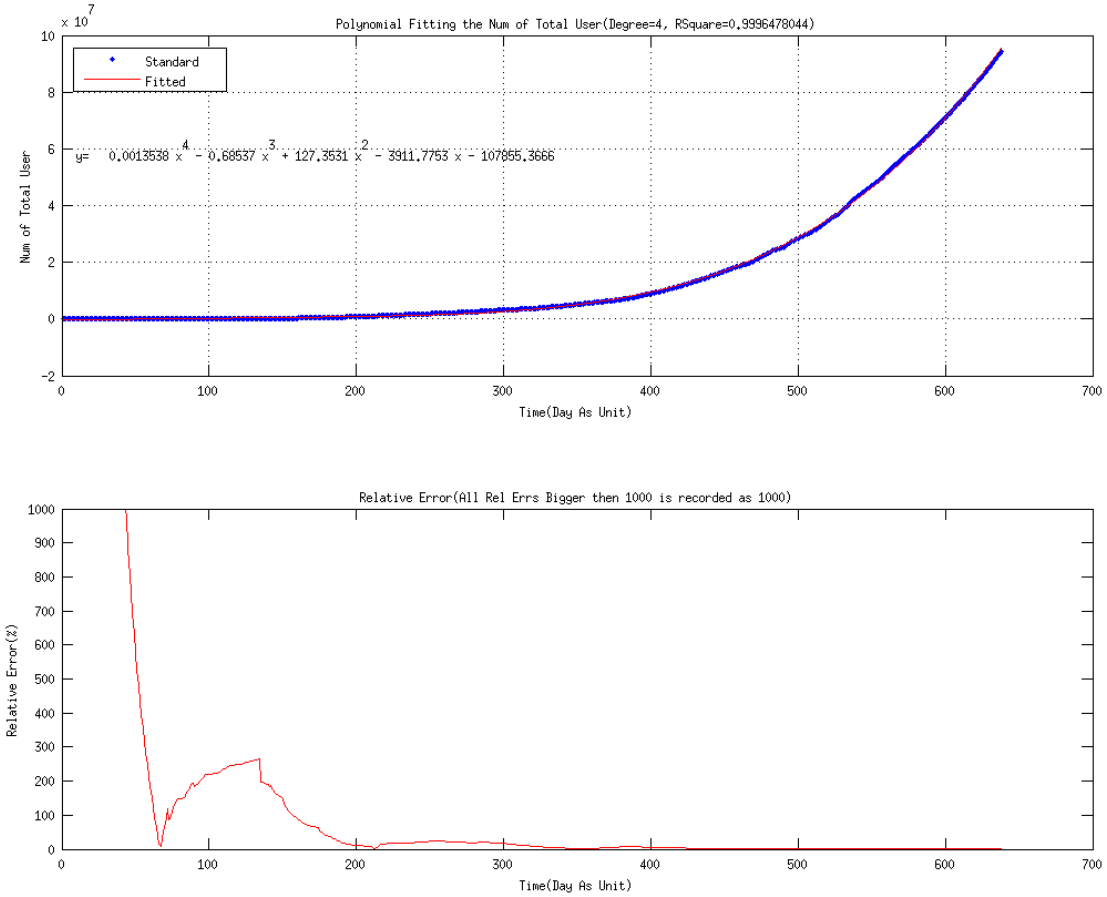
一、用户总数 数据模型

1. 用户总数的数据模型分析

观察用户总数随时间变化的曲线，曲线平滑、单调递增，适合用多项式进行拟合。图一是用4阶多项式进行拟合的结果：



图一 对用户总数随时间变化的曲线进行4阶多项式拟合(上图中的蓝点表示用户总数、红线表示拟合曲线；下图表示拟合值对真实值的相对误差)

观察图一，可以发现：

(1) 在初期(大约前200天内)，由于用户总数的基数小，使用多项式拟合的相对误差很大，但是在后期(大约在200天以后)，相对误差一直维持在一个很小的值。由于我们的主要目的是通过拟合已知的数据，来预测将来的数据走向，所以干脆舍去前200天内的数据，只对200天以后的数据进行拟合。

(2) 四阶多项式拟合曲线的确定系数(R Square)为0.999，说明多项式拟合得非常好，实际上，通过比较更高阶的多项式拟合结果，会发现随着阶数的升高，确定系数越来越高，多项式拟合得越来越好。但从拟合曲线能否正确预测将来的数据来看，高阶多项式的预测结果未必比低阶多项式好，因为高阶多项式十分容易发生抖动，后面会介绍一种判断拟合曲线预测能力好坏的方法。

2. 对用户总数的数据模型的分析和验证

根据前面对用户总数的数据模型的分析，建立如下的处理过程：

(1) 从文件中读取用户总数时间变化的数据，共 638 天的数据

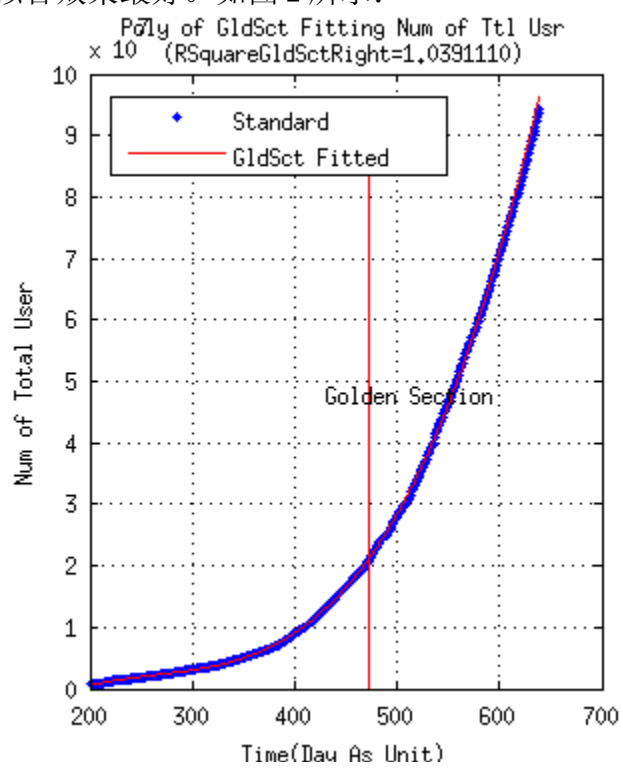
(2) 舍去前 200 天的数据，剩余第 200 天到第 638 天，共 439 天的数据。

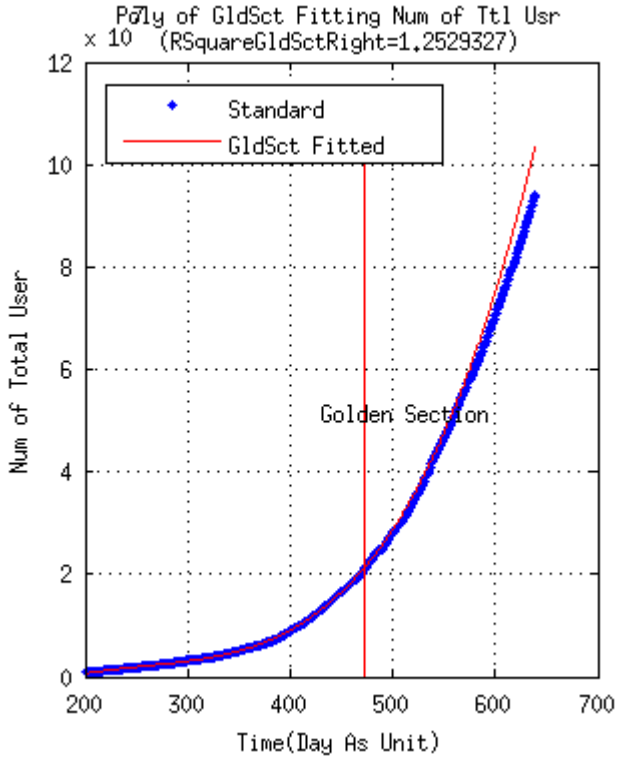
(3) 依次使用 2 阶到 6 阶的多项式对剩余的 439 个数据进行拟合，对拟合结果画图。

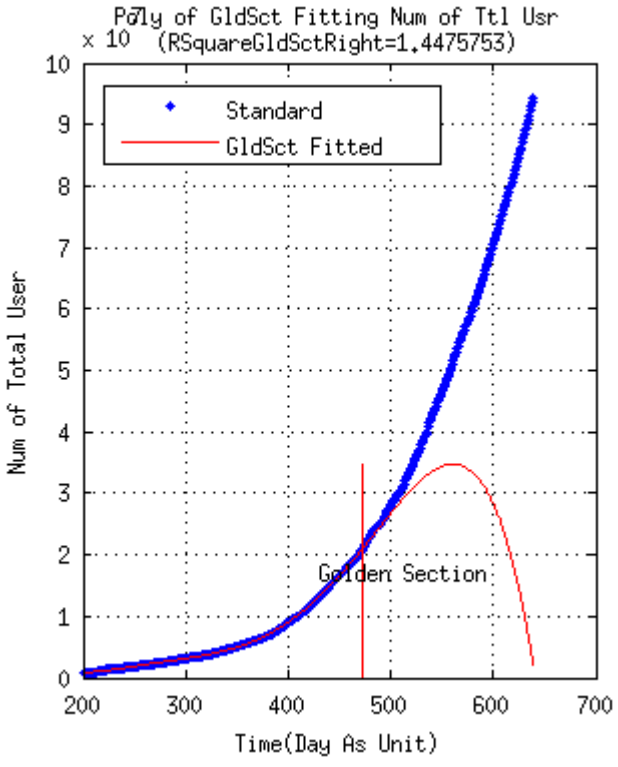
(4) 根据画图结果，分析能最优地预测未来的拟合曲线。

那么如何判断那一阶的多项式能最优地预测将来的数据呢?

可以通过黄金分割来分割已知的 439 天的数据，让前 271(=439×0.618) 天的数据参加同阶的多项式拟合，然后用拟合得到的函数与后 168(=439*0.382)天的数据进行比较,计算其确定系数(R Square)，最后用不同阶数的拟合函数算得的确定系数相比较，谁最接近 1,说明这一阶的拟合效果最好。如图 2 所示：







图二 左上方为 3 阶、右上方为 4 阶 、左下方为 5 阶多项式拟合的验证情况。可以看到，5 阶多项式在换进分割线左边虽然拟合得很好，但在黄金分割线右侧发生了很大的抖动，不能用来预测未来数据的变化情况；而 3 阶多项式拟合时，在黄金分割线右侧的拟合曲线的走势与实际数据基本相同，能很好得预测未来数据的走势。

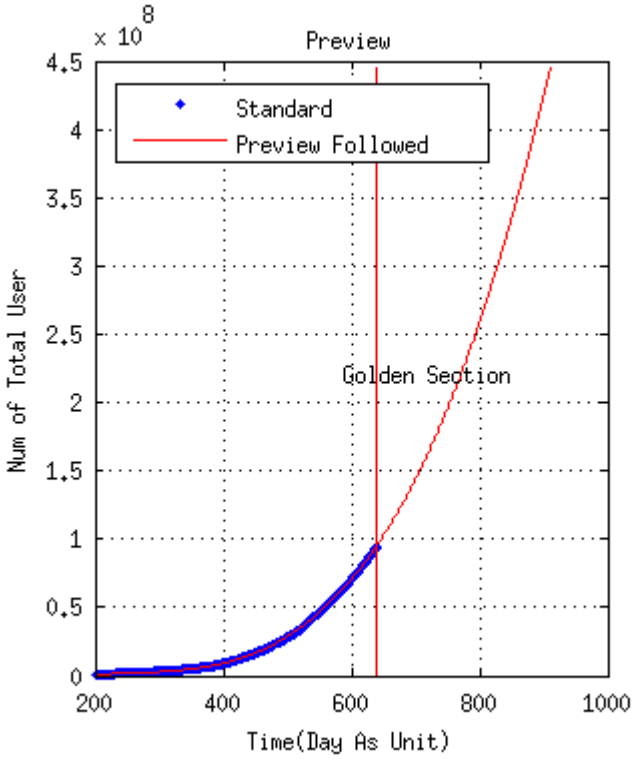
3. 用户总数的数据模型公式

根据前面的分析和验证，可以认为采用 3 阶多项式能很好得与已知的数据拟合、并用来推算未来的数据走势。三阶多项式的拟合函数为：

y = 1.5471*(x-200)^3 - 265.4165*(x-200)^2 + 32717.7053*(x-200) + 747569.7353

x 表示第几天(第一天的日期为 2010-10-20)，y 表示拟合的用户总数。

使用此函数对未来 271(439/0.618-439)天(第 639 到 909 天)进行预测,如图三所示：



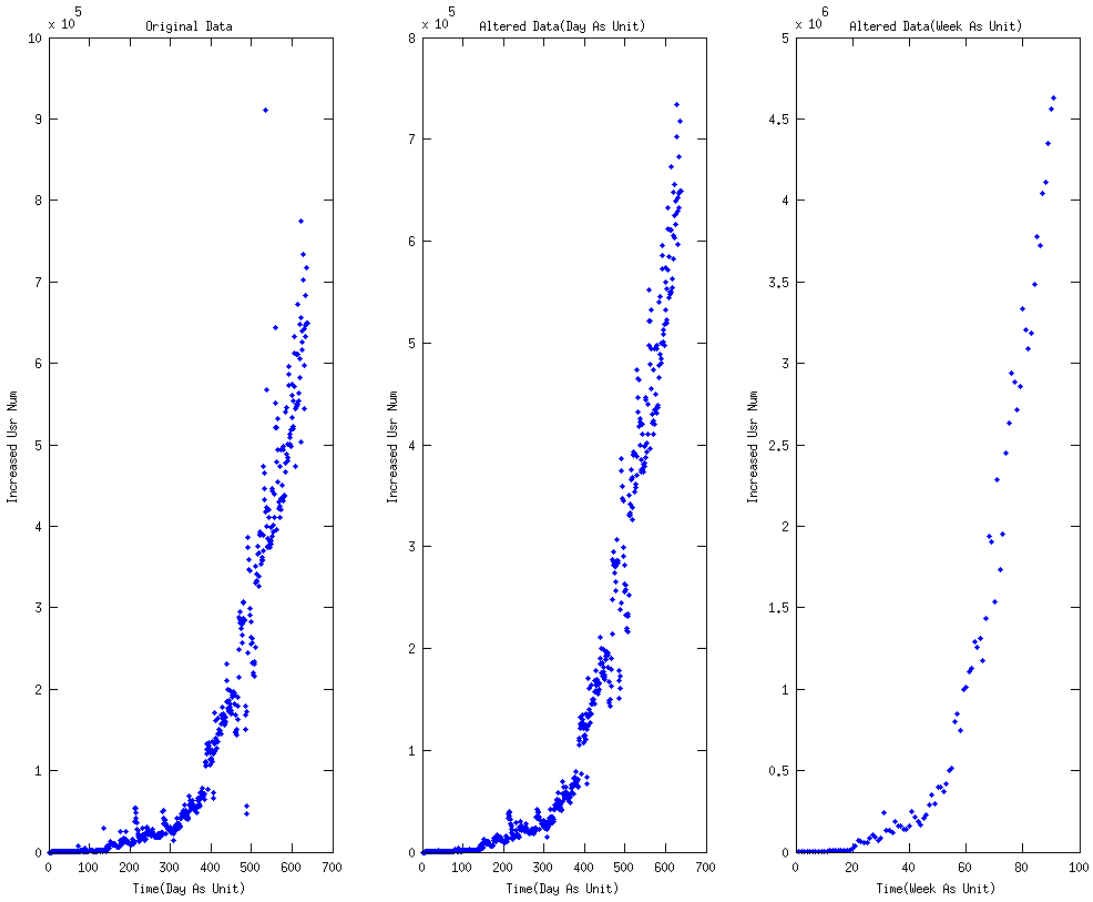
图三　使用拟合的三阶多项式函数对未来 271(439/0.618-439)天进行预测(黄金分割线右侧的红线表示将来 271 天内的预测结果)

二、新增用户数 数据模型

1. 新增用户数的数据模型分析

新增用户数随时间的变化不如用户总数的平滑，在个别点上出现抖动，但总体趋势还是平滑上升的，仍然可以用多项式的方法进行拟合。

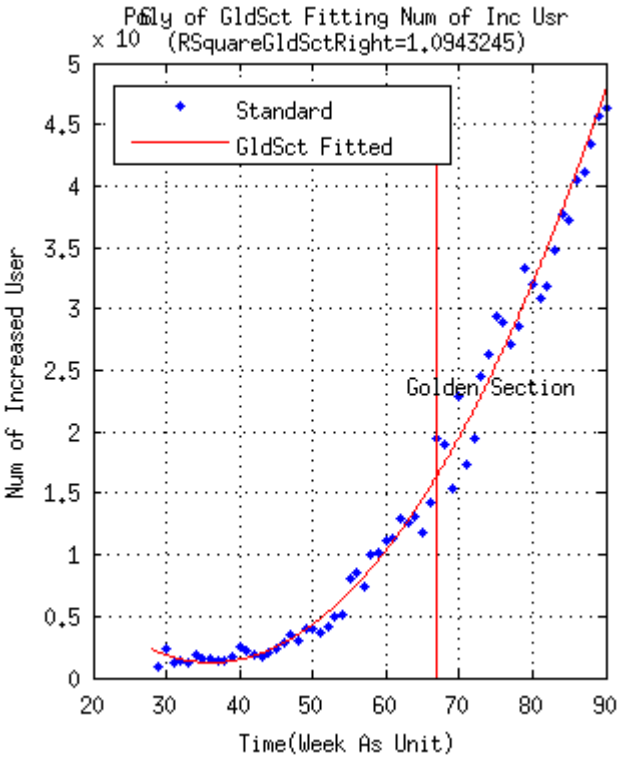
在拟合之前仍然需要对数据进行处理，首先对个别抖动的点按一定算法进行替换，使曲线更加平滑；然后使时间轴以星期做单位，即对每周新增用户数进行拟合，目的也是为了平滑曲线。如图四所示：

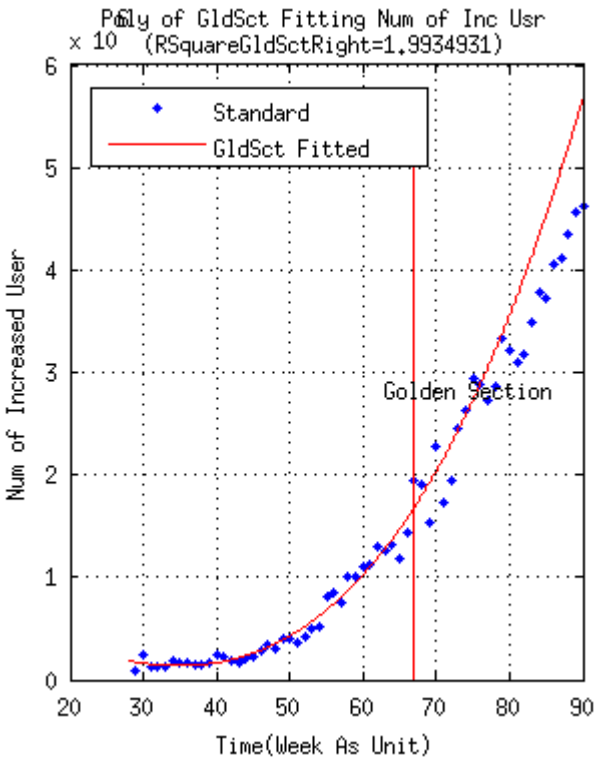


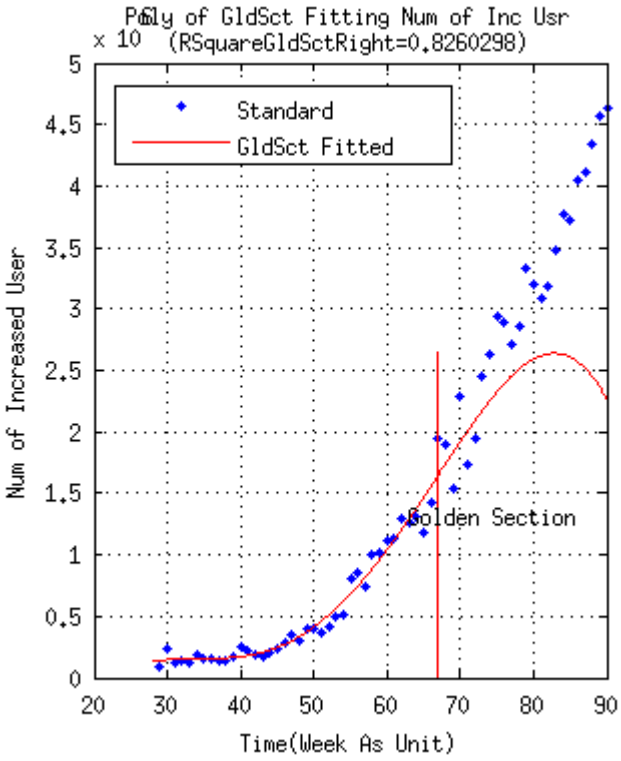
图四 对新增用户数进行预处理(第 1 副图为未处理的用户总数随时间变化的图像，第 2 副图为替换了个别抖动点后的图像，第 3 副图为使时间轴单位为星期)

2. 对每周新增用户总数的数据模型的分析和验证

对每周新增用户总数的数据模型的分析和验证方法与用户总数的方法类似，仍然采用多阶多项式拟合的方法进行拟合，并利用黄金分割的方法进行验证。黄金分割的验证结果，如图五所示：







图五 左上方为 2 阶、右上方为 3 阶、左下方为 4 阶多项式拟合的验证情况。可以看到，2 阶多项式拟合时，在黄金分割线右侧的拟合曲线的走势与实际数据基本相同，确定系数也最接近 1，能很好得预测未来数据的走势。

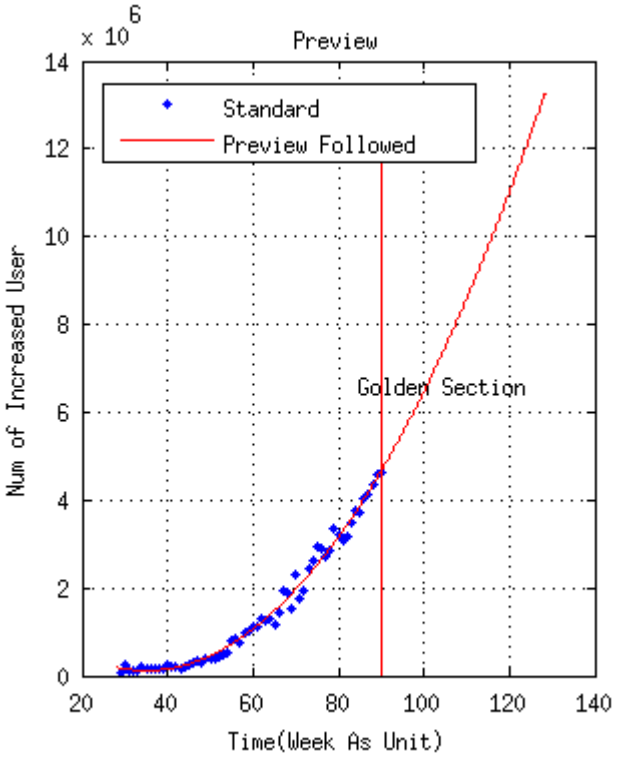
3. 新增用户数的数据模型公式

根据前面的分析和验证，可以认为采用 2 阶多项式能很好得与已知的数据拟合、并用来推算未来的数据走势。2 阶多项式的拟合函数为：

y = 1515.464*(x-28)^2 - 21845.6093*(x-28)+197565.6438

x 表示第几周，y 表示拟合的新增用户数。

使用此函数对未来 38(=271/7)周(第 92 周到第 129 周)进行预测,如图六所示：



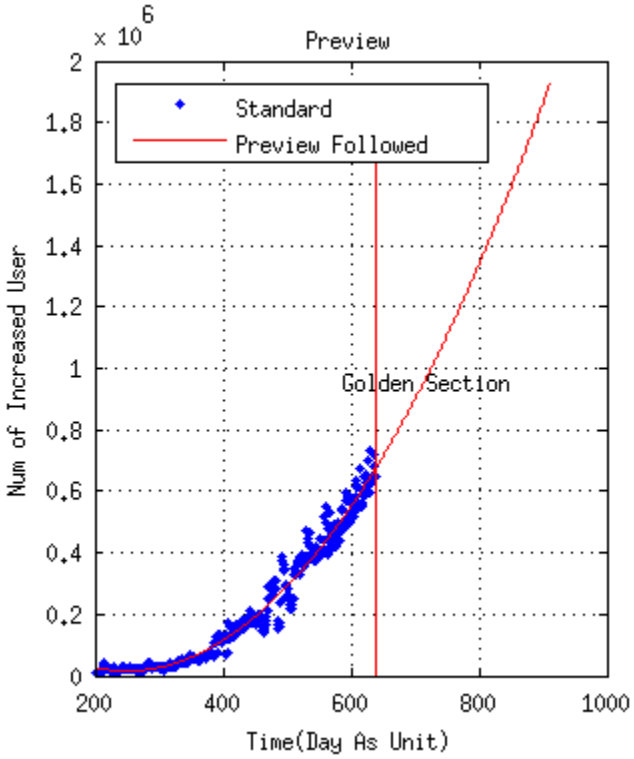
图六 对未来 38(=271/7)周(第 92 周到第 129 周)的每周新增用户数进行预测(黄金分割线右侧的红色曲线为预测曲线)

实际上，作为对比，我还将时间轴以天为单位建立了二阶多项式拟合函数：

$$y = 4.3832*(x-200)^2-445.554*(x-200)+27219.3335$$

x 表示第几天，y 表示新增的用户数。

使用此函数对未来 271(439/0.618-439)天(第 639 到 909 天)进行预测,如图七所示：

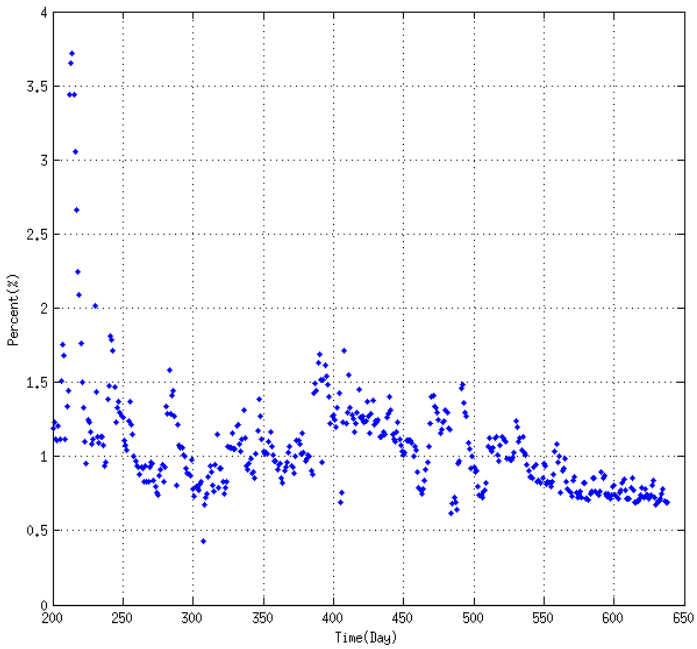


图七 未来 271(439/0.618-439)天(第 639 到 909 天)进行预测

三、新增用户百分比 数据模型

1. 新增用户百分比的数据分析

新增用户百分比随时间的变化如图八所示：

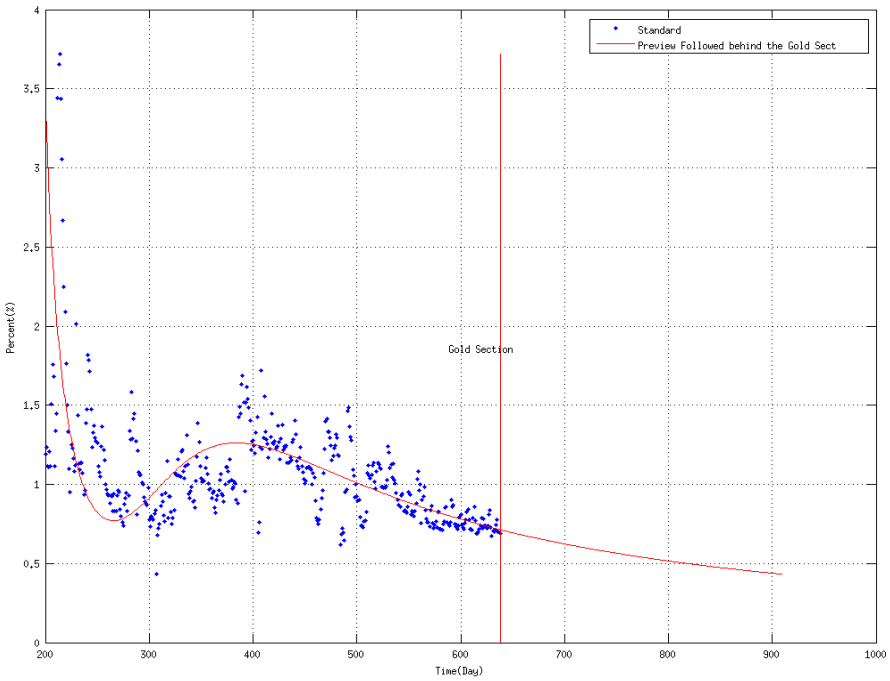


图八 新增用户百分比随时间的变化

可以看到，数据抖动特别大，且有两个明显的拐点，不适合做多项式拟合。

2. 新增用户百分比的数据模型

鉴于分段拟合的效果不太好，且难以用来预测未来新增用户百分比的变化，提出了一种新思路，即直接用前面对新增用户数随时间变化的拟合函数直接除以总用户的拟合函数，其拟合及预测结果如图九所示：



图八 对新增用户占总用户数的百分比进行拟合(以天为单位)