#### 历史版本1：

目标：1.播放时长大于5s 2.是否播放

样本：

总量：280万

正样本：

目标1：100万

目标2：180万

负样本：

目标1: 180万

目标2：90万

模型：gbdt

可用特征：

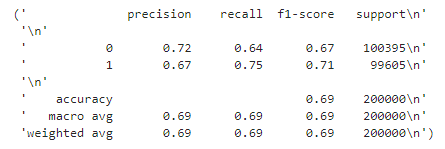
**去掉空置率大于50%的字段和中文字符字段，27个特征**

|  |  |
| --- | --- |
| 点赞数 | feed\_like\_count |
| 转发数 | feed\_share\_count |
| 本月视频播放人数 | cur\_play\_video\_cnt |
| 本月视频播放人数占比 | cur\_play\_video\_ratio |
| 本月平均播放时长 | cur\_avg\_watch\_time |
| 本月视频5s播放占比 | cur\_5s\_video\_ratio |
| 本月视频转化率（营销类视频） | cur\_handle\_rate |
| 上月视频播放人数 | pre\_play\_video\_cnt |
| 上月视频播放人数占比 | pre\_play\_video\_ratio |
| 上月平均播放时长 | pre\_avg\_watch\_time |
| 上月视频5s播放占比 | pre\_5s\_video\_ratio |
| 是否分省视频 | is\_prov\_video |
| 是否15秒以内 | is\_15s\_video |
| 是否60秒以内 | is\_60s\_video |
| 是否30分钟以内 | is\_30m\_video |
| 是否近3天上架 | is\_3d\_update |
| 是否近7天上架 | is\_7d\_update |
| 是否营销类视频 | is\_market\_video |
| 省份编码 | prov\_code |
| 近7日平均播放时长 | watch\_avg\_time\_7day |
| 近7日5s有效观看占比 | watch\_finish\_ratio\_7day |
| 近14日平均播放时长 | watch\_avg\_time\_14day |
| 近14日5s有效观看占比 | watch\_finish\_ratio\_14day |
| 近30日平均播放时长 | watch\_avg\_time\_30day |
| 近30日5s有效观看占比 | watch\_finish\_ratio\_30day |

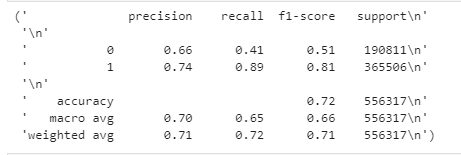
数据处理：原始数据

评估报告：

1. 5s播放时长



1. 是否播放



分析原因：特征过少，模型无法捕捉关系

#### 历史版本2：

目标：1.播放时长大于5s 2.是否播放

样本：

总量：280万

正样本：

目标1：100万

目标2：180万

负样本：

目标1: 180万

目标2：90万

模型：sharebottom、mmoe和deepFM

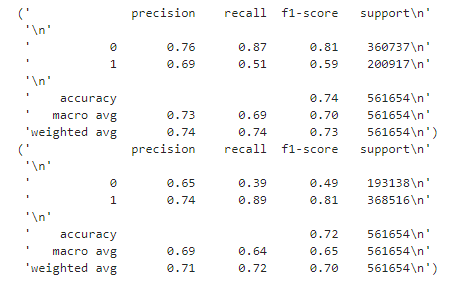
可用特征：

**去掉空置率大于50%的字段，29个特征**

|  |  |
| --- | --- |
| 点赞数 | feed\_like\_count |
| 转发数 | feed\_share\_count |
| 本月视频播放人数 | cur\_play\_video\_cnt |
| 本月视频播放人数占比 | cur\_play\_video\_ratio |
| 本月平均播放时长 | cur\_avg\_watch\_time |
| 本月视频5s播放占比 | cur\_5s\_video\_ratio |
| 本月视频转化率（营销类视频） | cur\_handle\_rate |
| 上月视频播放人数 | pre\_play\_video\_cnt |
| 上月视频播放人数占比 | pre\_play\_video\_ratio |
| 上月平均播放时长 | pre\_avg\_watch\_time |
| 上月视频5s播放占比 | pre\_5s\_video\_ratio |
| 是否分省视频 | is\_prov\_video |
| 是否15秒以内 | is\_15s\_video |
| 是否60秒以内 | is\_60s\_video |
| 是否30分钟以内 | is\_30m\_video |
| 是否近3天上架 | is\_3d\_update |
| 是否近7天上架 | is\_7d\_update |
| 是否营销类视频 | is\_market\_video |
| 省份编码 | prov\_code |
| 近7日平均播放时长 | watch\_avg\_time\_7day |
| 近7日5s有效观看占比 | watch\_finish\_ratio\_7day |
| 近14日平均播放时长 | watch\_avg\_time\_14day |
| 近14日5s有效观看占比 | watch\_finish\_ratio\_14day |
| 近30日平均播放时长 | watch\_avg\_time\_30day |
| 近30日5s有效观看占比 | watch\_finish\_ratio\_30day |
| 一级标签 | product\_type1 |
| 二级标签 | product\_type2 |

数据处理：归一化，标签化，类目embedding

评估报告：



分析原因：特征过少，模型无法捕捉关系

#### 历史版本3：

目标：1.播放时长大于5s

样本：

总量：600万

正样本：310万

负样本：290万

模型：din和gbdt

可用特征：

**去掉空置率大于50%的字段，加上有效播放序列，衍生特征，40个特征**

|  |  |  |
| --- | --- | --- |
| 视频id | video\_id | 0% |
| 手机号 | serv\_number | 0% |
| 是否分省视频 | is\_prov\_video | 0% |
| 是否15秒以内 | is\_15s\_video | 0% |
| 是否60秒以内 | is\_60s\_video | 0% |
| 是否30分钟以内 | is\_30m\_video | 0% |
| 是否近3天上架 | is\_3d\_update | 0% |
| 是否近7天上架 | is\_7d\_update | 0% |
| 一级标签 | product\_type1 | 2% |
| 二级标签 | product\_type2 | 2% |
| 是否营销类视频 | is\_market\_video | 2% |
| 省份编码 | prov\_code | 0% |
| 近7日平均播放时长 | watch\_avg\_time\_7day | 33% |
| 近7日5s有效观看占比 | watch\_finish\_ratio\_7day | 33% |
| 近14日平均播放时长 | watch\_avg\_time\_14day | 19% |
| 近14日5s有效观看占比 | watch\_finish\_ratio\_14day | 19% |
| 近30日平均播放时长 | watch\_avg\_time\_30day | 0% |
| 近30日5s有效观看占比 | watch\_finish\_ratio\_30day | 0% |
| 视频时长 | env\_video\_time | 0% |
| 点赞数 | feed\_like\_count | 3% |
| 转发数 | feed\_share\_count | 3% |
| 本月视频播放人数 | cur\_play\_video\_cnt | 28% |
| 本月视频播放人数占比 | cur\_play\_video\_ratio | 28% |
| 本月平均播放时长 | cur\_avg\_watch\_time | 28% |
| 本月视频5s播放占比 | cur\_5s\_video\_ratio | 28% |
| 上月视频播放人数 | pre\_play\_video\_cnt | 29% |
| 上月视频播放人数占比 | pre\_play\_video\_ratio | 29% |
| 上月平均播放时长 | pre\_avg\_watch\_time | 29% |
| 上月视频5s播放占比 | pre\_5s\_video\_ratio | 29% |
| 目标视频有效播放频次 | watch\_cnt |  |
| 目标视频类目1历史观看频次 | type1\_cnt |  |
| 目标视频类目2历史观看频次 | type2\_cnt |  |
| 历史最大播放时长 | max\_watch\_time |  |
| 历史最小播放时长 | min\_watch\_time |  |
| 历史视频时长最大值 | max\_video\_time |  |
| 历史视频时长最小值 | min\_video\_time |  |
| 历史视频时长均值 | avg\_video\_time |  |
| 历史有效播放视频id序列 | hist\_video\_id | 21% |
| 历史有效播放视频类目1序列 | hist\_product\_type1 | 21% |
| 历史有效播放视频类目2序列 | hist\_product\_type2 | 21% |

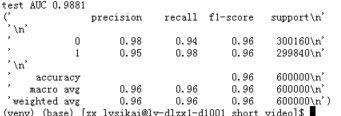
数据处理：

din：归一化，标签化，历史序列填充，衍生特征

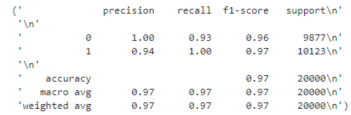
gdbt：类目emmbedding，历史序列训练word2vec，衍生特征

评估报告：

din:



gbdt:



#### 历史版本4：

目标：1.播放时长大于5s

样本：

总量：300万

正样本：150万

负样本：150万

权重：以观看时长作为样本权重

模型：din

召回：采用多路召回策略

规则：有浏览历史走模型，无历史走冷启动，计算视频热度

可用特征：

|  |  |  |
| --- | --- | --- |
| 视频id | video\_id | 0% |
| 手机号 | serv\_number | 0% |
| 是否分省视频 | is\_prov\_video | 0% |
| 是否15秒以内 | is\_15s\_video | 0% |
| 是否60秒以内 | is\_60s\_video | 0% |
| 是否30分钟以内 | is\_30m\_video | 0% |
| 是否近3天上架 | is\_3d\_update | 0% |
| 是否近7天上架 | is\_7d\_update | 0% |
| 一级标签 | product\_type1 | 2% |
| 二级标签 | product\_type2 | 2% |
| 是否营销类视频 | is\_market\_video | 2% |
| 省份编码 | prov\_code | 0% |
| 近7日平均播放时长 | watch\_avg\_time\_7day | 33% |
| 近7日5s有效观看占比 | watch\_finish\_ratio\_7day | 33% |
| 近14日平均播放时长 | watch\_avg\_time\_14day | 19% |
| 近14日5s有效观看占比 | watch\_finish\_ratio\_14day | 19% |
| 近30日平均播放时长 | watch\_avg\_time\_30day | 0% |
| 近30日5s有效观看占比 | watch\_finish\_ratio\_30day | 0% |
| 视频时长 | env\_video\_time | 0% |
| 点赞数 | feed\_like\_count | 3% |
| 转发数 | feed\_share\_count | 3% |
| 本月视频播放人数 | cur\_play\_video\_cnt | 28% |
| 本月视频播放人数占比 | cur\_play\_video\_ratio | 28% |
| 本月平均播放时长 | cur\_avg\_watch\_time | 28% |
| 本月视频5s播放占比 | cur\_5s\_video\_ratio | 28% |
| 上月视频播放人数 | pre\_play\_video\_cnt | 29% |
| 上月视频播放人数占比 | pre\_play\_video\_ratio | 29% |
| 上月平均播放时长 | pre\_avg\_watch\_time | 29% |
| 上月视频5s播放占比 | pre\_5s\_video\_ratio | 29% |
| 目标视频有效播放频次 | watch\_cnt |  |
| 目标视频类目1历史观看频次 | type1\_cnt |  |
| 目标视频类目2历史观看频次 | type2\_cnt |  |
| 历史最大播放时长 | max\_watch\_time |  |
| 历史最小播放时长 | min\_watch\_time |  |
| 历史视频时长最大值 | max\_video\_time |  |
| 历史视频时长最小值 | min\_video\_time |  |
| 历史视频时长均值 | avg\_video\_time |  |
| 历史有效播放视频id序列 | hist\_video\_id | 21% |
| 历史有效播放视频类目1序列 | hist\_product\_type1 | 21% |
| 历史有效播放视频类目2序列 | hist\_product\_type2 | 21% |

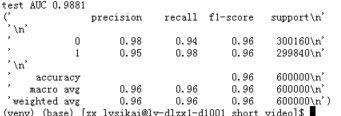
数据处理：

din：归一化，标签化，历史序列填充，衍生特征

模型推理优化：

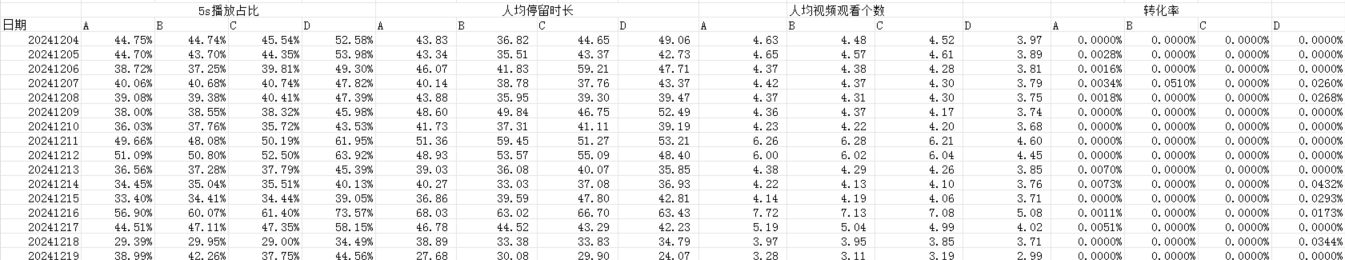
启用XLA、并行计算和模型预热

离线效果：



上线效果（模型D）：

5s播放占比相比以前历史版本C有明显提升，提升幅度6% - 11%



#### 历史版本5：

目标：1.播放时长大于5s

样本：

总量：300万

正样本：150万

负样本：150万

权重：以观看时长作为样本权重

模型：din

召回：采用多路召回策略，并增加多样性扩展

规则：有浏览历史走模型，无历史走冷启动，计算视频热度

可用特征：

结合历史版本特征，再加上视频效率数据特征，以及用户历史序列衍生特征

|  |  |  |
| --- | --- | --- |
| 视频id | video\_id | 0% |
| 手机号 | serv\_number | 0% |
| 是否分省视频 | is\_prov\_video | 0% |
| 是否15秒以内 | is\_15s\_video | 0% |
| 是否60秒以内 | is\_60s\_video | 0% |
| 是否30分钟以内 | is\_30m\_video | 0% |
| 是否近3天上架 | is\_3d\_update | 0% |
| 是否近7天上架 | is\_7d\_update | 0% |
| 一级标签 | product\_type1 | 2% |
| 二级标签 | product\_type2 | 2% |
| 是否营销类视频 | is\_market\_video | 2% |
| 省份编码 | prov\_code | 0% |
| 近7日平均播放时长 | watch\_avg\_time\_7day | 33% |
| 近7日5s有效观看占比 | watch\_finish\_ratio\_7day | 33% |
| 近14日平均播放时长 | watch\_avg\_time\_14day | 19% |
| 近14日5s有效观看占比 | watch\_finish\_ratio\_14day | 19% |
| 近30日平均播放时长 | watch\_avg\_time\_30day | 0% |
| 近30日5s有效观看占比 | watch\_finish\_ratio\_30day | 0% |
| 视频时长 | env\_video\_time | 0% |
| 点赞数 | feed\_like\_count | 3% |
| 转发数 | feed\_share\_count | 3% |
| 本月视频播放人数 | cur\_play\_video\_cnt | 28% |
| 本月视频播放人数占比 | cur\_play\_video\_ratio | 28% |
| 本月平均播放时长 | cur\_avg\_watch\_time | 28% |
| 本月视频5s播放占比 | cur\_5s\_video\_ratio | 28% |
| 上月视频播放人数 | pre\_play\_video\_cnt | 29% |
| 上月视频播放人数占比 | pre\_play\_video\_ratio | 29% |
| 上月平均播放时长 | pre\_avg\_watch\_time | 29% |
| 上月视频5s播放占比 | pre\_5s\_video\_ratio | 29% |
| 近3天点赞量 | like\_3d\_cnt |  |
| 近3天分享量 | share\_3d\_cnt |  |
| 近3天播放次数 | play\_video\_3d\_cnt |  |
| 近3天播放占比 | play\_video\_3d\_ratio |  |
| 近3天平均观看时长 | avg\_3d\_watch\_time |  |
| 近3天5s播放占比 | video\_5s\_3d\_ratio |  |
| 近7天点赞量 | like\_7d\_cnt |  |
| 近7天分享量 | share\_7d\_cnt |  |
| 近7天播放次数 | play\_video\_7d\_cnt |  |
| 近7天播放占比 | play\_video\_7d\_ratio |  |
| 近7天平均观看时长 | avg\_7d\_watch\_time |  |
| 近7天5s播放占比 | video\_5s\_7d\_ratio |  |
| 近14天点赞量 | like\_14d\_cnt |  |
| 近14天分享量 | share\_14d\_cnt |  |
| 近14天播放次数 | play\_video\_14d\_cnt |  |
| 近14天播放占比 | play\_video\_14d\_ratio |  |
| 近14天平均观看时长 | avg\_14d\_watch\_time |  |
| 近14天5s播放占比 | video\_14s\_7d\_ratio |  |
| 星期 | week |  |
| 时间段 | time\_period |  |
| 时间间隔 | interval |  |
| 目标视频有效播放频次 | watch\_cnt |  |
| 目标视频类目1历史观看频次 | type1\_cnt |  |
| 目标视频类目1历史3天观看频次 | type1\_3day\_cnt |  |
| 目标视频类目1历史7天观看频次 | type1\_7day\_cnt |  |
| 目标视频类目1历史14天观看频次 | type1\_14day\_cnt |  |
| 目标视频类目2历史观看频次 | type2\_cnt |  |
| 目标视频类目2历史3天观看频次 | type2\_3day\_cnt |  |
| 目标视频类目2历史7天观看频次 | type2\_7day\_cnt |  |
| 目标视频类目2历史14天观看频次 | type2\_14day\_cnt |  |
| 历史最大播放时长 | max\_watch\_time |  |
| 历史最小播放时长 | min\_watch\_time |  |
| 历史视频时长最大值 | max\_video\_time |  |
| 历史视频时长最小值 | min\_video\_time |  |
| 历史视频时长均值 | avg\_video\_time |  |
| 历史有效播放视频id序列 | hist\_video\_id | 21% |
| 历史有效播放视频类目1序列 | hist\_product\_type1 | 21% |
| 历史有效播放视频类目2序列 | hist\_product\_type2 | 21% |

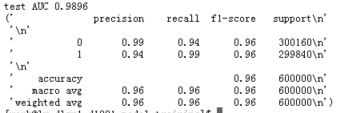
数据处理：

din：归一化，标签化，历史序列填充，衍生特征

模型推理优化：

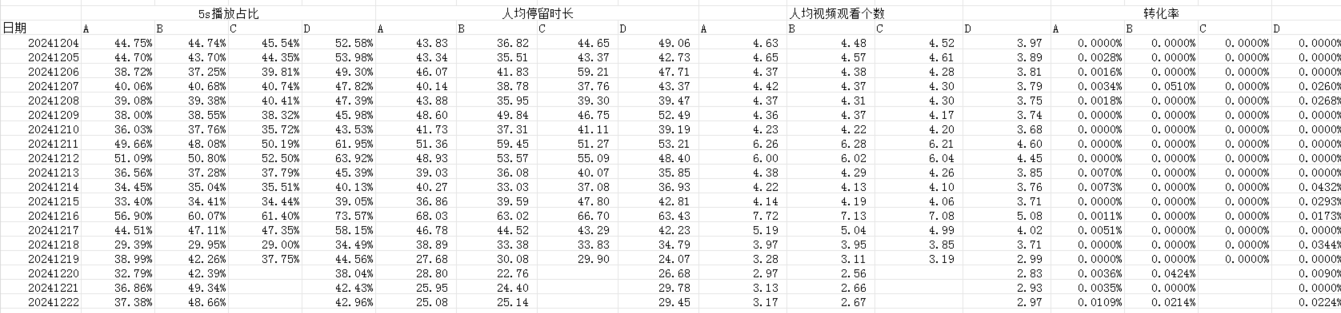
启用XLA、并行计算和模型预热

离线效果：



上线效果（模型B）：

5s播放占比相比以前历史版本D有明显提升，提升幅度6%



# 模型复盘和优化计划

1. 数据方面1

数据要区分用户数据，还是产品数据，目前看，数据还是没用完全使用。

2、数据方面2

1. 完全没有用户基础画像数据（如年龄、性别、职业、使用套餐价格、兴趣标签、）
2. 没有用户的行为的视频类标信息（比如点赞视频、分享视频，观看视频），即与视频发生的时间关系。
3. 完全没有产品的属性数据
4. 确实用户与某个视频之间的关系，比如这条样本，只知道部分用户观看信息，视频本身的播放信息，但是这个用户对这个视频的行为信息没有。
5. 缺少实时数据，用户进来后的实时行为，无法传递给模型侧。

这两类数据的缺失，导致用户目前的行为数据，很难与实际产品产生关系，两者之间没有形成关联。导致模型是无法学习和挖掘到用户的兴趣和产品的客群的。

3、模型方面

1、目前是业务规则召回，然后模型排序。这导致模型对整个推荐的影响较弱。这些按照业务召回的产品，可能本身就不是用户的兴趣爱好。

2、召回或者排序模型中，没有冷启动策略，新用户也默认按照业务规则召回和模型排序。

3、加入规则，统计热门视频，对新用户直接推荐热门产品。

TODO：

1、新版本模型上线，及效果分析。@威龙

2、模型全业务流程的梳理（召回、排序、展示）。@威龙@昆鹏

3、数据梳理和数据优化

（1）当前使用数据的详细准确梳理：目前有哪些用户数据、哪些产品数据、哪些构造的数据。搞到一个字典表中，不同类型的数据分别标识出来。@威龙@吴晨

（2）数据需求：针对目前的数据情况，对模型优化还需要哪些数据？（比如用户的年龄、性别、职业、使用套餐价格、兴趣标签、播放分享点赞列表、实时数据等），考虑有历史行为无历史行为，查阅资料从尽可能多的维度丰富数据。@威龙@吴晨

4、模型优化。@威龙@王聪

（1）召回侧的多路召回业务优化，提供详细的召回方案。

（2）冷启动策略：无历史行为用户的冷启动、新上线视频的推荐

（3）规则模型：比如热销规则（热门类别，热门列别下的热门视频）