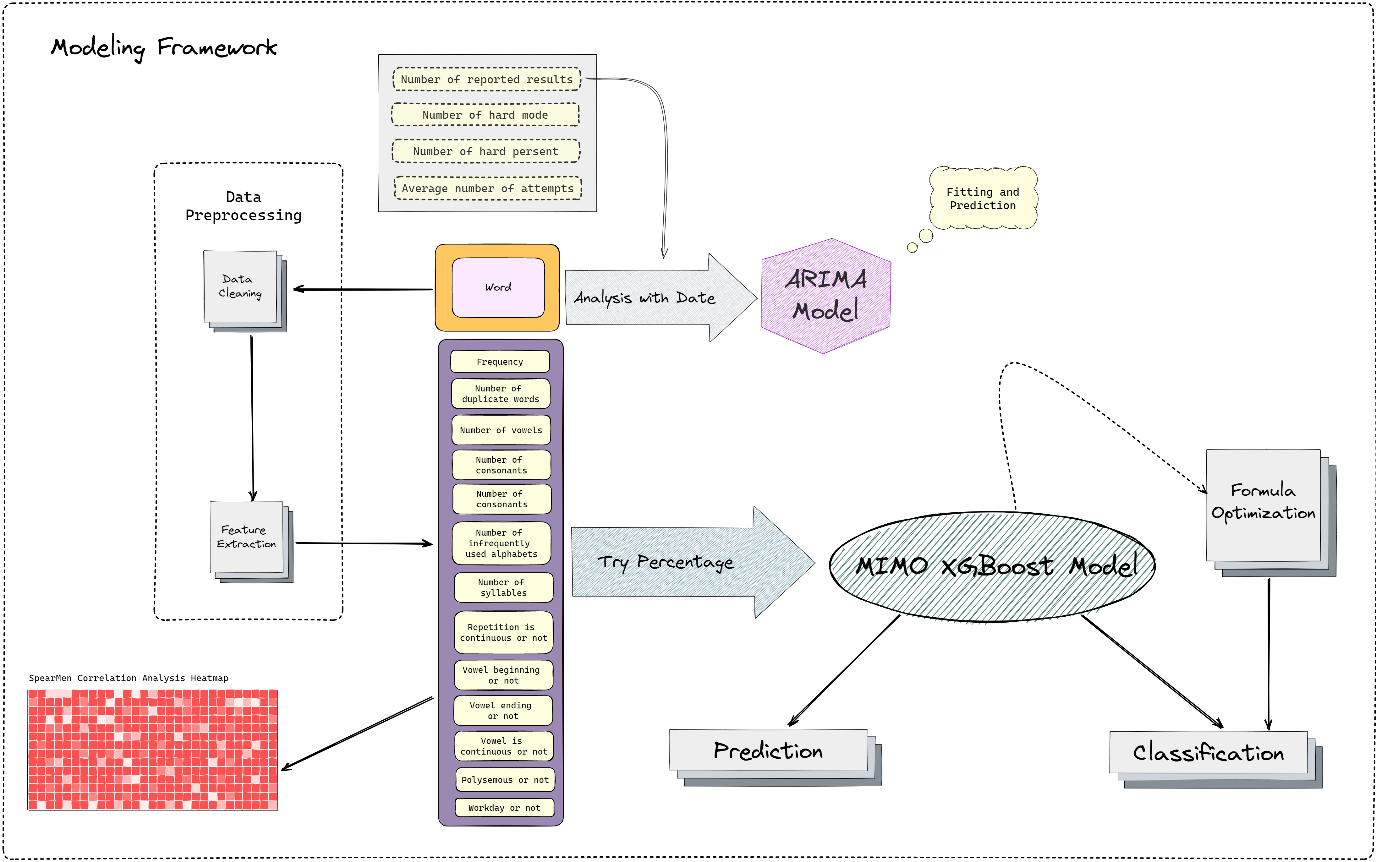
After the standardization step, the standard data of uniform interval is obtained (The results for mummy words are shown in Table 3)。

Table 3？：Standardized data



### 3.4 Modeling Framework

Our modeling framework can be illustrated as shown in Figure 1.



## 4. ARIMA Model

In this section, we use the Autoregressive Integrated Moving Average(ARIMA) prediction model as a data simulation measure for the future online number of Wordle. Based on the number performance of Wordle in historical time, At the same time, the lagged moving average is used to obtain a smooth time series, so as to make a rough estimate of the number of people online at a certain time in the future. We then performed feature extraction on the word attributes, using the similarity matrix to analyze whether any attribute of the word affects the percentage of the total number of people participating in the difficult pattern.

## Model Review

### 1.Auto Regression(AR)模型

The autoregressive model describes the relationship between the current value and the historical value and uses the historical time data of the variable to predict itself.

In general, the formula of the P-order autoregressive process is defined as follows:

y\_{t}=\mu+\sum\_{i=1}^{p} \gamma\_{i} y\_{t-i}+\epsilon\_{t}

手表上有字

中度可信度描述已自动生成

General P-order autoregressive model AR:

x\_{t}=\alpha\_{1} X\_{t-1}+\alpha\_{2} X\_{t-2}+\ldots+a\_{p} x\_{t-p}+μ\_{t}



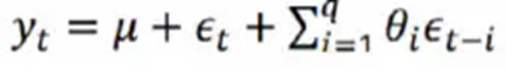
As can be seen from the formula, the current value is predicted by the historical value, and p is an order in the autoregressive model, which indicates that the historical value of several epochs is used to predict the current value.

### 2. Moving Average(MA)模型

Moving averages focus on the accumulation of error terms in autoregressive models.

Among them, the formula of the q-order self-MA model is defined as follows:

y\_{t}=\mu+\epsilon\_{t}+\sum\_{i=1}^{q} \theta\_{i} \epsilon\_{t-i}



In the AR model, if it is not white noise, it is usually considered to be a moving average of order q:

u\_{t}=\varepsilon\_{t}+\beta\_{1} \varepsilon\_{t-1}+\ldots+\beta\_{q} \varepsilon\_{t-q}



The moving average method can effectively eliminate random fluctuations in the forecast

## 3. 自回归移动平均(ARMA)模型

The autoregressive moving average model consists of two parts: the autoregressive part and the moving average part. The regression equation is expressed as follows:

y\_{t}=\mu+\sum\_{i=1}^{p} \gamma\_{i} y\_{t-i}+\epsilon\_{t}+\sum\_{i=1}^{q} \theta\_{i} \epsilon\_{t-i}

黑色的钟表

中度可信度描述已自动生成

It can be seen from the regression equation that the autoregressive moving average model combines the advantages of AR and MA. In the ARMA model, the autoregressive process is responsible for quantifying the relationship between the current data and the previous data, and the moving average process is responsible for solving the problem of random changes.

## 4.差分自回归移动平均(ARIMA)模型-ARIMA(p，d，q)

The Autoregressive model (AR), the moving average model (MA), and the difference method (I) are combined to obtain the differential autoregressive moving average model ARIMA (p, d, q), where d is the order of the difference to be performed on the data, and ARIMA is the ARMA model after the difference

## Model Analysis

Step size selection:

Considering that the ARIMA model has high uncertainty for the result of too long prediction, we add a sliding window to explore the number of future players, and the step size is set to 7 so that the time unit of prediction is changed from days to weeks, which can greatly reduce the number of predictions.

文本

描述已自动生成

1. ARIMA model requires that sequences meet stationarity. Check the results of the ADF test, and analyze whether it can significantly reject the hypothesis of sequence instability according to the analysis t value (P<0.05).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ADF test table | | | | | | | |
| variables | Order of difference | t | P | AIC | Critical value | | |
| 1% | 5% | 10% |
| Average Number | 0 | -3.276 | 0.016\*\* | 651.242 | -3.616 | -2.941 | -2.609 |
| 1 | -4.002 | 0.001\*\*\* | 627.622 | -3.616 | -2.941 | -2.609 |
| 2 | -3.867 | 0.002\*\*\* | 620.775 | -3.621 | -2.944 | -2.61 |
| Note: \*\*\*, \*\*, \* represent significance levels of 1%, 5%, and 10% respectively | | | | | | | |

By looking up the table and comparing the difference sequence images, we conclude that the best difference value is order 1. When the difference is order 1, the significance P value is 0.001\*\*\*, showing significance on the level, rejecting the null hypothesis, and the series is a stationary time series.

图表, 折线图

描述已自动生成

Optimal difference map

2. Check the comparison chart of the data before and after differences to determine whether it is stable. At the same time, the time series is biased (autocorrelation analysis), and its p and q values are estimated according to the censoring situation, as shown in the following figure:

（两图合并）

图表

描述已自动生成

**Final differential data autocorrelogram (ACF)**

图表

描述已自动生成

**Final differential data Partial Autocorrelation Map (PACF)**

We finally determined the ARIMA model as ARIMA(2,1,0). The model parameter table is as follows. The table shows the results of this model test, including the sample number, degrees of freedom, Q statistics, and the goodness of fit of the information criterion model.

|  |  |  |
| --- | --- | --- |
| **ARIMA model (1,1,0) test table** | | |
| **item** | **Symbol** | **value** |
|  | Df Residuals | 45 |
| Number of samples | N | 48 |
| Q statistic | Q6(P value) | 3.317(0.069\*) |
| Q12(P value) | 13.838(0.031\*\*) |
| Q18(P value) | 18.141(0.111) |
| Q24(P value) | 18.315(0.435) |
| Q30(P value) | 18.384(0.784) |
| Information criterion | AIC | 1041.317 |
| BIC | 1046.867 |
| Goodness of fit | R² | 0.977 |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | | |

Finally, we obtain the time series plot and prediction interval plot for the simulated average number of people, as shown in the following figure:

（两个图和一个）

图表, 折线图

描述已自动生成

*（放大右下角的图）*

图表, 折线图

描述已自动生成

The prediction interval is obtained by predicting the minimum number of players and the maximum number of players per week and taking their numbers as the left and right critical values of the interval to obtain a more reliable value.

## The similarity matrix

Conjecture: Since in the Wordle game, the player cannot know which word is until the end of the game, this should not affect the player's choice of a difficult mode. Based on this conjecture, we carry out the following analysis.

In analyzing the correlation between word attributes and the percentage of people who participated in difficulties, we chose to use Spearman correlation analysis. Pearson correlation analysis is mainly used to analyze the relationship between two quantitative variables that meet normal distribution. If the two variables include rank variables, or the variables do not meet normal distribution, or the distribution type of the variables is unknown, Spearman correlation analysis is a more appropriate analysis method. The basic idea of Spearman correlation analysis is: rank transformation is performed on two variables X and Y respectively, denoted by rank order RX and RY, and the average rank is taken as the rank of each data. The calculation formula of Spearman correlation coefficient rs is:

Analyze the following figure:

图表, 表格, 树状图

描述已自动生成

It can be clearly seen from the figure that there is a very weak correlation and the irrelevant relationship between the percentage of people who participate in difficult mode and the attributes of words, so it can be concluded that the attributes of words do not affect the number of people who choose to participate in difficult mode on the day of the game.

## 5. MIMO XGBoost Model

The model should not only use the feature values extracted in Section 3.1 but also the time dimension is an important influencing factor. We summarize the $IF\_w$ attribute according to the change in the activity of the social network over time. In summary, we optimize the XGBoost regression prediction model and based on the conclusion of Section 3.1, we take the attributes of words and time attributes as input values, predict the distribution of (1,2,3,4,5,6, X), and quantitatively test the accuracy of the model.

### 5.1 XGBoost Model Analysis

### 5.1.1 Decision Tree Ensembles

A tree ensemble model consists of a set of classification and regression trees (CART), mathematically speaking, we can write the model in the following form:

Where K is the number of trees, f is the space of functions of F, and F is the set of CART possibilities. The objective function of the above model is given by the following equation：

Where the first term is the loss function, and the second term is the regularization parameter.

### 5.1.2 Tree Boosting

For the current tree model, the learning method is to define the objective function and optimize it. We h\ave defined the objective function above. Considering the difficulty of learning the tree structure, we also adopt the addition strategy to optimize it, adding a new tree each time:

Then select the tree that can optimize the objective function, and consider using the mean square error (MSE) as the loss function:

The loss function is obtained by Taylor expansion to second order and removing all constants:

Under this definition, the value of the objective function is determined only by 𝑔 𝑖 and ℎ 𝑖. This is how XGBoost supports custom loss functions.

### 5.1.3 The Structure Score

After reformulating the tree model, we can write the target value as:

If the structure part q of the tree is known, the objective function can be used to find the optimal Wj and obtain the optimal objective function value. Its essence can be reduced to the problem of solving the minimum value of quadratic function. The solution is:

## 5.2 Model optimization

### 5.2.1 Deficiency of XGBoost regression prediction model

The XGBoost model we are using here is a multi-input single-output model that can only predict one/at a time. If only this model is used, the different/are not correlated and may result in the total percentage sum of the forecast being much greater than or less than 100, with the subscript of the word black:

Application of XGBoost to the word black

|  | 1 try | 2 tries | 3 tries | 4 tries | 5 tries | 6 tries | 7 or more tries (X) | total |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **percentage** | 0 | 8.9793 | 27.9642 | 32.1828 | 15.0472 | 5.0703 | 1.0029 | 90.2467 |

Using a simple XGBoost regression forecasting model to forecast the world, find the sum of forecast from 100, the deflection, in other words, 𝑅\_{ti}lack of correlation between.

### 5.2.2 Optimize the XGBoost model

In practice, the relationship between R\_{ti} is as follows:

Because of rounding, the error value of the final data is allowed to be 1.

This, in turn, can calculate the difference between the predicted values and 100 value V\_d, computation formula is as follows:

For this model, each predicted value has an impact on V\_d, and their respective weights are w\_i:

Based on the weight of the predicted value and difference value, the influence value of each predicted value on the deviation value (V\_{di}) is given. The relevant formula is as follows:

Formula 1 is added into the objective function of the XGBoost regression model (Formula 3) as a supplement to the cost function, then the optimized XGBoost regression model can be obtained, and the optimized model can conform to the correlation between each R\_{ti}, the formula is as follows:

Taking the Taylor expansion of the loss function to second order and removing all the constants yields:

The value of the objective function only depends on g\_i and h\_i. In line with the way XGBoost supports custom loss functions.

## 3.模型评估

### 3.1 model measurement

Using 70% of the data in the word\_normalization file as the training set and the rest as the test set, the Percentage of the i tries () in the test set is calculated, and the predicted value of each tries is obtained. In Table i, the left side is the predicted value, and the right side is the real value.

3 tries Partial table of test samples versus real samples

|  | predicted | true value |
| --- | --- | --- |
| **begin** | 26.021 | 26 |
| **being** | 21.982 | 22 |
| **berth** | 23.769 | 24 |
| **black** | 30.904 | 31 |
| **booze** | 6.988 | 7 |

From the above table, we can see that all the predicted results and the true results are very different, only 0.24 difference at most.

### 3.2 quantitative evaluation

We use several common statistical parameters to quantitatively test the conclusions obtained from the data. Root Mean Square Error () is the first measure we choose to use, which can effectively test the accuracy of the predicted value. The related formula is as follows:

手机屏幕截图

中度可信度描述已自动生成

In addition, we choose the Mean Absolute Error (MAE) as the accuracy measure. MAE can reflect the actual situation of the predicted value error, and the relevant formula is as follows:

墙上的钟表

中度可信度描述已自动生成

Finally, the conventional goodness-of-fit R^2 is selected as the overall evaluation of model fitting. The following table shows the test result statistics of the test set and the training set:

EERIE的评估结果

|  | RMSE | MAE | R^2 |
| --- | --- | --- | --- |
| **training set** | 2.907 | 0.807 | 0.95 |
| **test set** | 2.208 | 1.094 | 0.94 |

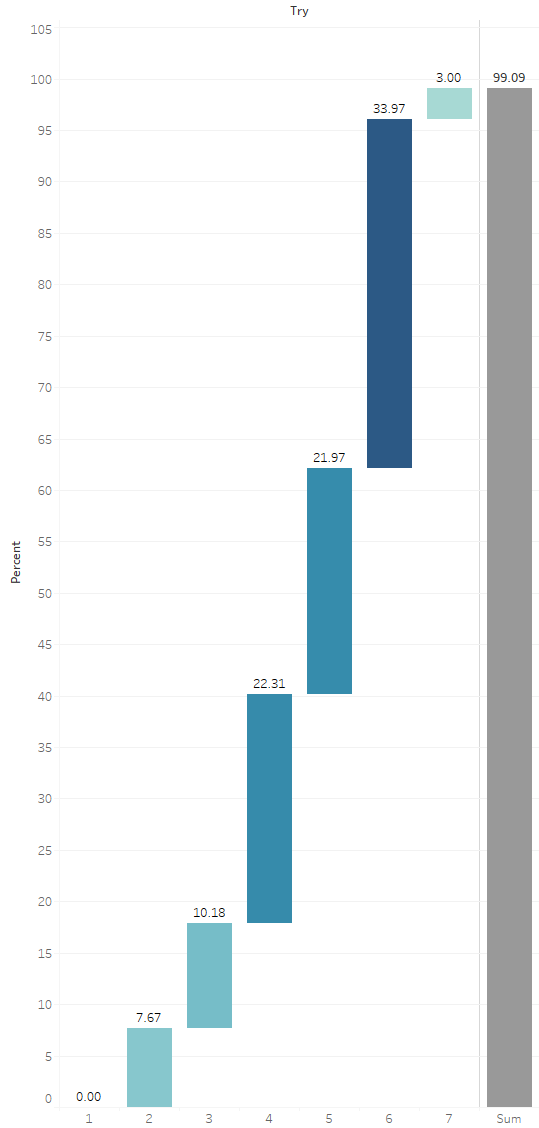
The evaluation indexes of the training set were RMSE=2.907, MAE=0.807, fig:，It can be considered that our optimized model has a good fit.

The evaluation indexes of the test set are RMSE=2.208, MAE=1.094, and the goodness of fit indicates that the prediction of the model is very accurate.

### 3.3 预测EERIE的结果

Using the model above, we produce a prediction for EERIE on March 1, 2023 (the result for total may not match 100% due to rounding)

Plot of the prediction results for EERIE



Based on the evaluation conclusions of Section 3.2, we have strong confidence in the prediction of the distribution of EERIE words $R\_{ti}$

### 3.4 uncertainty

Although our prediction model has a good degree of fit and prediction accuracy, it is a theoretical model trained based on the assumptions in Section 2.1. The real situation is more complex, and there are many uncontrollable factors that may impact our prediction results. For example, the act of Posting grades on Twitter is highly subjective. If the word EERIE is too complicated for many people, it may reduce people's grades, and bad grades may cause them not to upload their grades on Twitter. But if we ignore this subjective impact, I have strong confidence in our predictive model in the theoretical case.

6. Empirical Rule Based Classification Model

In Section 5.2, we used predictive models to measure the influence of a word versus date pair (1,2,3,4,5,6, X), based on more than a dozen attributes extracted from features. In this section, we introduce a direct word difficulty metric called Average number of attempts, which intuitively describes the average number of times it takes a person to guess a word of Wordle. Therefore, we analyzed the distribution according to ATN, and divided different word difficulty by ATN. Finally, applying the model from section 5.1, we can classify the difficulty of a word.

### Principle

The normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is widely used in statistics and probability theory. Many natural phenomena, such as test scores, follow a normal distribution.

In addition, the central limit theorem states that the sum of many independent random variables, regardless of their distribution, tends to follow a normal distribution, provided that the sample size is large enough.

Based on our previous findings (H\_r is relatively stable) and the fact that Wordle randomly selects words for each day, it is reasonable to guess that the distribution of attempts is acceptable as a normal distribution. Therefore, we consider dividing the difficulty of words according to the properties of the normal distribution.

To verify the above conjecture, we need to test the normality of the number of player attempts. However, the dataset only contains the distribution of 1 try-7 or more tries (X), which makes it difficult to intuitively reflect the number of attempts a player makes on a particular word.

Therefore, we propose the Average number of attempts ATN to quantify the number of attempts made by a player in a global perspective:

ATN = ...

Where the weight of 7 or more tries (X) is set to 10。Based on this, our goal is transformed into normality test for ATN.

## Test for normality

Generally, there are two test methods for normal distribution. One is Shapiro-Wilk test, which is suitable for small sample data (sample size &le 5000). The other is the Kolmogorov-Smirnov test, which is suitable for large samples (sample size &gt 5000). Since the dataset contains 353 records (cleaned), we use the S-W test:

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

[Shapiro–Wilk test - Wikipedia](https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test)

### 详细结论

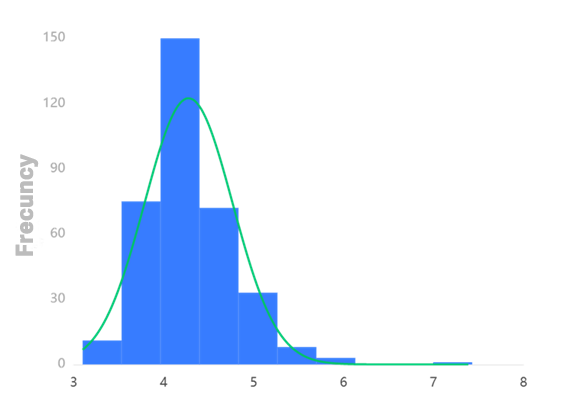
**Output result 1: general description result**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **variable** | Sample size | **median** | Average value | Standard deviation | skewness | 峰度 | S-W检验 | K-S检验 |
| ATN | 353 | 4.21 | 4.277 | 0.489 | 1.359 | 5.171 | 0.931(0.000\*\*\*) | 0.077(0.027893892156622635) |
| Note: \*\*\*, \*\* and \* represent the significance level of 1%, 5% and 10% respectively | | | | | | | | |

**Chart description:**

$ATN$ was tested by S-W test, the significance P value was 0.000\*\*\*, the level showed significance, and the null hypothesis was rejected, so the data did not meet the strict normal distribution. However, it is usually difficult to meet the test in real research situations, the absolute value of $ATN$ kurtosis (5.171) is less than 10 and the absolute value of skewness (1.359) is less than 3, and the normal distribution histogram, P-P plot or Q-Q plot should be combined for further analysis.

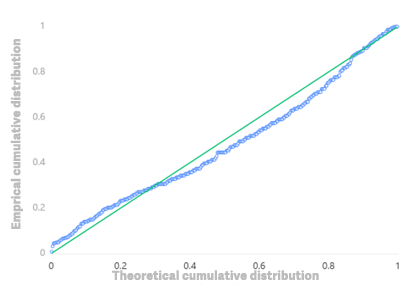
**Output 2: Normality test histogram**



**Chart description:**

The above plot shows the normality test histogram for the ATN data.If the normal plot has a bell-shaped shape (high in the middle, low at both ends), the data is not absolutely normal, but it is generally accepted as normally distributed.

**Output 3:** **Normality test P-P plot**

****

**Chart description:**

The above figure shows the fit of the cumulative probability (P) of the observation calculated by ATN to the normal cumulative probability (P). The higher the degree of fit, the more normal distribution.

**Output result 4: Normality test Q-Q plot**

图表, 折线图

描述已自动生成

**Chart description:**

The Q-Q Plot stands for "Quantile Quantile Plot". Test for normal distribution by comparing the probability distributions of different quantiles of the observed and predicted values (assuming normal distribution). The actual data is taken as the X-axis, and the quantile of the assumed normal data is taken as the Y-axis to make a scatter plot. The higher the coincidence degree between the scatter and the line, the more normal distribution is followed, and the larger the difference between the scatter points, the less normal distribution is followed.

In summary, the PP plot and QQ show a better situation of you and sum, and the ATN can be accepted as normal distribution.

**Measure of difficulty**

The normal distribution is defined by two parameters: its mean μ and standard deviation σ. The mean determines the center of the distribution and the standard deviation determines the spread or width of the distribution.

According to the above normality test, we can conclude that the mean value μ = 4.277 and the standard deviation σ = 0.489 for ATN. We divide ATN into three intervals according to the empirical rule:

图表, 直方图

描述已自动生成

[4.277-3\*0.489, 4.277-0.489)

[4.277-0.489, 4.277+0.489)

[4.277+0.489, 4.277+3\*0.489)

Based on the empirical rule, the probability that ATN is distributed in the above three intervals is 99.73%.

68.27% of the data of ATN fall within [4.277-0.489, 4.277+0.489), we believe that the ATN in this range reflects the middle difficulty of the word -- normal, and based on Wordle's mechanism, we believe that word difficulty is positively correlated with ATN. Since then, we are proud to announce our difficulty classification model:

if ATN 属于 xx: easy

elif ATN 属于 xx: normal

elif ATN 属于 xx: hard

## 难度的预测

Thanks to the work in X.X, we can already predict with high accuracy the distribution of player tries (1 try-7 or more tries (X)) for a given word on a given day.Applying the formula X.X, we can get the ATN for a given word, and the difficulty classification can be obtained from the model x.x.

实例：预测2023.3.1日单词 EERIE 的难度分类

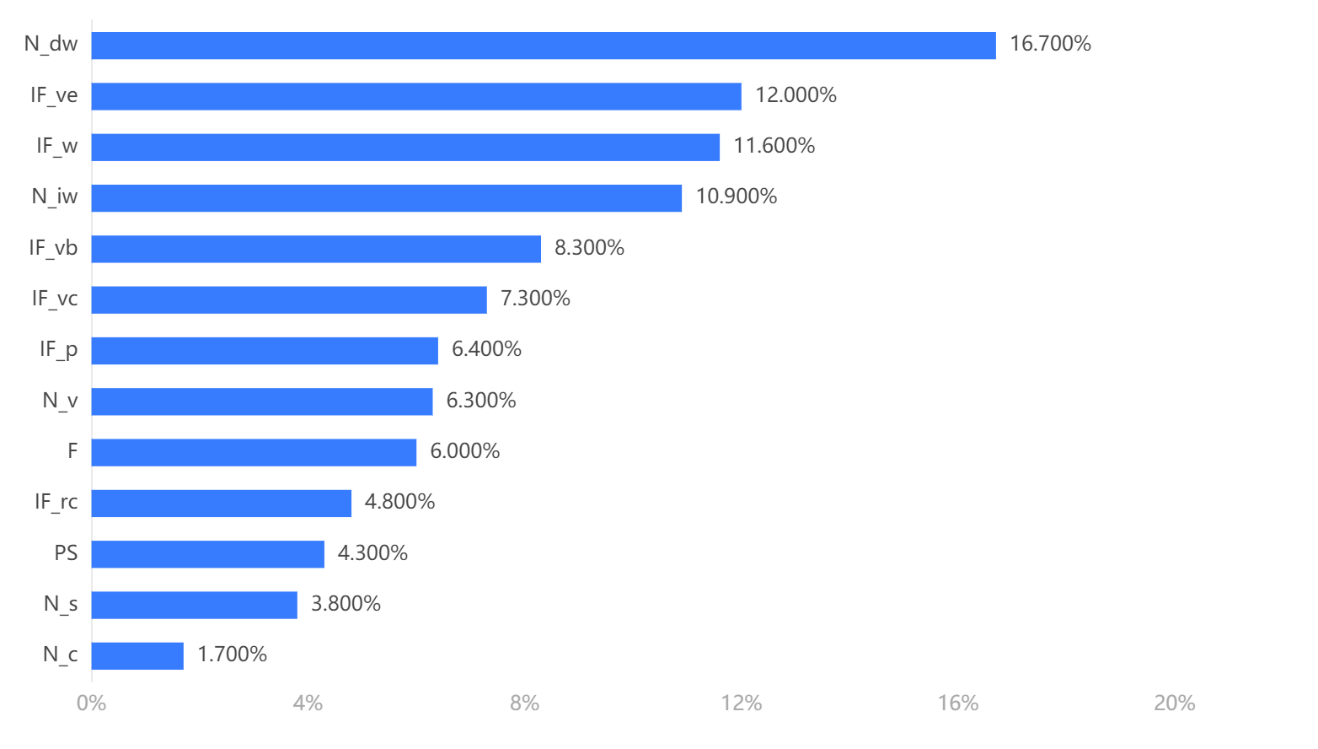
First, using our feature extractor, we get the 13 attributes of EERIE:

In the second step, the 13 variables are used as input to the prediction model x.x to obtain the distribution of attempts for EERIE:

Step 3: Calculate the ATN of EERIE according to formula x.x = (7.67\*2+10.18\*3+22.31\*4+21.97\*5+33.97\*6+3\*10) \*0.01= 4.79

In the fourth part, the difficulty classification of EERIE is obtained as hard according to the classification model x.x.

Identify important attributes



中The upper bar chart shows the proportion of importance of each attribute in the difficulty classification, where The Number of duplicate words N\_dw, Vowel beginning or not IF\_ve, Workday or not IF\_w and The number of hours used alphabets N\_iw accounted for 16.7% respectively, 12.0%, 11.6% and 10.9%. This is in line with our general understanding, and also has a high compatibility with the Wordle recipes on social networking sites.

### Model evaluation

Since the classification model x.2 depends on the ATN and the ATN depends on the prediction model x.1. So the classification model accuracy of model x theoretically depends on x.1. Without further elaboration, please refer to the work in section x.x.

### 模型评估

Since the classification model x.2 depends on the ATN and the ATN depends on the prediction model x.1. So the classification model accuracy of model x theoretically depends on x.1. Without further elaboration, please refer to the work in section x.x.

## 7. Data Insights

图表, 折线图

描述已自动生成

The above three graphs are the data graphs of the number of people who participate in the game, the number of people who choose the difficult mode, and the percentage of people who choose the difficult mode over time. From the figure, we can easily see that no matter what mode, the number of daily players first increases and then decreases, and the popularity of the game is lower than before. At the same time, we notice that the percentage of people in difficult mode is increasing slightly. Combined with the above two figures, we can roughly think that it is mainly caused by the decline of the total number of people, but this percentage also reflects that there are still a number of loyal users in difficult mode who are keen to challenge themselves in the word guessing game every day.

图表, 表面图

描述已自动生成

The graph above shows the numerical relationship between the number of attempts, the date, and the percentage. When we look at the number of attempts on a given day, we can see that the distribution is roughly normal, with the daily high points of the percentage concentrated between 2, 3, 4, and 5 attempts, which confirms our conjecture about the distribution of the data. So the number of 1 and 7 attempts can be considered to be very low when excluding the exception of individual players, and this feature is in good agreement with our previous prediction for words.

## 8. SENSITIVITY ANALYSIS

图表, 折线图

描述已自动生成



The above two graphs are the zeroth-order difference graph and the first-order difference graph in the ARIMA model. Although both of them perform well on the ADF test table, we can still see from the images that the first-order difference graph reflects a more stable time series, which is verified by our subsequent model data.

## 9. Strengths and Weaknesses

### 9.1 Strengths

-Make the most of the information: when designing our measure, we considered other useful properties of the words in the provided data;

- Excellent model performance: Our high model fit and low error values indicate that our model performs well and that our measurements are accurate;

-Right method choice: We reveal a link between the number of no-tries, thanks to our choice of method;

-High robustness: In general, our model is not sensitive to changes in the values of the parameters.

### 9.2 Weaknesses

- Some degree of randomness in feature extraction: we used random thinking when extracting some feature values;

- Limitations of the training word set: The word set we used to train in the XGboost model lacks scale to some extent, which may lead to biased prediction results;

## 10. LETTER

From: Team #2307004 To: Puzzle Editor of the New York Times Date: February 20, 2023

Dear Puzzle Editor of the New York Times,

Thank you very much for inviting us to conduct a data analysis of this interesting word-guessing game by Wordle. After understanding your specific requirements, we fully evaluated the feasibility of the task and are pleased to share our results with you.

Firstly, after reading Wordle's rules in detail and processing and visualizing the data you provided, we extracted a set of attributes for the words, which we guessed would be an important measure of word difficulty through the extraction of word features. When analyzing the characteristics of the dataset we saw trends over time in data such as the number of players per day, which inspired us to use the Autoregressive Integrated Moving Average (ARIMA) model to fit the current player data and predict the interval of change in the number of players over time. We believe that we can achieve the same excellent results. Also, in response to your question about whether word attributes have an effect on difficulty mode selection rates, we have used Spearman's Correlation Analysis to make it clear that there is no significant correlation between the two.

We then concluded that the average number of attempts (ATN) could be used to some extent as a characterization of word difficulty, and after performing the Shapiro-Wilk normality test, we used the Empirical Rule to classify words as easy, normal and hard, which also meant that the task of classifying word difficulty was transformed into a prediction task for ATN, which allowed us to use the above MIMO XGBoost Model again, further demonstrating the reasonableness and superiority of the model.

Finally, in the course of our analysis we found that the number of Wordle players, in either mode, tends to rise and then fall in 2022, and in order to re-increase the flow of the game, we have summarized the following recommendations.

1. propose two words per day, one for difficult words and the other for easy or medium words, thus making the game a more friendly experience for both beginners and expert word-guessers.
2. hold regular special events on Wordle, such as holiday-related word-guessing games on special holidays.
3. add a social aspect to the current game mode, such as PK with friends, to make the game more interesting.

As we quote at the beginning of our paper, "The purpose of computation is insight, not numbers.", we offer more than just cold models and numbers, but through our analysis, we want to help Wordle re-engage players around the world and make word guessing more accessible to more people.

Thanks for taking the time out of your busy schedule to read my letter. If you would like to know more about the results of the analysis, we would be glad to provide you with further assistance.

[1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.

[2] Wang Yan. Application of time series analysis [M]. Beijing: China Renmin University Press 2005.

[1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.

[2] Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[J]. ACM, 2016.

[1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.

[2] 徐维超. 相关系数研究综述[J]. 广东工业大学学报,2012,29(3):12-17.