

推荐模型的融合

赵海臣

als的结果分段数据

```
> *****推荐[_alsForTest]*****
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.263635][0.005302][5113319.486571][0.799960][0.982658][1.000000][13.903810][22.347600][0.000060],[0.000890] *****
> *****mainpageTOP召回*****
> ***** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
> ***** [0.084694][0.001640][1642737.132984][0.006001][0.012002][25.000000][38.000000] *****
> *****allTOP召回*****
> ***** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
> ***** [0.204731][0.003984][3503187.016149][0.006001][0.012002][21.000000][24.000000] *****
> *****sevenDaysTOP召回*****
> ***** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
> ***** [0.230771][0.004397][4031258.416872][0.006001][0.012002][19.000000][25.000000] *****
> *****推荐[_alsForTest]*****
> lowRank = 0, highRank = 19
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.142660][0.002735][2731375.603641][0.891661][1.000000][1.000000][7.057920][7.057920][0.061270],[0.213260] *****
> *****推荐[_alsForTest]*****
> lowRank = 20, highRank = 39
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.048997][0.001023][976280.101107][0.972332][1.000000][1.000000][9.097980][9.097980][0.011990],[0.067320] *****
> *****推荐[_alsForTest]*****
> lowRank = 40, highRank = 59
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.030629][0.000651][606777.331005][0.988560][1.000000][1.000000][10.054870][10.054870][0.003600],[0.029050] *****
> *****推荐[_alsForTest]*****
> lowRank = 60, highRank = 79
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [1.000000][1.000000][10.677040][10.677040][0.001300],[0.014880] *****
> *****
> lowRank = 80, highRank = 99
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.017704][0.000386][11161090][11.161090][1.000000][1.000000][11.161090][11.161090][0.000720],[0.008080] *****
```

cf的结果分段数据

```
> *****推荐[_CF_test]*****
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.410682][0.008173][7398654.254959][0.407381][0.791179][1.000000][13.830320][21.700710][0.003680],[0.017200] *****
> *****mainpageTOP召回*****
> ***** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
> ***** [0.084227][0.001634][1618157.142221][0.006001][0.012002][25.000000][38.000000] *****
> *****allTOP召回*****
> ***** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
> ***** [0.205926][0.003956][3387730.564634][0.006001][0.012002][21.000000][24.000000] *****
> *****sevenDaysTOP召回*****
> ***** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
> ***** [0.232887][0.004383][3898925.295349][0.006001][0.012002][19.000000][25.000000] *****
> *****推荐[_CF_test] *****
> lowRank = 0, highRank = 19
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.319774][0.006098][5710689.353359][0.752470][1.000000][1.000000][7.390960][7.390960][0.077730],[0.233180] *****
> *****推荐[_CF_test] *****
> lowRank = 20, highRank = 39
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.040336][0.000924][751235.990700][0.756600][1.000000][1.000000][8.178901][8.178901][0.059834],[0.195712] *****
> *****推荐[_CF_test] *****
> lowRank = 40, highRank = 59
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.021764][0.000495][401719.640393][0.759026][1.000000][1.000000][8.598032][8.598032][0.043233],[0.157681] *****
> *****推荐[_CF_test] *****
> lowRank = 60, highRank = 79
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [1.000000][1.000000][9.032133][9.032133][0.031693],[0.124332] *****
> *****
> lowRank = 80, highRank = 99
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.012507][0.000283],[0.000000][0.000000][9.451328][9.451328][0.023884],[0.098825] *****
```

分段结果分析

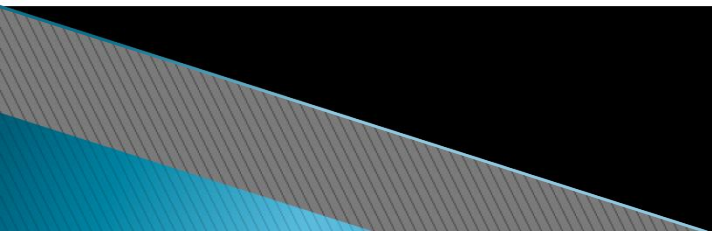
- ▶ 根据两种推荐模型分段结果，考虑使用Blending交叉融合的方式进行组合：
 - 两种不同来源的推荐结果，从分段数据上皆体现出排名越高，召回率/精确度/销售额都远高于排名低的，因此可以想办法融合两种模型，让二者分段融合，使得两个模型高分段保持高分，中分段保持中分，低分段保持低分。

融合方法

- ▶ 由于cf与als的原始分数分布大部分在1以下，cf偶尔能超过1，因此融合方法主要是针对不同段的分数加上一个相应的大值权重，使得不同的层次之间达到不可逾越的鸿沟。
 - Class A：二者相同user-productId部分基础分以个性模型als的recommendationScore为基准，加上60的偏置权重；
 - Class B：二者剩下的分别处于最高1~20位的，加上50的偏置权重；
 - Class C：二者剩下的分别处于最高21~40位的，加上40的偏置权重
 -
 - Class F：二者剩下的分别处于最高80~100位的，加上10的偏置权重

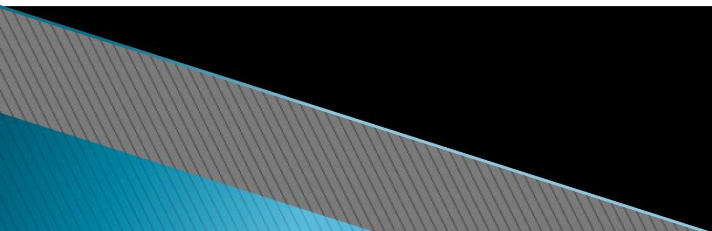
融合方法

- ▶ 加上分层偏置权重后，层与层之间很难跨越，因此可以达到层与层之间分别融合的目的
- ▶ 最后将各个CLASS A/B/C/D/E/F的DataFrame进行unionAll，达到拼接成整体推荐排序的目的，推荐分按照排序给即可。



分数重映射

- ▶ 由于处理过的分数有明显的分层性质，需要将分数进行合理化，可以考虑按照排序映射成长尾分布分数。
 - 对每个用户的整体分数进行排序，使用zipf函数进行分数重分布，重新生成成长尾分布。



融合后整体评估结果

```
> *****推荐[_Combine_test]*****
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.425164][0.008322][7669285.037242][0.283748][0.728360][1.000000][13.994330][22.738280][0.000490],[0.002870] *****
> *****mainpageTOP召回*****
> ***** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
> ***** [0.084440][0.001612][1625083.342823][0.006001][0.012002][25.000000][38.000000] *****
> *****allTOP召回*****
> ***** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
> ***** [0.210416][0.003974][3454065.455631][0.006001][0.012002][21.000000][24.000000] *****
> *****sevenDaysTOP召回*****
> ***** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
> ***** [0.235369][0.004375][3947741.876448][0.006001][0.012002][19.000000][25.000000] *****
> *****推荐[_Combine_test]*****
> lowRank = 0, highRank = 19
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.196824][0.003599][3539871.283162][0.648627][1.000000][1.000000][7.011310][7.011310][0.079310],[0.244210] *****
> *****推荐[_Combine_test]*****
> lowRank = 20, highRank = 39
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.081972][0.001648][1453752.131136][0.720601][1.000000][1.000000][8.539005][8.539005][0.026880],[0.113831] *****
> *****推荐[_Combine_test]*****
> lowRank = 40, highRank = 59
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.076463][0.001577][1411597.081268][0.716526][1.000000][1.000000][9.286083][9.286083][0.013390],[0.070611] *****
> *****推荐[_Combine_test]*****
> lowRank = 60, highRank = 79
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [1.000000][1.000000][9.939749][9.939749][0.007230],[0.048621] *****
> *****
> lowRank = 80, highRank = 99
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.020403][0.000437][1030435.000000][0.000000][0.000000][1.000000][1.000000][10.327537][10.327537][0.004900],[0.034851] *****
```


基于CF评估用户的评估结果

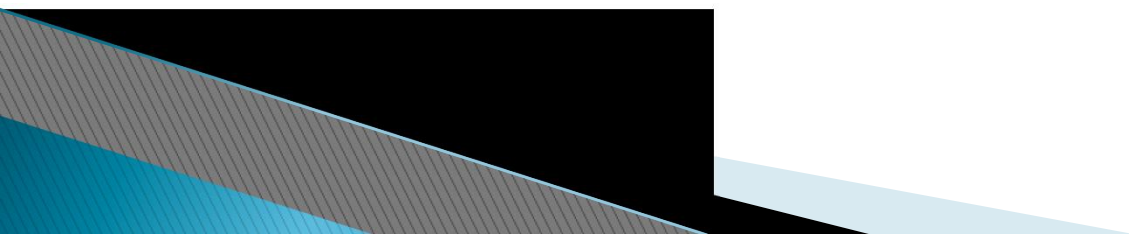
```
> *****推荐[_Combine_test]*****
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.458966][0.009114][8273833.726410][0.299980][0.739817][1.000000][14.502850][23.633090][0.000430],[0.003040] *****
> *****mainpageTOP召回*****
> ***** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
> ***** [0.084227][0.001634][1618157.142221][0.006001][0.012002][25.000000][38.000000] *****
> *****allTOP召回*****
> ***** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
> ***** [0.205926][0.003956][3387730.564634][0.006001][0.012002][21.000000][24.000000] *****
> *****sevenDaysTOP召回*****
> ***** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
> ***** [0.232887][0.004383][3898925.295349][0.006001][0.012002][19.000000][25.000000] *****
> *****推荐[_Combine_test]*****
> lowRank = 0, highRank = 19
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.206166][0.003819][3723617.512632][0.657462][1.000000][1.000000][7.236210][7.236210][0.074170],[0.229040] *****
> *****推荐[_Combine_test]*****
> lowRank = 20, highRank = 39
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.089329][0.001824][1594521.261141][0.730395][1.000000][1.000000][8.645436][8.645436][0.026420],[0.112631] *****
> *****推荐[_Combine_test]*****
> lowRank = 40, highRank = 59
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [0.088204][0.001818][1610859.851355][0.725041][1.000000][1.000000][9.323146][9.323146][0.014290],[0.074301] *****
> *****推荐[_Combine_test]*****
> lowRank = 60, highRank = 79
> ***** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
> ***** [1.000000][1.000000][9.974719][9.974719][0.008410],[0.052862] *****
```

基于ALS评估用户的评估结果

```
*****推荐[_Combine_test]*****
**** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
****  [0.417887][0.008403][7763040.417944][0.265653][0.725157][1.000000][14.013650][22.719520][0.000050],[0.001190] *****
*****mainpageTOP召回*****
**** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
****  [0.084694][0.001640][1642737.132984][0.006001][0.012002][25.000000][38.000000] *****
*****allTOP召回*****
**** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
****  [0.204731][0.003984][3503187.016149][0.006001][0.012002][21.000000][24.000000] *****
*****sevenDaysTOP召回*****
**** 召回,准确率,销售额,top50覆盖率,top100覆盖率,top50多样性,top100多样性*****
****  [0.230771][0.004397][4031258.416872][0.006001][0.012002][19.000000][25.000000] *****
*****推荐[_Combine_test]*****
lowRank = 0, highRank = 19
**** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
****  [0.180248][0.003436][3432953.053371][0.631128][1.000000][1.000000][7.033860][7.033860][0.074390],[0.237210] *****
*****推荐[_Combine_test]*****
lowRank = 20, highRank = 39
**** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
****  [0.085646][0.001729][1542845.451438][0.717184][1.000000][1.000000][8.564650][8.564650][0.023000],[0.104760] *****
*****推荐[_Combine_test]*****
lowRank = 40, highRank = 59
**** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
****  [0.079989][0.001670][1470267.611741][0.715785][1.000000][1.000000][9.301630][9.301630][0.010960],[0.063790] *****
*****推荐[_Combine_test]*****
lowRank = 60, highRank = 79
**** 召回,准确率,销售额,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比*****
****  [1.000000][1.000000][9.956960][9.956960][0.006280],[0.043310] *****
```

分析

- ▶ 融合后的总体召回率/精确度/销售额都比单个模型要好。
- ▶ 并且分段结果相比于之前聚集在top20分布得相对均匀。



The end