如何度量预测用户付费的误差

在广告，电商，游戏等行业中，预测用户付费是核心的业务场景，能直接帮助提升收入，利润等核心业务指标，堪称预测中的明星。在预测用户付费的系列文章中，结合作者理论和工程实践经验，深入探讨如何更好更准地去预测用户付费。

- ## MAE和RMSE

  传统的回归预测通常使用MAE，RMSE等指标去评价预测误差。

MAE全称Mean Absolute Error，指平均绝对值误差，是对预测值$\hat{y}_i$和真实值y_i的绝对差值计算平均值，其计算公式是：

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}$$

计算绝对值的好处在于能避免正负误差抵消的情况。例如，有三个误差值$(-1 \quad 0 \quad 1)$，不取绝对值的话，计算平均误差为0，与实际情况不符。而取绝对值后，计算平均误差为0.667。

RMSE全称Root Mean Square Error，指均方根误差，是计算所有预测值$\hat{y}_i$和真实值y_i的样本标准差，即对预测值$\hat{y}_i$和真实值y_i的差值取平方再计算平均值再开根号，其计算公式是：

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$

由于在误差计算中引入了平方计算，和MAE比较，RMSE会放大较大的误差。例如，有两组误差值，

$$(10 \quad 10 \quad 10)$$

$$(10 \quad 10 \quad 1000)$$

计算可得，

第一组MAE=10.0，RMSE=10.0，比例1:1

第二组MAE=340.0，RMSE=577.408，比例1:1.698

可见较大的误差值对RMSE影响更显著，换句话说，使用RMSE指标度量误差，会更多地去惩罚较大的误差，从而避免出现特别明显的预测误差。因为RMSE是光滑可微函数，所以很多回归模型都使用RMSE作为默认损失函数。

- ## MAPE和WAPE

  尽管MAE和RMSE是机器学习回归模型的默认度量指标，但不太适合付费预测的业务场景。例如，对于同样的真实值，

$$(10 \quad 10 \quad 100)$$

有以下两组不同的预测值，

$$(0 \quad 0 \quad 90)$$

$$(10 \quad 10 \quad 70)$$

计算得到以下两组不同的误差值，

$$(10 \quad 10 \quad 10)$$

$$(0 \quad 0 \quad 30)$$

第一组MAE=10.0，RMSE=10.0

第二组MAE=10.0，RMSE=17.32

只看MAE，两组误差一样大。只看RMSE第二组误差更大。但对于付费预测业务来说，第一组，

1. 虽然更准确地预测了付费值100的用户。
2. 但将两个付费值10元的用户都预测为0，会损失两个付费用户。

第二组，

1. 虽然预测付费值100的误差更大，但预测为70也能给予相当的信号。
2. 同时完全准确预测了两个付费值10的用户，在付费用户数上3:1领先于第一组。

在对业务的帮助上，第二组明显更好，理应认为第二组误差更小。

显然，MAE和RMSE不适合用来度量用户付费预测的误差。对于这样的情况，对误差引入百分比计算，将误差值计算转化为相对误差计算。把误差定义为预测误差占真实值的百分比，则真实付费值10，预测误差值1，和真实付费值100，预测误差值10，尽管数值上有10倍的差异，但在百分比上都是误差10%。上文的例子，按百分比误差计算得到以下两组新的误差值，

(100% 100% 10%)

(0% 0% 30%)

通常使用MAPE计算百分比误差。MAPE全称Mean Absolute Percentage Error，指平均绝对百分比误差，是预测值$\hat{y}_i$和真实值y_i的绝对差值，除以真实值y_i，得到绝对百分比误差，再求其平均值。其计算公式是：

$$MAPE = \frac{\sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{y_i}}{n}$$

按MAPE计算，

第一组MAPE=70%

第二组MAPE=10%

可见第二组比第一组误差更小。所以MAPE更适合付费金额预测这样的场景，即倾向于每一个真实值都能预测得比较准确，即使大额付费用户的误差值大一些，也不影响对整体预测准确程度的评估。但MAPE指标也存在一个问题，如果真实值y_i为0，则出现除数为0的情况，无法计算，不能度量预测值$\hat{y}_i$>0并且真实值y_i=0这种情况的误差。

对于MAPE做一个改进，用绝对误差总和去除以真实值总和，可以避免除数为0的问题。这样的指标叫WAPE，全称Weighted Absolute Percentage Error，指加权绝对百分比误差，是预测值$\hat{y}_i$和真实值y_i的绝对差值之和，再除以真实值y_i之和。其计算公式是：

$$WAPE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{\sum_{i=1}^{n} y_i}$$

按WAPE计算，

第一组WAPE=(10+10+10)/(10+10+100)=25%

第二组WAPE=(0+0+30)/(10+10+100)=25%

细心的读者已经发现，这两组的误差一模一样。因为WAPE统计的是总体误差，而无法区分具体误差的分布。

考虑到WAPE不能完全体现具体误差分布，在实际工程实践中，一般会综合评价MAPE和WAPE两个指标，先用WAPE看总体误差，再用MAPE看具体误差。如果业务对不同付费区间的误差敏感程度不一样，还要看相应付费区间的MAPE和WAPE，最简单的区间划分是十分位，即看十分位划分的MAPE和WAPE。

如果需要WAPE能对不同情况的误差进行区别，则需要对不同情况的误差进行加权处理，从而得到加权后的指标，叫WMAPE，全称Weighted Mean Absolute Percentage Error，指加权平均绝对百分比误差，是预测值$\hat{y}_i$和真实值y_i的绝对差值乘以加权系数w_t之和，再除以真实值y_i乘以加权系数w_t之和。其计算公式是：

$$WMAPE = \frac{\sum_{i=1}^{n} w_t|\hat{y}_i - y_i|}{\sum_{i=1}^{n} w_t y_i}$$

假设我们希望增加付费更小用户的权重，设真实值y_i=10，加权系数w_t=1.0，真实值y_i=100，加权系数w_t=0.8，则得到WMAPE的值为，

第一组WMAPE=
(10*1.0+10*1.0+10*0.8)/(10*1.0+10*1.0+100*0.8)=28%

第二组WMAPE=
(0*1.0+0*1.0+30*0.8)/(10*1.0+10*1.0+100*0.8)=24%

计算结果是第二组误差更小，能说明第二组对权重更高的用户付费预测更准确。

使用MAPE和WAPE代替MAE和RMSE度量预测用户付费误差，并指导模型进行优化之后，会得到如下效果：

1. 显著降低总体预测误差。

2. 真实值较小用户的预测误差降低最为明显。
3. 显著减少过预测（预测值大于真实值）的情况。

同时也会存在如下问题：

1. 真实值较大用户的预测误差可能不降反增。
2. 整体预测总值偏低，大部分预测值都是欠预测（预测值小于真实值）。
3. 欠预测会导致给模型下游系统的信号值偏低，影响业务效果。

- **单位信号量误差**

  为什么使用MAPE和WAPE会导致欠预测？因为MAPE和WAPE的计算中，真实值y_i是分母，相当于乘以$\frac{1}{y_i}$，y_i越大，乘数越小，y_i越小，乘数越大。相对于平均值回归曲线$y=\bar{y}$，MAPE在平均值回归曲线下方的点，y_i更小，乘数更大，有更大的权重，会把MAPE整体往平均值回归曲线的下方去拉，会有更多的点在平均值回归曲线下方，使得更多预测值低于平均值，从而导致欠预测。

为了解决欠预测的问题，引入信号量的概念，信号量等于预测平均值除以真实平均值。预测完全准确的情况下，信号量等于1。

$$semaphore=\frac{\bar{y_i}}{\bar{y}}$$

误差和信号量在实际分布上有局部或者全局最优解，在最优解上误差最小，在最优解附近，信号量变大，误差也会变大，信号量变小，误差也会变大。但如果将信号量按区间分组总体来看，信号量偏低的分组，误差会比信号量偏高的分组误差要低，这跟总体欠预测的情况是一致的。

对于实际业务场景，如果只优化误差，导致信号量变小，给下游系统的信号值不够，也会影响下游系统的业务，得不偿失，所以需要对误差和信号量综合度量。对于给定的信号量，误差越小越好。对于给定的误差，信号量越大越好。

引入新的指标，单位信号量误差，指误差和信号量的比值，是WAPE和MAPE的加权之和，再除以信号量。可以根据实际情况，对WAPE和MAPE取不同的权重，默认都取0.5。其计算公式是：

$$error=\frac{0.5\times WAPE+0.5\times MAPE}{semaphore}$$

使用单位信号量误差，能综合度量总体误差，付费用户误差和信号量三种情况，在实际业务中取得了最好的效果。

- ## 其他度量指标
  - ### 线性系数R-Square

    在预测用户付费系统的下游系统也是机器学习系统的场景里，预测值$\hat{y}_i$和真实值y_i的线性相关性非常重要。和真实值y_i存在显著线性关系的预测值$\hat{y}_i$，能让下游的机器学习系统学得更好，取得更好的业务效果。

一般用R-Square指标计算线性相关性，其计算公式是：

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- ## 多分类指标

  如果业务允许预测金额存在一定误差，更关注是否能将用户付费预测到所在付费区间，比如，预测用户是高价值用户，一般价值用户，低价值用户。这时，不妨将预测用户付费金额的回归问题，转化为预测用户付费金额所在区间的多分类问题，使用多分类评价指标来度量预测用户付费的误差。