

# 视频初始缓冲时延机器学习分析报告

廖文哲

# 视频业务驱动-视频初始缓冲时延机器学习分析报告

1 视频感知理论体系研究



2 明确需求，制定计划

3 数据采集与数据预处理

4 决策树算法实现与模型评估

5 多轮数据分析与模型参数调整

6 经验总结

# 影响视频感知的网络因素

据某权威调研机构表明，影响用户在线视频体验的TOP5因素为：视频流畅度、流量消耗、初始缓冲时延、画质清晰度和视频是否收费。

和网络侧KPI息息相关的三大因素分别为：

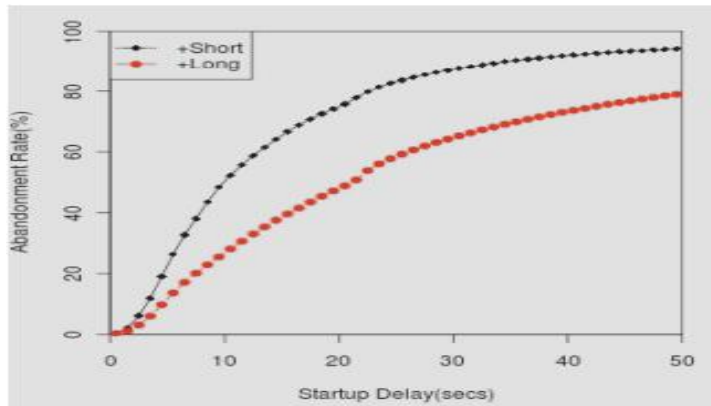


初始缓冲时延公式：

$$\text{初始缓冲时延} \text{delay} = \frac{t \cdot V_{\text{编码速率}} - \text{Volume}_{\text{SlowStart}}}{V_{\text{initial}}} + (n1 + n2) * RTT \quad (1)$$

- $n1$  视频头文件交互阶段需要 $n1$ 个RTT
- $n2$  TCP慢启动阶段需要 $n2$ 个RTT,  $n=n1+n2$
- $\text{Volume}_{\text{SlowStart}}$  慢启动过程下载的数据量, 为 $m \cdot 1310 \cdot 8$

## 视频2S初始缓冲区的由来

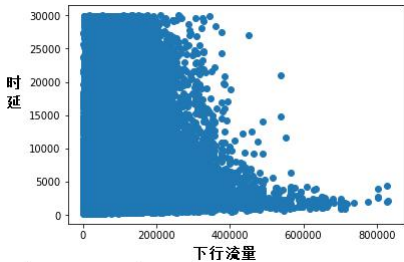


美国麻省大学教授拉梅什·西塔拉曼(Ramesh Sitaraman)的一项研究显示，对很大一部分用户来说，如果视频无法在2秒钟内完成加载，那么这些用户将放弃观看该视频。

# 几种常用的相关分析方法

相关分析（Analysis of Correlation）是数据分析中经常使用的分析方法之一。通过对不同特征或数据间的关系进行分析，发现数据中的关键影响及驱动因素。并对业务的发展进行预测。相关分析的方法很多，初级的方法可以快速发现数据之间的关系，如正相关，负相关或不相关。中级的方法可以对数据间关系的强弱进行度量，如完全相关，不完全相关等。高级的方法可以将数据间的关系转化为模型，并通过模型对未来的业务发展进行预测。

## 1.图表相关分析（折线图及散点图）



## 2.协方差及协方差矩阵

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

将数据进行可视化处理，简单的说就是绘制图表。优点是对相关关系的展现清晰，缺点是无法对相关关系进行准确的度量，缺乏说服力。并且当数据超过两组时也无法完成各组数据间的相关分析。

协方差通过数字衡量变量间的相关性，正值表示正相关，负值表示负相关。但无法对相关的密切程度进行度量

## 3.相关系数

|                        |  |
|------------------------|--|
| INTER_FREQ_ATT         | 0.032685259501369, 1.996817437139217e-290) |
| VIDEO_SERVER_AVG_RTT   | (0.491452, 0.0)                            |
| INIT_BUFFER_DL_RATE_4S | (-0.31521819098137355, 0.0)                |
| 下行PRB平均利用率             | (0.26404402950126554, 0.0)                 |
| NERTEDL_QPSK           | (0.25933547895455866, 0.0)                 |
| NERTEDL_16QAM          | (0.2548904786233423, 0.0)                  |
| DL_PRB_USED            | (0.25474960827928617, 0.0)                 |

相关系数的优点是可以通过数字对变量的关系进行方向性的度量，1表示正相关，-1表示负相关，越靠近0相关性越弱。缺点是无法利用这种关系对数据进行预测，简单的说就是没有对变量间的关系进行提炼和固化，形成模型。

## 4.一元回归及多元回归

$$y = b_0 + b_1x$$

回归分析是确定两组或两组以上变量间关系的统计方法。回归分析按照变量的数量分为一元回归和多元回归，两个变量使用一元回归，两个以上变量使用多元回归，回归分析可对数据进行预测，但对特征值的数量有所限制。

## 5.信息熵及互信息（决策树）

信息熵是接收的每条消息中包含的信息的平均量，消息代表来自分布或数据流中的事件、样本或特征。在信息世界，熵越高，则能传输越多的信息，熵越低，则意味着传输的信息越少。对于特征值特多别的数据，度量这些文本特征值之间相关关系的方法就是互信息。通过这种方法我们可以发现哪一类特征与最终的结果关系密切。

1 视频感知理论体系研究

2 明确需求，制定计划



3 数据采集与数据预处理

4 决策树算法实现与模型评估

5 多轮数据分析与模型参数调整

6 经验总结

## 明确需求，制定计划

## 项目需求

前期与客户会议讨论，确认了本次专题的需求：

目标1：基于SDK详单数据进行初缓时延定界；

目标2：关联指标的拐点分析；

## 专题进度计划

本次专题计划时间一个月，主要利用机器学习对视频初始缓冲时延进行科学分析，找出影响时延好坏的OMC无线指标，同时输出指标门限值，便于后续优化KPI，提高用户观看视频体验。

| 时间/工作内容安排     | 2018/4/4 | 2018/4/8 | 2018/4/12 | 2018/4/16 | 2018/4/20 | 2018/4/24 | 2018/4/28 | 2018/5/2 |
|---------------|----------|----------|-----------|-----------|-----------|-----------|-----------|----------|
| 明确数据源         |          |          |           |           |           |           |           |          |
| 采集数据源         |          |          |           |           |           |           |           |          |
| 探讨明确分析需要      |          |          |           |           |           |           |           |          |
| 确定分析模型        |          |          |           |           |           |           |           |          |
| 检查数据完整性       |          |          |           |           |           |           |           |          |
| 数据整理清洗        |          |          |           |           |           |           |           |          |
| 分析模型搭建        |          |          |           |           |           |           |           |          |
| 分析模型代码实现      |          |          |           |           |           |           |           |          |
| 进行模型仿真        |          |          |           |           |           |           |           |          |
| 调整模型参数，输出最优结果 |          |          |           |           |           |           |           |          |
| 结果可视化         |          |          |           |           |           |           |           |          |
| 结果呈现          |          |          |           |           |           |           |           |          |
| 整理汇报（PPT）     |          |          |           |           |           |           |           |          |
|               |          |          |           |           |           |           |           |          |
|               | 已完成      |          |           |           |           |           |           |          |
|               | 进行中      |          |           |           |           |           |           |          |
|               | 未完成      |          |           |           |           |           |           |          |
| 采用周报汇报进度      |          |          |           |           |           |           |           |          |

## 模型选择

## 数据特征分析

本次机器学习的研究目标数据为视频初始缓冲时延，其是连续性的因变量，来自于DO上用户面单数数据

与之对应的指标是网管上15分钟粒度的OMC小区级指标。特征值总共80多项。

特征值多 (80多项)

## 特征值属性不统一

针对此次项目的需求结合数据特征的分析，最终采用了决策树算法作为本次专题的主要研究手段。决策树具有以下优势：

- 1: 理解和解释起来简单, 且决策树模型可以想象 (便于理解)
- 2: 决策树算法的时间复杂度较低
- 3: 能够处理多种数字和数据的类别 (处理多特征值)
- 4: 能够处理多输出的问题
- 5: 能使用统计检验来验证模型可靠性



1 视频感知理论体系研究

2 明确需求，制定计划

3 数据采集与数据预处理



4 决策树算法实现与模型评估

5 多轮数据分析与模型参数调整

6 经验总结

# 数据采集与数据预处理

## 数据采集

采集一周的视频话单数据与OMC15分钟粒度的无线指标关联，并存储到DO数据库中。关联后的表名为：sdk\_omc\_relevance，通过PL/SQL查询表，直接下载CSV文件到本地进行机器学习数据分析。

|     |         |                        |
|-----|---------|------------------------|
| 行数  | 1648011 | 用户级/小区级                |
| 列数  | 185     | 多种属性（字符，整数，小数）         |
| 自变量 | 80项     | OMC相关指标                |
| 因变量 | 1项      | CREATE_TO_PLAY_TIMEOTT |

原始数据示例：



关联表字段明细

关联脚本：



关联脚本

## 数据预处理

### 数据质量

- 准确性（检查数据是否是我们需要的数据）
- 完整性（检查数据是否采集完整）
- 时效性（确认数据的采集时间是否正确）

### 数据清理

- 缺失值（对空值进行数据删除）
- 异常值（对异常值进行变换或者删除）
- 数据类型（对不合适的数据类型进行转化）

### 数据离散化

- 视频初始缓冲时延连续属性离散化，将时延界定为两种类别（卡或不卡）



1 视频感知理论体系研究

2 明确需求，制定计划

3 数据采集与数据预处理

4 决策树算法实现与模型评估



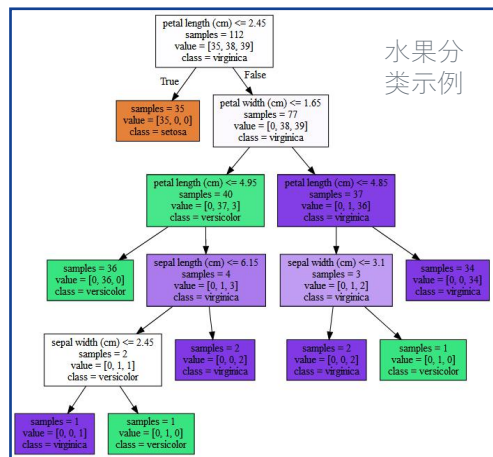
5 多轮数据分析与模型参数调整

6 结果呈现，经验总结

# 决策树算法实现与模型评估

## 决策树模型介绍

决策树（Decision Tree）是一种基本的分类与回归方法。决策树模型呈树形结构，在分类问题中，表示基于特征对实例进行分类的过程。它可以认为是if-then规则的集合，也可以认为是定义在特征空间与类空间上的条件概率分布。



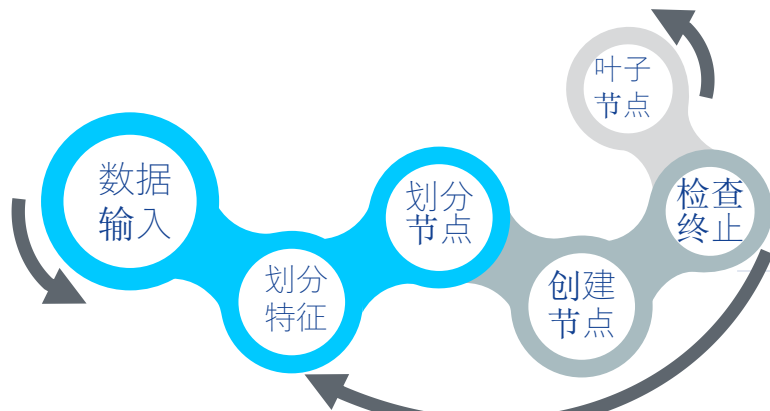
分类的时候，从根节点开始，对实例的某一个特征进行测试，根据测试结果，将实例分配至其子结点；此时，每一个子结点对应着该特征的一个取值。如此递归向下移动，直至达到叶结点，最后将实例分配到叶结点的类中。

## 决策树算法框架

主函数：本质上是个递归函数，该函数主要功能是根据某种规则生长出决策树的各个分支节点，并根据终止条件结束算法。

数据输入：分类的数据集和类别标签。

划分过程：1) 根据某种分类规则（依据信息增益进行特征选取和分裂）得到最优的划分特征，并创建特征的划分节点。2) 按照该特征的每个取值划分数据集为若干部分。3) 根据划分子函数的计算结果构建出新的节点，作为树生长出的新分支。4) 检验是否符合递归终止条件（如果增益够大，就分裂为左右子树，如果增益很小，就停止分裂，将这个节点直接作为叶子节点）。5) 将划分的新节点包含的数据集和类别标签作为输入，递归执行上述步骤。



# 决策树算法实现与模型评估

## 机器学习模型评价指标-混淆矩阵

**True Positive**：真正类(TP)，样本的真实类别是正类，模型预测成为正类

**False Negative**：假负类(FN)，样本的真实类别是正类，模型预测成为负类

**False Positive**：假正类(FP)，样本的真实类别是负类，模型预测成为正类

**True Negative**：真负类(TN)，样本的真实类别是负类，模型预测成为负类

|                |          | Predicted condition           |                       |                   |
|----------------|----------|-------------------------------|-----------------------|-------------------|
|                |          | positive                      | negative              |                   |
| True condition | positive | True Positive                 | False Negative (预测错误) | Recall=TP/(TP+FN) |
|                | negative | False Positive (预测错误)         | True Negative         | FPR=FP/(FP+TN)    |
|                |          | $F=(2 \cdot P \cdot R)/(P+R)$ | $P=TP/(TP+FP)$        | $FOR=FN/(FN+TN)$  |

F：综合了Precision和Recall的一个判断指标，F1-Score的值是从0到1的，1是最好，0是最差

Precision：查准率，在模型预测为正类的样本中，真正为正类的样本所占的比例。一般情况下，查准率越高，说明模型的效果越好。

FOR:表示在模型预测为负类的样本中，真正的正类所占的比例。一般情况下，错误遗漏率越小，模型的效果越好。

Recall:召回率，模型预测为正类的样本的数量，占总正类样本数量的比值。一般情况下，Recall越高，说明有更多的正类样本被模型预测正确，模型的效果越好。

FPR：模型预测为正类的样本中，占模型负类样本数量的比值。一般情况下，假正类率越低，说明模型的效果越好。

Accuracy：模型的精度，即模型预测正确的个数/样本的总个数。一般情况下，模型的精度越高，说明模型的效果越好。

1 视频感知理论体系研究

2 明确需求，制定计划

3 数据采集与数据预处理

4 决策树算法实现与模型评估

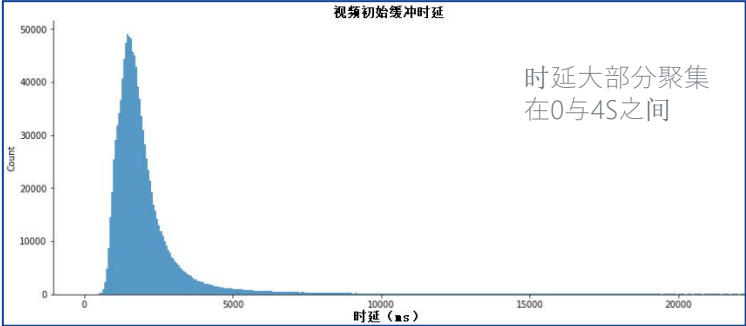
5 多轮数据分析与模型参数调整



6 经验总结

# 多轮数据分析与模型参数调整

## 视频初始缓冲时延数据统计分析



| 用户维度           | 3月份     | 4月份     |
|----------------|---------|---------|
| 总数据量           | 1284311 | 1046925 |
| 携带IMSI的用户条数    | 1282294 | 1045577 |
| IMSI数量         | 104463  | 95765   |
| 条数均值（每个用户贡献条数） | 13.09   | 11.61   |
| 时延大于4S的条数      | 83512   | 72421   |
| 时延大于5S的条数      | 60608   | 53780   |
| 时延均值（大于5S的用户）  | 9.5509  | 10.5138 |
| 时延大于5S的用户数     | 4101    | 3783    |
| 时延均值（全量）       | 2.30423 | 2.4275  |

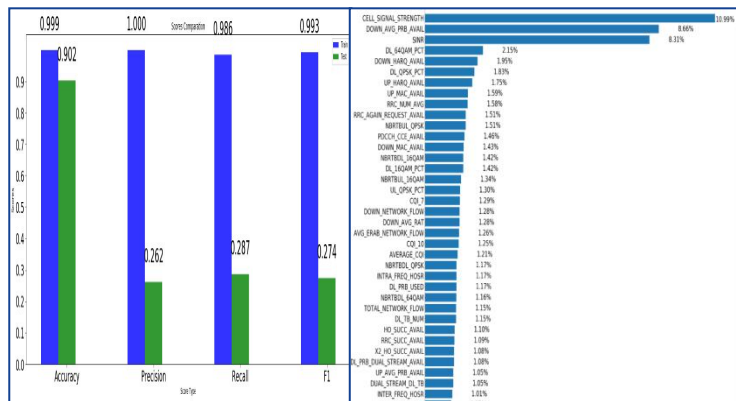
| 字段                     | 相关系数                        |
|------------------------|-----------------------------|
| CREATE_TO_PLAY_TIMEOTT | (1,0)                       |
| VIDEO_SERVER_AVG_RTT   | (0.491452, 0.0)             |
| INIT_BUFFER_DL_RATE_4S | (-0.31521819098137355, 0.0) |
| 下行PRB平均利用率             | (0.26404402950126554, 0.0)  |
| NBRTBDL_QPSK           | (0.25933547895455866, 0.0)  |
| NBRTBDL_16QAM          | (0.2548904786233423, 0.0)   |
| DL_PRB_USED            | (0.25474960827928617, 0.0)  |
| CELL_SIGNAL_STRENGTH   | (-0.22969810475109215, 0.0) |
| 下行传输TB数                | (0.22903685523798467, 0.0)  |
| PDCCH信道CCE占用率          | (0.22752045548619756, 0.0)  |

对原始数据做统计性分析，3月份总数据量在128万左右，总共用户数在10万左右，平均每个用户上报13条左右话单信息，平均时延等于2.3S,时延大于4S的样本数为8万左右，同时对KQI与KPI进行皮尔逊相关系数分析，发现视频初始缓冲时延与VIDEO\_SERVER\_AVG\_RTT与较强的线性关系，与其他的KPI指标相关系数都在0.25以下，两者之间不存在强的线性相关。

## 第一轮数据分析：

因变量：视频初始缓冲时延

class\_weight : None



### 重要度排名（占比大于1%）

## 调整决策树模型参数：

因变量：视频初始缓冲时延

class\_weight :none/balanced

|           | criterion |         |
|-----------|-----------|---------|
| 评估分项      | gini      | entropy |
| Accuracy  | 0.902     | 0.902   |
| Precision | 0.262     | 0.262   |
| Recall    | 0.287     | 0.279   |
| F1        | 0.274     | 0.273   |

|           | class_weight |          |
|-----------|--------------|----------|
| 评估分项      | none         | balanced |
| Accuracy  | 0.902        | 0.905    |
| Precision | 0.262        | 0.265    |
| Recall    | 0.287        | 0.264    |
| F1        | 0.274        | 0.265    |

|           | max_depth |       |       |       |       |       |       |       |       |       |
|-----------|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 评估分项      | 4         | 5     | 6     | 7     | 8     | 9     | 10    | 11    | 12    | 13    |
| Accuracy  | 0.936     | 0.936 | 0.937 | 0.937 | 0.937 | 0.936 | 0.936 | 0.935 | 0.935 | 0.934 |
| Precision | 0.541     | 0.542 | 0.555 | 0.563 | 0.563 | 0.564 | 0.568 | 0.570 | 0.572 | 0.462 |
| Recall    | 0.165     | 0.179 | 0.174 | 0.183 | 0.186 | 0.202 | 0.213 | 0.236 | 0.237 | 0.234 |
| F1        | 0.253     | 0.269 | 0.265 | 0.276 | 0.280 | 0.297 | 0.310 | 0.334 | 0.335 | 0.311 |

**NOKIA 上海贝尔**

# 多轮数据分析与模型参数调整

## 第二轮数据分析：

样本判定门限：4S

数据量：128万条

自变量：KPI指标 (top12)

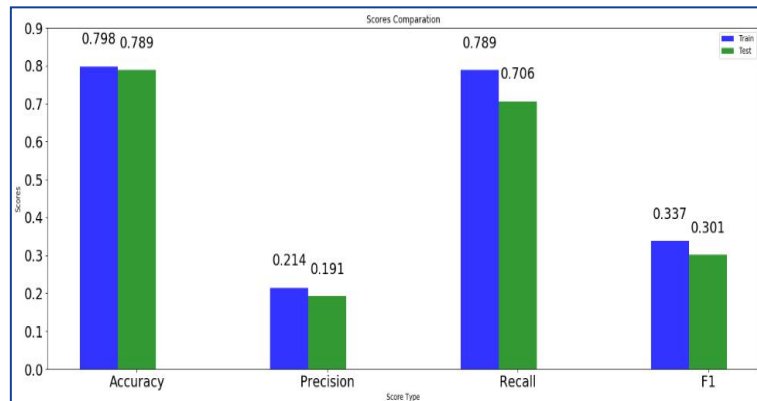
因变量：视频初始缓冲时延

模型参数：

Criterion：gini

max\_depth：12

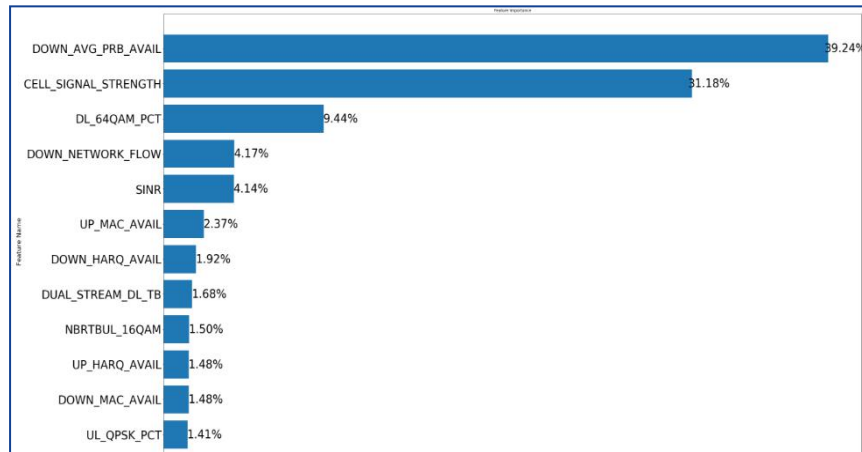
class\_weight :None



仿真结果四项评估得分

通过筛选top12的KPI指标以及依据第一次的参数优化设置模型参数进行第二轮数据分析，评估得分部分提高，Recall得分大比例提升，同时训练集与测试集评估得分差值缩小了很多，说明模型的泛化性得到了提升，但是P值还是只有0.2,需要进一步提升。

## KPI重要度排名：

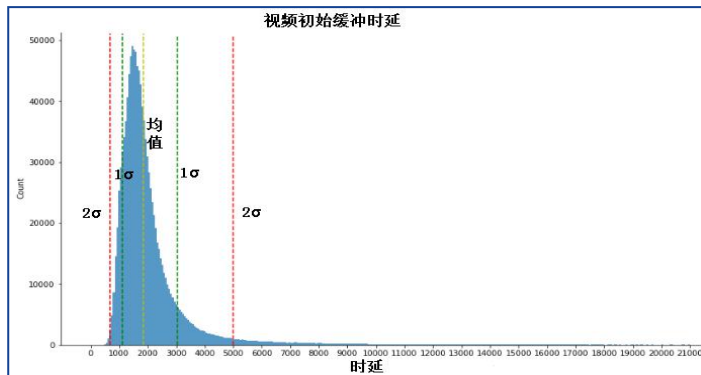


分析KPI指标重要度，对视频初始缓冲视频影响的KPI排名为：下行PRB利用率，RSRP，DL\_64QAM\_PCT，下行流量，SINR，MAC层上行误块率，下行HARQ重传比例，双流下行传输TB数，NBRTBUL\_16QAM，上行HARQ重传比例，MAC层下行误块率，UL\_QPSK\_PCT

# 多轮数据分析与模型参数调整

## 第三轮数据分析：

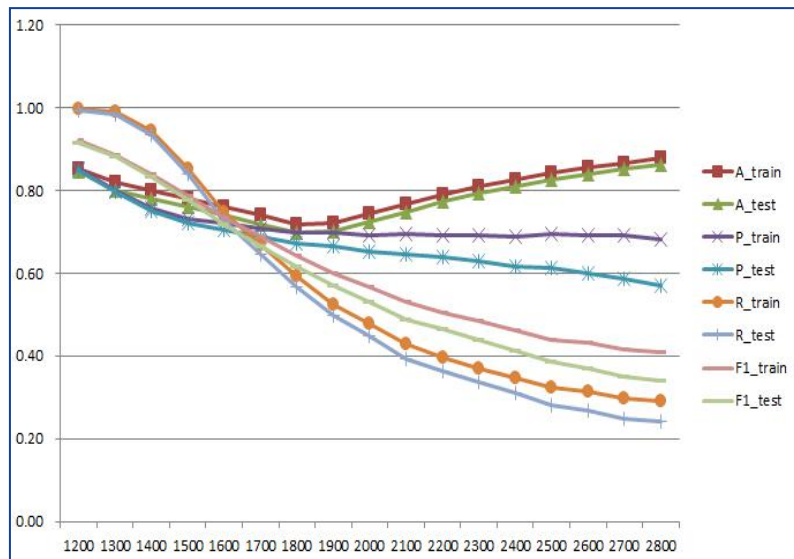
### 初缓时延 正太分布 分析



|                             |         |
|-----------------------------|---------|
| 时延均值                        | 7.53    |
| 时延标准差                       | 0.49    |
| 时延均值对应的时延                   | 1854.51 |
| 时延 $1\sigma$ 左边界            | 1130.91 |
| 时延 $1\sigma$ 右边界            | 3041.10 |
| 时延 $2\sigma$ 左边界            | 689.64  |
| 时延 $2\sigma$ 右边界            | 4986.93 |
| 时延大于 $2\sigma$ 的数据量         | 60608   |
| 时延小于 $2\sigma$ 的数据量         | 6306    |
| $2\sigma$ 内数据量占比            | 94.79%  |
| 时延大于 $2\sigma$ 的用户数（时延均值维度） | 4101    |
| 时延小于 $2\sigma$ 的用户数（时延均值维度） | 55      |
| $2\sigma$ 内用户数占比（时延均值维度）    | 96.02%  |

通过对初缓时延正太分布变化，观察统计数据，我们发现数据在 $2\sigma$ 内的数据量与用户数占比95%以上，大于 $2\sigma$ 的数据量占比为4.72%，用户数占比为3.92%。对于时延超过 $2\sigma$ 数据量有可能是外界因素导致的时延加大（如手机硬件卡顿，网站传输慢）

为了探寻初始缓冲时延界定卡顿的门限值，我们采取不同卡顿门限值进行仿真训练。通过多轮数据分析，发现随着界定门限的增加，模型的泛化性逐步减弱，2S之后泛化性减弱趋势明显加大，同时模型的准确度在2S之前一直在减小，2S之后逐渐增大，可见2S是模型准确度的拐点。





# 多轮数据分析与模型参数调整

## 第四轮数据分析：

样本判定门限：2S

数据量：121万条(2 $\sigma$ )

自变量：KPI指标 (top12)

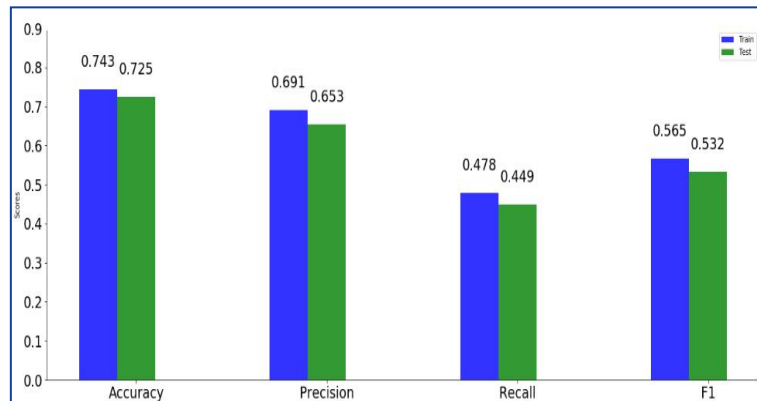
因变量：视频初始缓冲时延

模型参数：

Criterion：gini

max\_depth：12

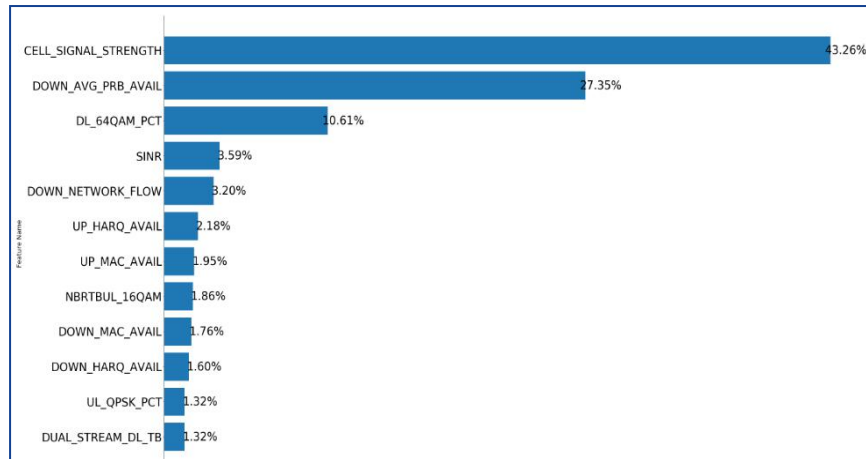
class\_weight :None



仿真结果四项评估得分

2 $\sigma$ 外的数据我们认为存在较大的其他外界干扰，所以我们筛选出2 $\sigma$ 内的数据，同时选择top12的KPI指标以及将取2S为界定门限，评估得分得到很大的提升，，同时训练集与测试集评估得分差值缩小了很多，模型的泛化性的泛化性较强，此模型可较好的判断出初缓时延的好坏。

## KPI重要度排名：



分析KPI指标重要度，对视频初始缓冲视频影响的KPI排名为：RSRP，下行PRB利用率，DL\_64QAM\_PCT，SINR,下行流量，上行HARQ重传比例,MAC层上行误块率，NBRTBUL\_16QAM, MAC层下行误块率,下行HARQ重传比例，UL\_QPSK\_PCT,双流下行传输TB数，

# 多轮数据分析与模型参数调整

第四轮数据分析：通过决策树规则，可筛选出准确度较高的辨别规则门限值

| KPI                  | 门限值    |
|----------------------|--------|
| CELL_SIGNAL_STRENGTH | -100   |
| DOWN_AVG_PRB_AVAIL   | 36     |
| DL_64QAM_PCT         | 33     |
| SINR                 | 5      |
| UP_HARQ_AVAIL        | 2      |
| UP_MAC_AVAIL         | 13     |
| NBRTBUL_16QAM        | 308123 |
| DOWN_MAC_AVAIL       | 1      |
| DOWN_HARQ_AVAIL      | 10     |
| UL_QPSK_PCT          | 20     |
| DUAL_STREAM_DL_TB    | 60     |

上表为采用2S界定门限而输出的top11KPI的门限值。

| KPI                  | 门限值    |
|----------------------|--------|
| DOWN_AVG_PRB_AVAIL   | 49     |
| CELL_SIGNAL_STRENGTH | -105   |
| SINR                 | 2.5    |
| DL_64QAM_PCT         | 21     |
| DOWN_HARQ_AVAIL      | 13     |
| NBRTBUL_16QAM        | 308123 |
| UP_MAC_AVAIL         | 14     |
| UL_QPSK_PCT          | 29     |
| UP_HARQ_AVAIL        | 5      |
| DUAL_STREAM_DL_TB    | 20     |
| DOWN_MAC_AVAIL       | 0.24   |

上表采用4S界定门限而输出的top11KPI的门限值。

1 视频感知理论体系研究

2 明确需求，制定计划

3 数据采集与数据预处理

4 决策树算法实现与模型评估

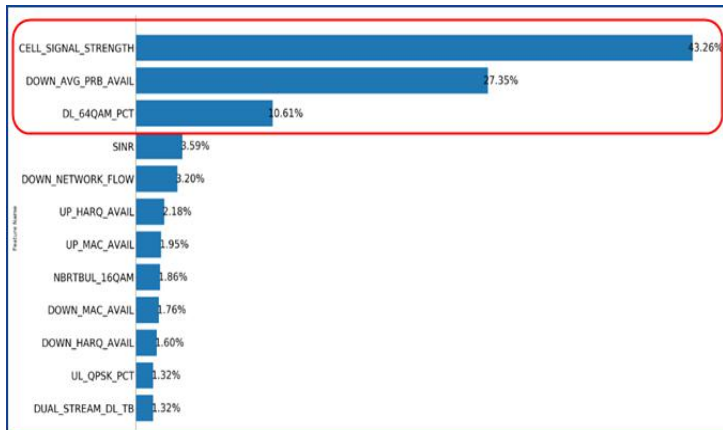
5 多轮数据分析与模型参数调整

6 经验总结



# 经验总结

进一步筛选，将KPI重要度占比较大的TOP3作为重点优化对象。（RSRP，下行PRB利用率， DL\_64QAM\_PCT）



KPI  
门限值

| KPI                  | 门限值(2S) | 门限值(4S) |
|----------------------|---------|---------|
| CELL_SIGNAL_STRENGTH | -100    | -105    |
| DOWN_AVG_PRB_AVAIL   | 36      | 49      |
| DL_64QAM_PCT         | 33      | 21      |



TOP3全规则图



2S卡顿界定门限  
规则

## 创新点1

引入机器学习进行数据分析，发掘传统人工发现不了问题



## 创新点2

探寻视频缓冲时延界定门限，2S的时候卡顿与不卡顿更加容易辨别



## 创新点5

通过决策树规则我们输出了视频KQI劣化的KPI门限



## 创新点4

通过决策树算法我们实现了视频初缓时延与KPI的重要度关联



## 创新点3

当视频缓冲时延超过某个阈值时，KPI与时延的关联性减小，即KPI对时延的影响有一定的范围作用。



---

# THANKS

---