**游戏数据挖掘报告内容和建模思路**

**数据挖掘报告内容：**

**1. 版本变更内容说明**

**2. 数据可视化**

**3. 统计建模**

**3.1 RF模型**

**3.2 AB test模型**

**4. 结论报告**

**=========================================================================**

**1. 版本变更说明**

说明游戏的版本变动内容以及依据这些变动确定参数矩阵(包含直接受影响的参数矩阵 + 激励和插页观看次数矩阵)，同时描述游戏的基础信息。实际统计建模过程中只选择**reward + inter + total** 这三个metrics.

**2. 数据可视化**

此部分分为两块，第一块是根据游戏的基础数据信息（留存率，广告点击次数之类的重要指标）进行可视化，此处会用到最为基础的柱状图，饼图，漏斗模型图等。第二块是服务于RF，聚类模型和AB test的可视化，使用到RF模型的用户散点图，基于参数矩阵的频率分布柱状图，箱型图，交互作用图等。

**3. 统计建模（数据中台内部使用）**

对所选择的模型的大概描述和介绍，比如，目前做AB test所选择的 ANOVA 和 wilcoxon mann whitney test。然后给出具体统计模型结果表。具体建模过程以appendix的形式放在最后以供验收和校正使用。数据中台内部需要把统计模型的结论和经验主义所做出的判断一起与实际情况进行比较，不断修正模型。

**4. 结论报告**

基于统计结果做出结论。比如版本之间的广告平均点击次数是否有显著性差异，直接受影响的参数矩阵是否影响了最终的目标矩阵，一次版本调整是否提升了激励广告点击率，而激励广告的点击率升高是否会直接使激励的点击次数提升。

**建模思路：**

**RF模型 & 聚类**

1. RF模型的目的是给所有游戏用户做分类找出 “有问题但有救用户”。该类用户是在游戏时长或者点击次数某一维度存在问题的用户，找出该部分用户是为了进一步做归因分析以及AB test里可能涉及到的抽样设计，归因分析可作为后续结合经验调整版本的依据。用户在游戏时长或者点击次数都很低的不会被作为目标用户是因为该类用户在游戏中的行为过少无法分析，这部分用户只能凭借经验主义去优化。同时，游戏时长和点击次数都高的“优秀用户”也会被监控，确保版本的变化不会影响该部分用户。
2. 通过RF模型的分类找出目标群体后，使用K-means方法或clustering方法给目标群体进行聚类分群，使用不同的游戏内指标（比如每一关的激励点击次数，死亡次数，游玩时长等）一起作为分类标准，最终把目标群体按照他们游戏内的不同行为分成几组，而后通过每组内群体的游戏行为的不同给出每一类用户的标签，并能推测他们游戏时长或点击次数不足的原因。

**版本比较 ( AB test )**

1. 首先，分别在A，B两个版本内做AA test，此处是A版本和B版本内部的比较。如果有必要使用RF模型中的分类进行抽样设计，此处使用分层抽样法。之后选择全部用户或样本进行建模，目前选择 repeated measure model。使用该模型分别检验A版本的metric是否有显著性差异和B版本的metric是否有显著性差异。理想情况是，A版本并无明显显著性差异，B版本有显著性差异。
2. 选择reward，inter，total 三个metrics做AB test，此处是对A，B版本间比较。目前使用 2 sample t test和 nonparametric test (wilcoxon mann whitney test) 结合winsorization method，按不同天数检验A版本的metric和B版本的metric是否有显著性差异。理想情况是，A和B版本在t test和 nonparametric test上都得出具有显著性差异的统计结果。
3. 相关性分析，直接受影响的参数矩阵和激励和插页观看次数矩阵是否存在正相关。目前合适的模型还在选择过程中。以下方法都在考虑范围内。（**该部分脱离出AB test部分**）

linear regression，kendall test，spearman sign test，chisquare test，kendall coeficient，spearman coeficient rank correlation

1. 优化：缩短AB test时间周期，增加factor的维度（从两个版本比较变为多个版本比较），增加影响因子factor（除版本外的其它因素）