宜信 CreditEase 10周年 2006-2016

科 技 让 金 融 更 美 好

算法大赛技术分享

郑大念、蔡建明、朱治柳、王三鹏

2017-04-12

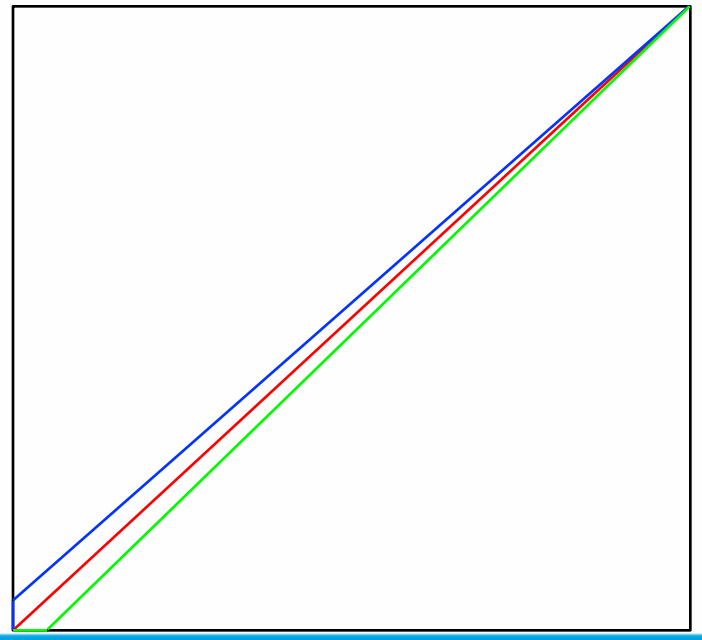
Task-1，标签1的占比

数据集

训练集：20085，标签1的占比＝0.3136

测试集：9513，标签1的占比＝0.7182

小技巧——通过一次提交推导测试集上1的占比(赛后分析)



如果预测全0，AUC＝0.5，ROC曲线如图中红线。

如果预测全0，任意修改一个样本为1

- 当1正确时，ROC曲线如图中蓝线
  AUC＝ 1-0.5*(N_+-1)/N_+ = 0.5+0.5/N_+
- 当1错误时，ROC曲线如图中绿线
  AUC = 0.5*(N_--1)/N_- = 0.5-0.5/N_-

实验：

- 当1正确时，0.500073=0.5+0.5/6832
- 当1错误时，0.499813=0.5-0.5/2681

Task-1，数据集分布和特征处理

数据集分布

- 重采样
- 删除样本
- 添加样本

  例子，20085(train)+20098(废弃test) = 40183

  样本越少，分布越不稳定；样本越多，分布越稳定可靠。

特征处理

- **特征选择**

  无用特征，值全0的、值0占比非常高的

  分布差异大，观察训练集和测试集的特征取值分布图有非常大差异，如手机型号

  不稳定特征，train上有效果，test上无效果，例子app_cat2_43、app_cat2_191、...

Task-1，特征处理（续）

特征处理（续）

- **特征对齐**

  划分bins

  长尾截断

  ...

- **特征构造**

  构造新的特征

  例子，构造ip_app_same: 若有ip相同并且app特征完全相同，并且只要有经纬度，经纬度要相同时， 则定义为1；否则定义为0。一般为刷机用户，不会购买。

  有ip_app_same时，单分类器 AUC = 0.775533

  无ip_app_same时，单分类器 AUC = 0.756612

Task-1，多分类器和集成

决策树类的分类器

- CART
- RF、AdaBoost
- GBDT、XGBoost、LightGBM、LBT (LossBoostedTrees)

比赛时提交的最后一版

GBM	RF	LBT	Mean Ranks
0.775014	0.770385	0.770588	0.779161

分类器之间需要有互补性，测试集上的概率或排序的相关系数度量。

多分类器集成

- 排序平均，Mean Ranks

$$score = (rank_1+rank_2+...+rank_m)/(m*test\_num)$$

Task-1，赛后分析——LBT

宜信 CreditEase 10周年

不同配置下的LBT，在测试集上的AUC值

param	split_node	loss=C_LOGISTIC	loss=C_SMOOTH_HINGE	loss=C_SQUARE_HINGE
max_depth=2 min_leaf_samples=100 learning_rate=0.1 num_trees=50	<, >= ==, !=	0.772429	0.771243	/
	<, >=	0.774065	0.774565	0.776443
max_depth=2 min_leaf_samples=100 learning_rate=0.05 num_trees=100	<, >=	0.774995	0.772435	0.776858

- app_cat类特征，好样本占比曲线一般是单调的，非枚举型，不适合用==/!=分裂。
- 有待完善，因开发不久，如loss=C_SQUARE_HINGE是赛后补开发的，还缺少一些重要功能，如feature bins、feature fraction等。

Task-1，赛后分析——分类器集成

以下数据分析都是基于最后提交的一版，非三分类器的最佳状态。

1、排序平均——Mean Ranks

0.779161, baseline

2、概率平均——Mean Probs

0.777888, bad

3、Stacking Ranks by L2-LR+

0.777901, bad

wb = [0.4007 0.3683 0.4244 -1.3398]

4、Stacking Probs by L2-LR+

0.777858, bad

wb = [0.0751 0.0489 0.0431 -0.7810]

当使用Stacking时，需要有正则项，并适当调节。

Task-1，赛后分析——分类器集成（续）

宜信 CreditEase 10周年

5、高级特征+原始特征，Stacking by L2-LR+

高级特征：GBM / RF / LBT，Probs

原始特征：212维，每维特征独自构建弱分类器 CART with Gini，depth=1

0.780167

原始特征：从212维选最佳9维，CART with Gini，depth=1

0.781596

原始特征：从212维选最佳9维，CART with Gini，depth=2

0.783993

原始特征：从212维选最佳9维，CART with Gini，depth=3

0.784611

原始特征：从212维选最佳9维，CART with Gini，depth=4

0.784880

注：CART预测输出，Gini impurity = sum $p_i*(1-p_i)$

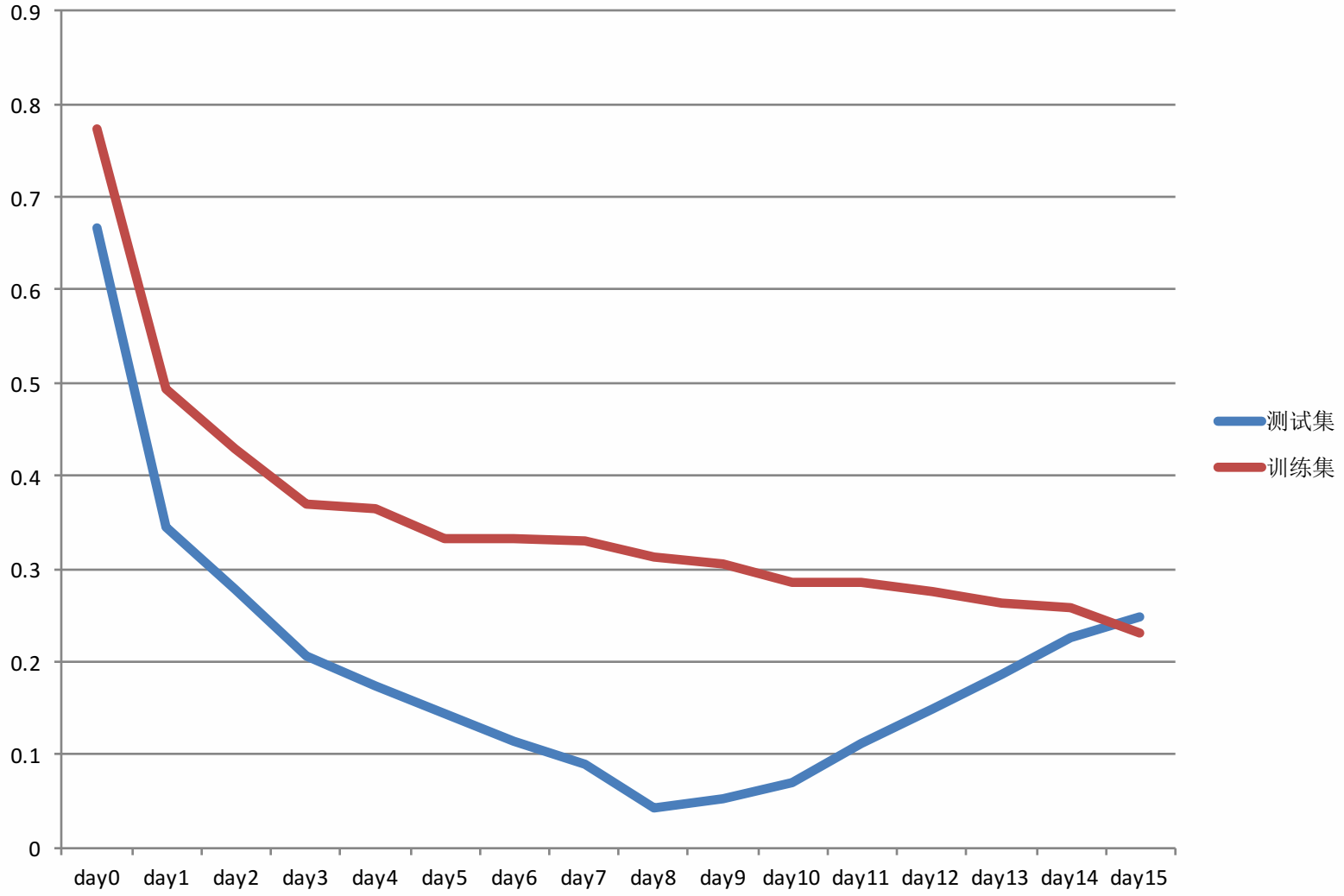
if leaf_node=1, output 1-impurity; otherwise impurity.

Task-2，数据分析

1、零的占比
training set: 0.7672 → 猜测test set的零占比
test set: 0.7437
制定分类-回归策略

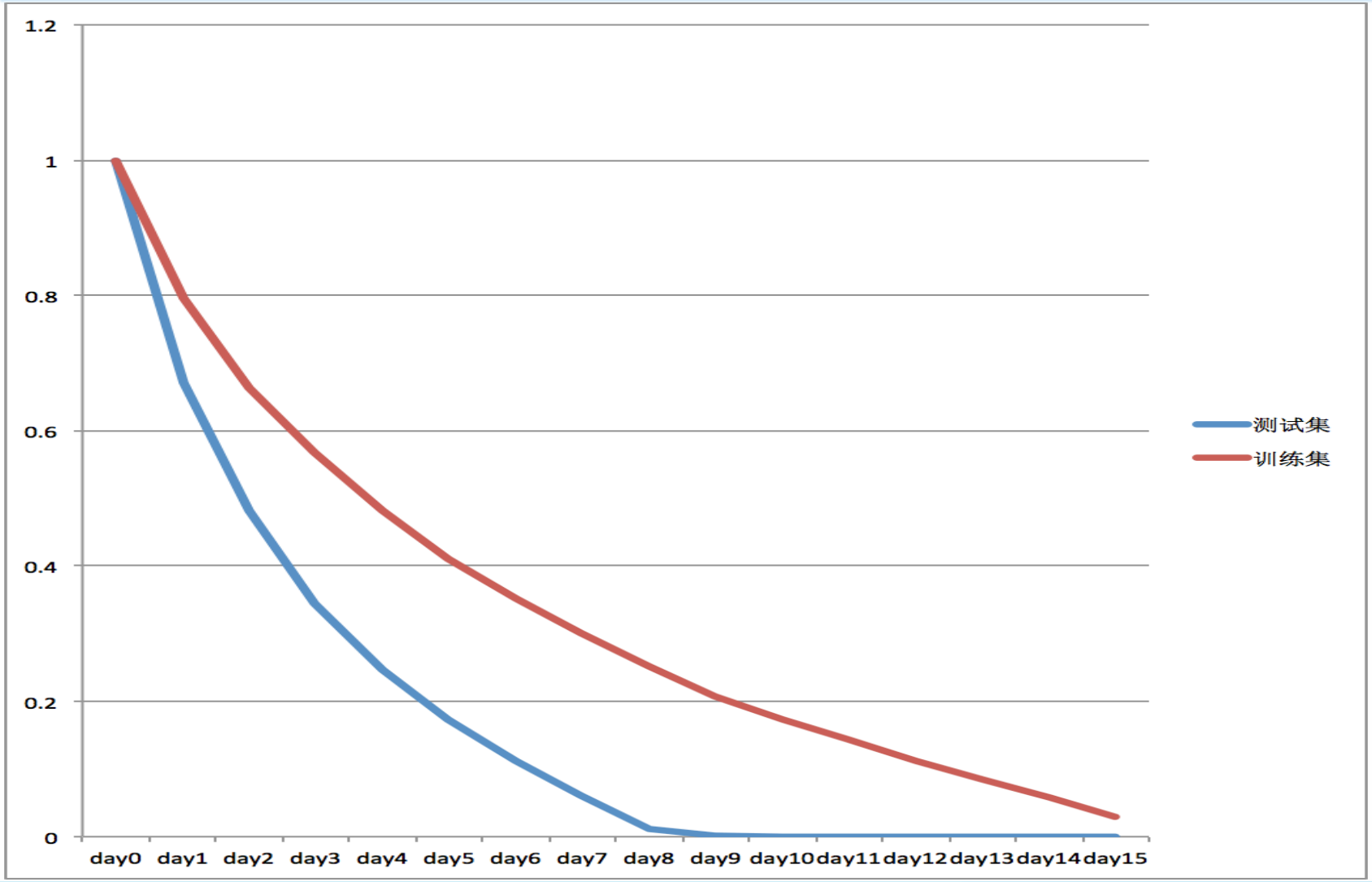
2、重要特征及其分布
GBDT feature importance 筛选特征 → app_day

Task-2，数据分析——时间序列分布不一致



训练集和测试集app_day{n}的非零个数占比

Task-2，数据分析——特征对齐



排序后，训练集和测试集app_day{n}的非零个数占比

Task-2，特征工程

1、特征选择
删除冗余特征amount, transaction_count……

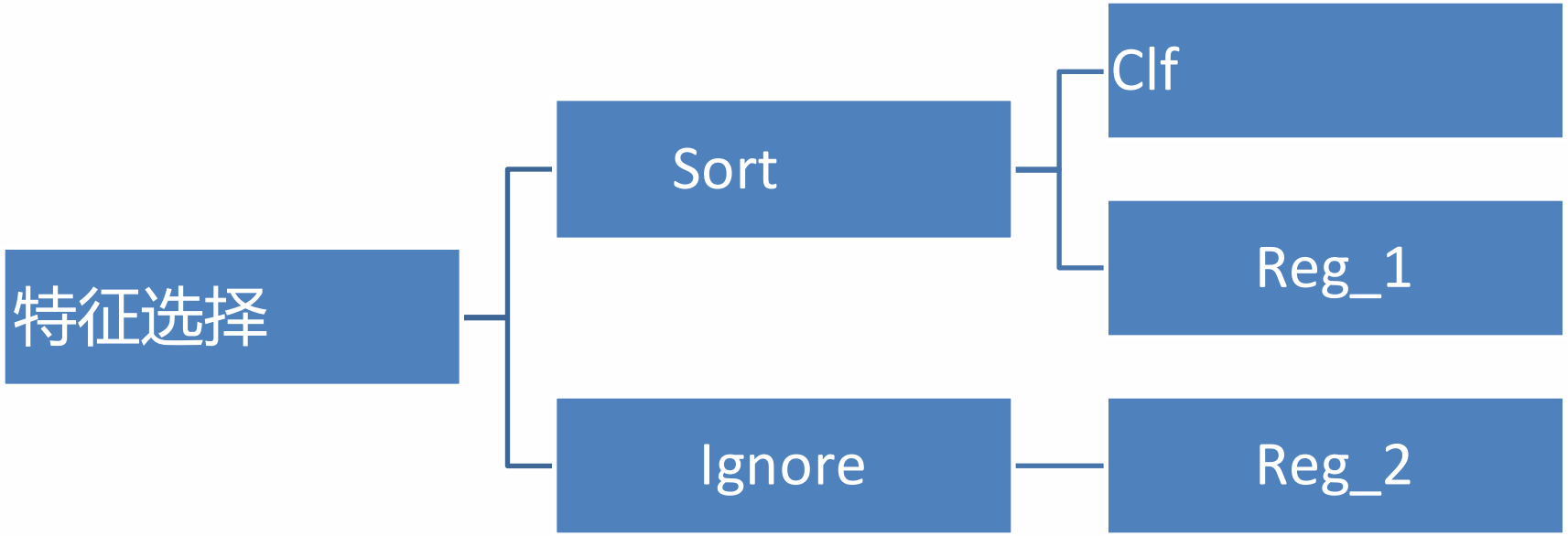
2、特征预处理
a. 每行数据app_day{n}用排序后的替代 → reg_1
b. 删除分布不一致的app_day{7-10} → reg_2

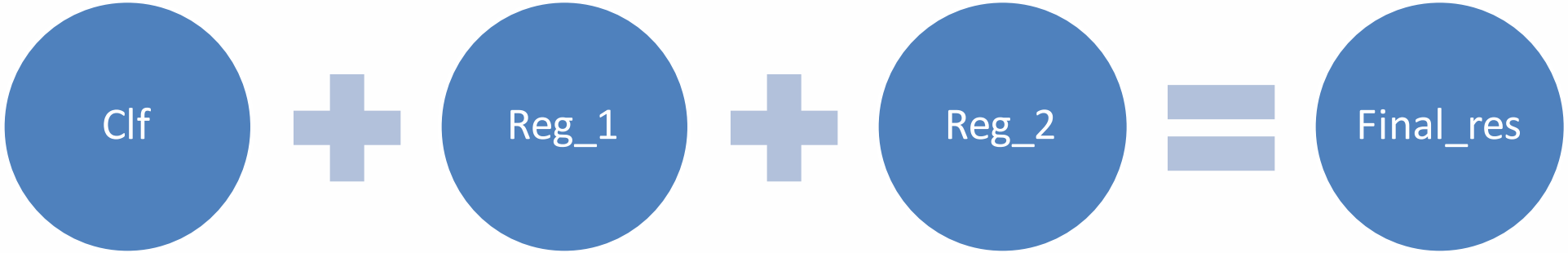
Task-2，特征工程——特征离散化

```
for (int i = 0; i < num_values - 1; ++i) {
  if (!is_big_count_value[i]) {
    rest_sample_cnt -= counts[i];
  }
  cur_cnt_inbin += counts[i];
  // need a new bin
  if (is_big_count_value[i] || cur_cnt_inbin >= mean_bin_size ||
    (is_big_count_value[i + 1] && cur_cnt_inbin >= std::max(1.0, mean_bin_size * 0.5f))) {
    upper_bounds[bin_cnt] = distinct_values[i];
    if (bin_cnt == 0) {
      cnt_in_bin0 = cur_cnt_inbin;
    }
    ++bin_cnt;
    lower_bounds[bin_cnt] = distinct_values[i + 1];
    if (bin_cnt >= max_bin - 1) { break; }
    cur_cnt_inbin = 0;
    if (!is_big_count_value[i]) {
      --rest_bin_cnt;
      mean_bin_size = rest_sample_cnt / static_cast<double>(rest_bin_cnt);
    }
  }
}
```

3、特征离散化

Task-2，模型流程





Task-2，模型参数

```
learning_rate = 0.02
num_iterations = 700
is_unbalance = true
lambda_l1 = 1
lambda_l2 = 1
```

总体思想是提高模型的泛化能力

Task2，复盘

1、特征选择
2、Sort
3、Rank Scale
　　按照排序序数scale至[0,1]。
4、离散化

Single model: Reg → 1.20177
Clf + Reg → 1.1968

宜信 CreditEase 10周年 2006-2016

科 技 让 金 融 更 美 好

Thank you