NMF乘客留存分层

以30天以上没有订单的口径来衡量，目前平台沉默乘客约5000W+，占总体乘客体系的10%以上。

对于个体来说，在沉默之前都会存在一定的行为变化，从平台的角度来看，这种行为变化会体现在各种方向，比如：频次降低，间隔变长，行为不稳定等等状态。

如果在乘客沉默之前就抓到这种状态，并且及时对乘客进行一定程度的“刺激”，是否可以改变乘客的行为轨迹，使其重新对平台产生依赖性？

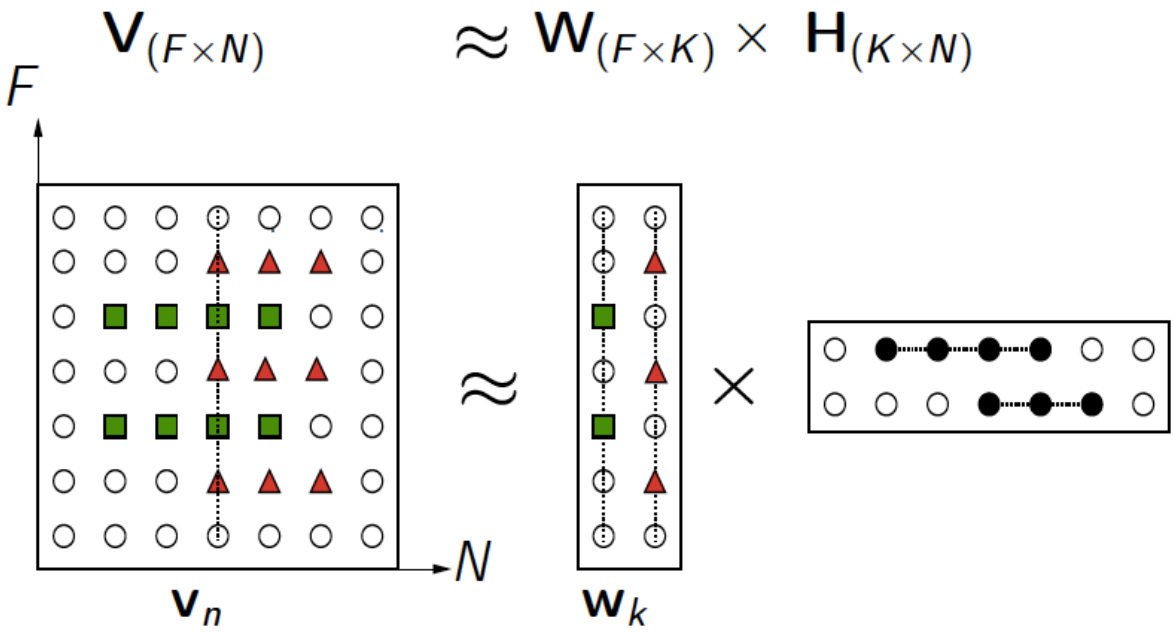
怀着这样的疑问，对乘客的行为状态进行探索性分层，使用非负矩阵分解的方法，对乘客行为进行了分层计算，期望抓住行为不稳定或行为突变的乘客。

下文将详细讲解非负矩阵分解的方法及该方法在本次探索中的应用情况。

NMF原理浅析

NMF全称为non-negative matrix factorization，中文名称为“非负矩阵分解”。

NMF属于一个无监督学习的算法，适用于稀疏矩阵分解限制条件是W和H中的所有元素都要大于0。



其中：

- V：F*N的原始矩阵，F为特征，N为观察值或特征向量，${V_n}$表示集合N中的第n个特征向量
- W：F*K的权重矩阵，K为分层个数，${\omega_k}$表示第k个元素的基向量
- H：K*N的扩展矩阵，${H_n}$表示列向量，${H_k}$表示行向量

对于NMF来说，相对难点在于K的设定上，适用于分类场景来看，K的定义倾向于选择要分几层。不同的K对于模型影响的情况如下：

- 从数据拟合角度来看，K越大数据拟合效果约好，分解矩阵的还原误差越小，但实际上随着K变大，误差将趋于平稳
- 从模型复杂度来看，K越小模型越简单，易于预测，迭代次数少

实际应用中，对于K的选择，有矩阵还原误差值可以进行辅助选择，但个人认为更多的应该针对于业务场景进行设定。

NMF实际应用

上面简单讲述了NMF的数据原理，接下来的部分将重点讲解一下，对于嘀嘀的乘客打车行为应用时，NMF代表的含义及应用效果。

NMF乘客留存分层模型

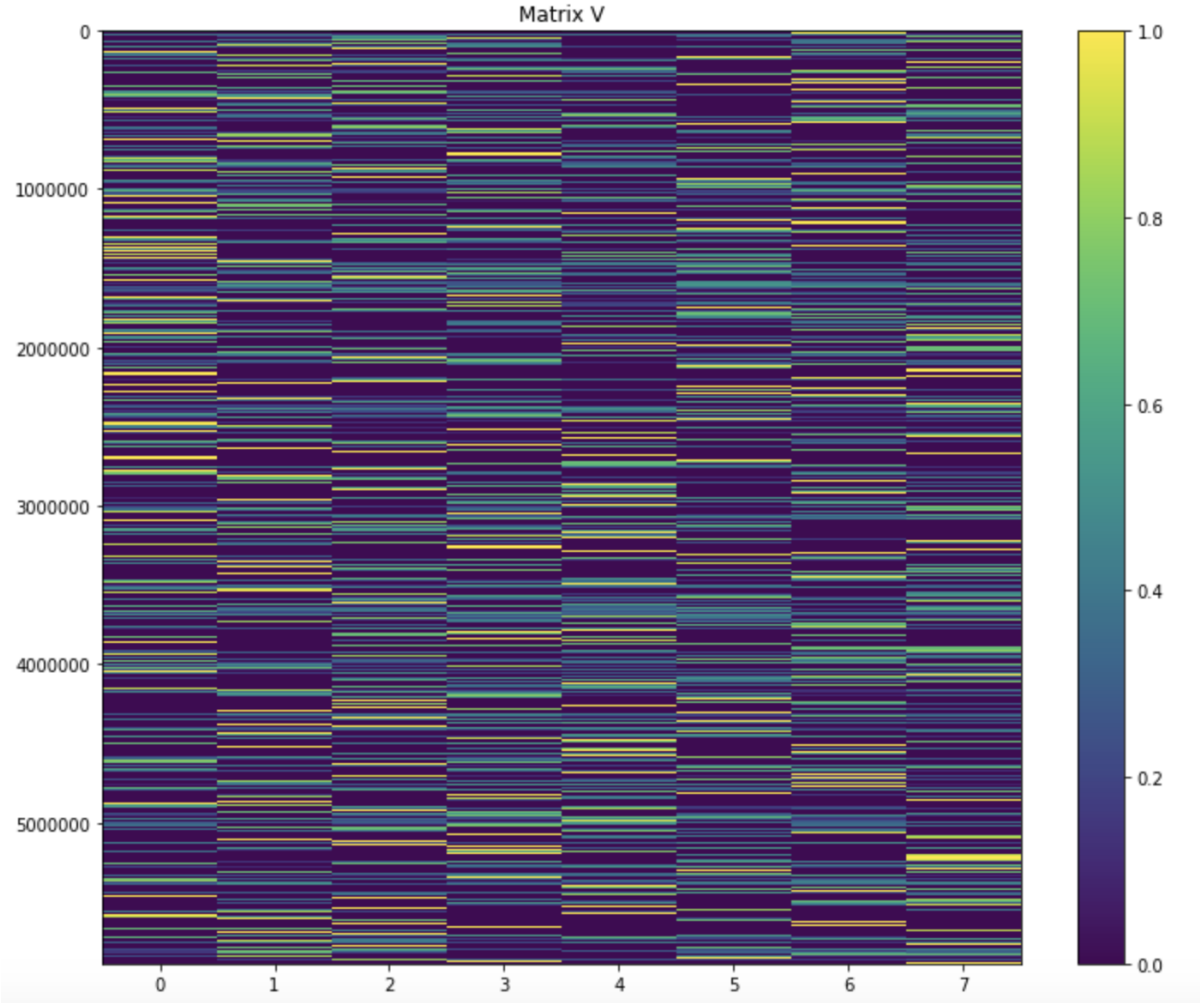
以乘客8周的打车频次分布作为训练数据，8周的打车频次分布作为预测数据。训练数据与预测数据存在月维度的重叠，按周维度更新迭代。

对于乘客的打车频次数据来看，存在极度稀疏且分布不均的情况，为避免分层计算时忽略掉低频乘客数据。

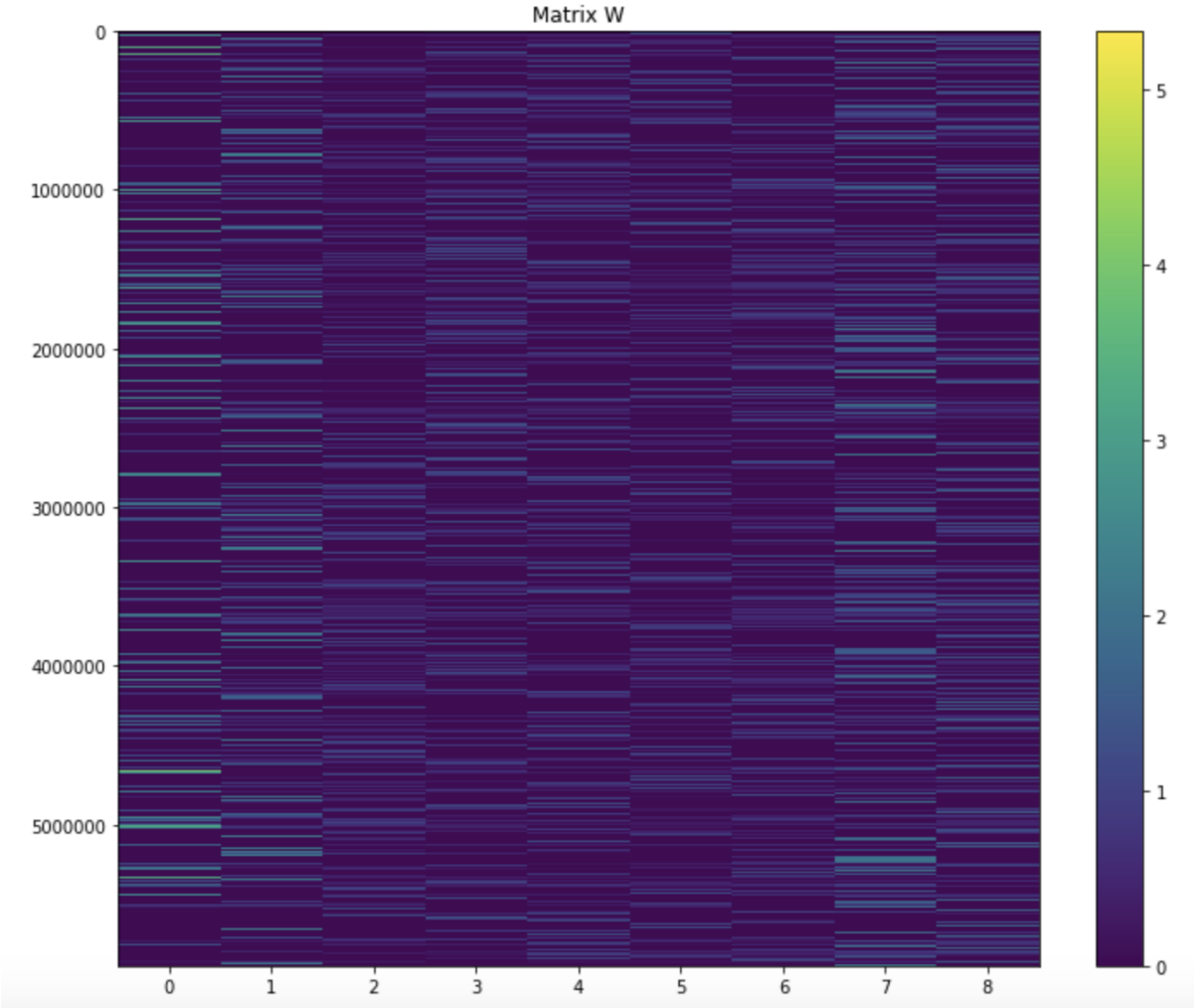
优先使用TF-IDF对矩阵进行调权，通过调权的方式调整乘客个体区分点的权重影响。

应用于乘客打车行为时，NMF矩阵如下：

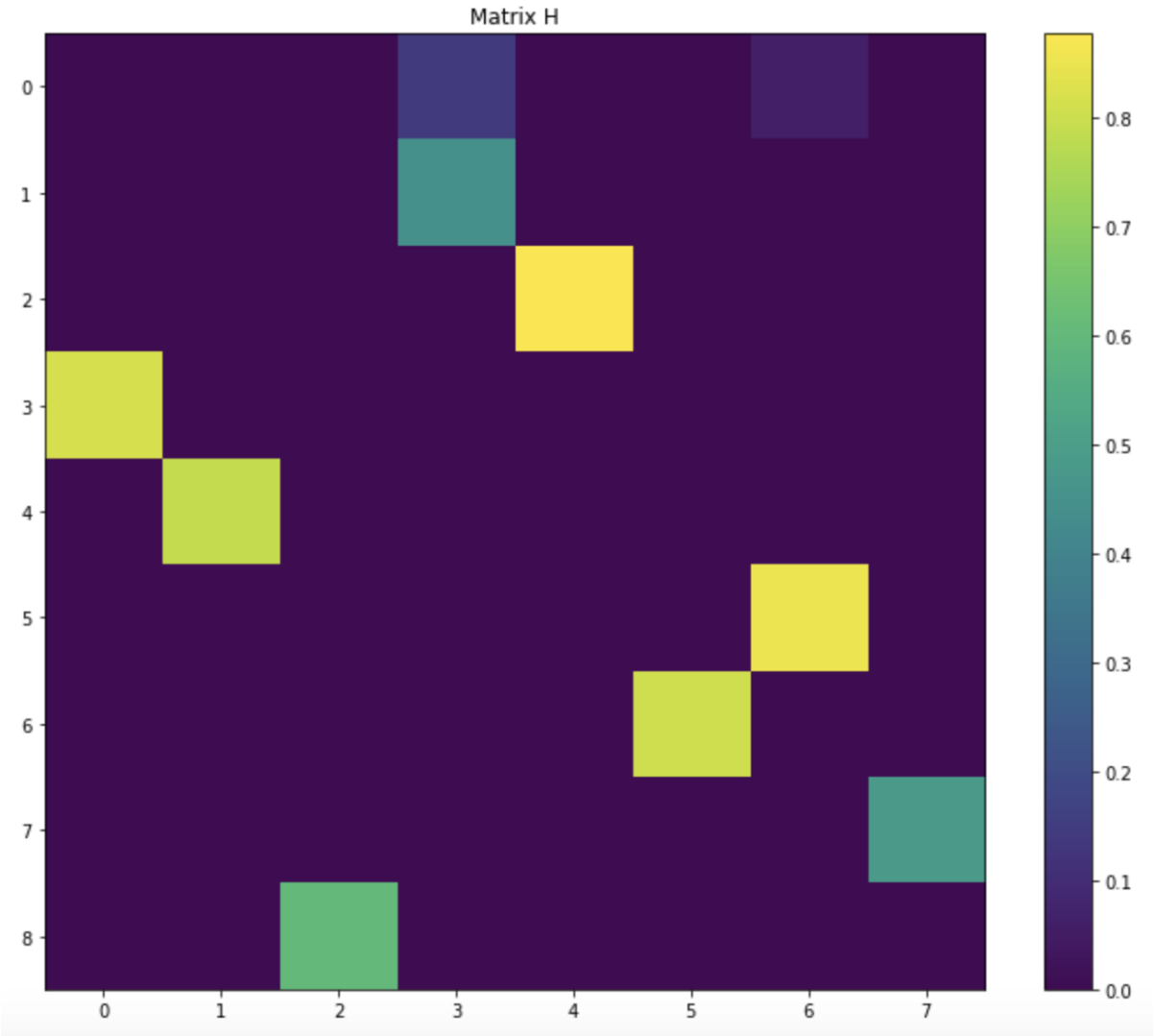
- V：原始矩阵，行表示目标乘客，列表示乘客周打车频次，数据为乘客8周的打车频次数据，N=8



- W：权重矩阵，行表示目标乘客，列为抽象出的乘客分层权重，分层为9时矩阵还原误差最小，K=9

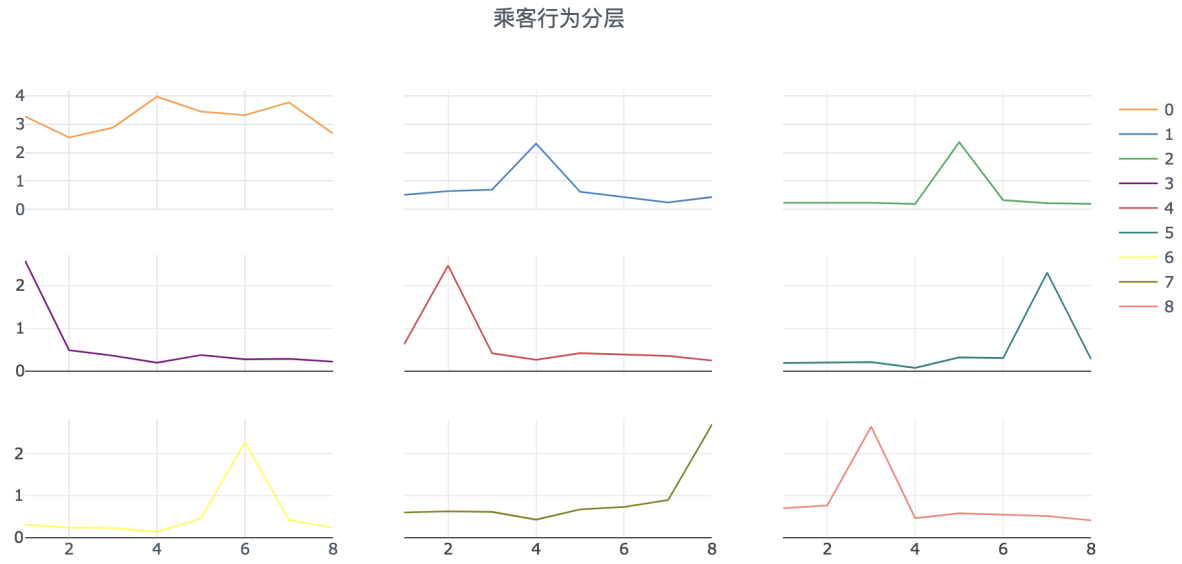


- H：降维矩阵，行表示每个分层特征值，列表示抽象出的某一分层的乘客的打车频次，H为9x8的矩阵



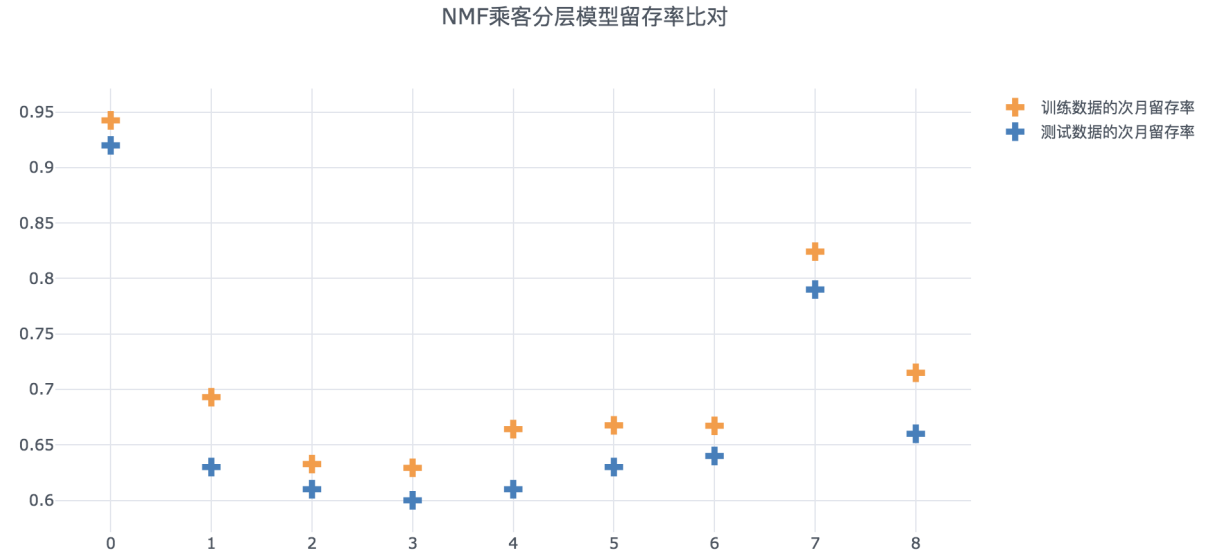
模型结果展示

将不同分层的乘客行为抽象出来，通过NMF分层之后可以得到的9类乘客的打车行为，如下图所示：

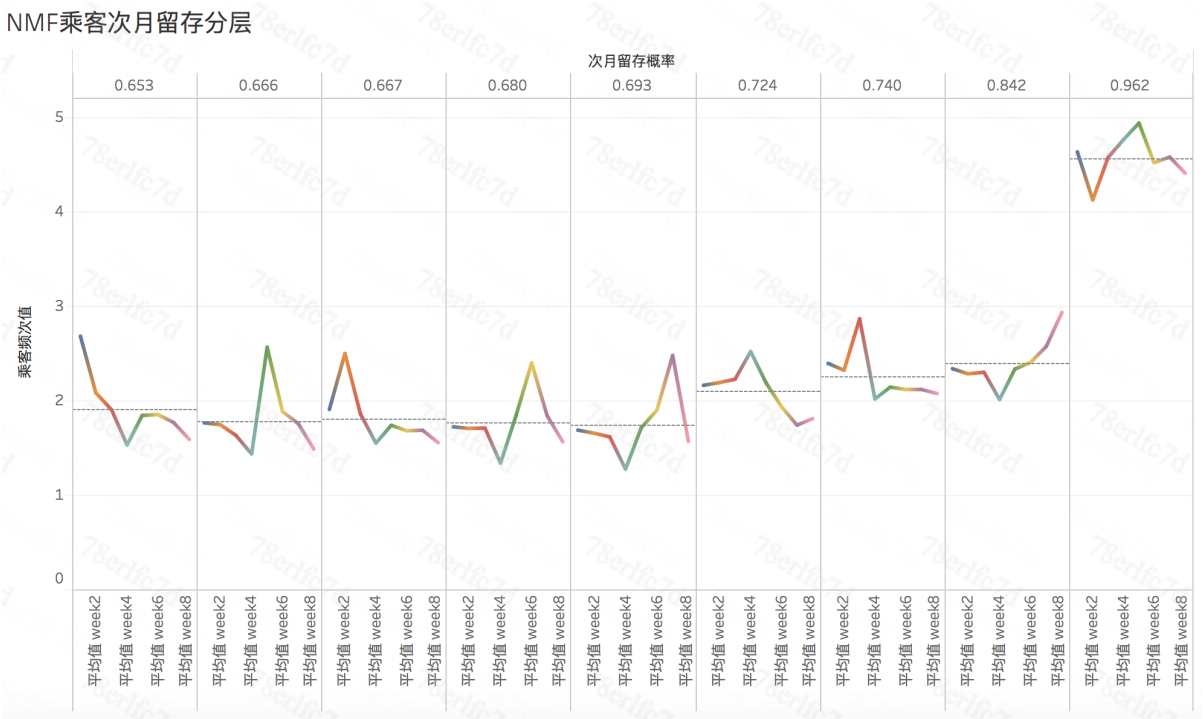


按照当前沉默口径，用训练数据之后一个月的数据计算每一层级乘客的次月留存概率。训练数据及预测数据的乘客实际留存率分布如下。

对比训练数据与测试数据的乘客流失概率来看，变动趋势基本一致。



针对每一个城市进行单独的模型训练，可得城市分层结果，下图为北京2018-07-15日更新的模型结果：



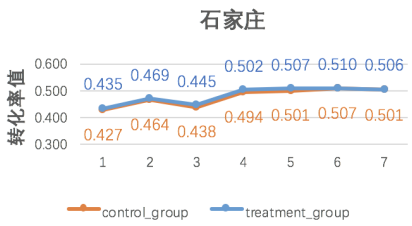
图中每一类别代表该层级乘客次月留存率的数值，8个数据点分别代表该层级周完成订单数 / 周完单乘客数，以层级为单位查看频次分布。
从图中各层级的波动趋势可以看到，稳定高频的乘客分层次月留存概率较好，打车行为极不稳定或前期频次较高后期下降的乘客层级次月留存率较差。

NMF乘客留存分层模型实验结果展示

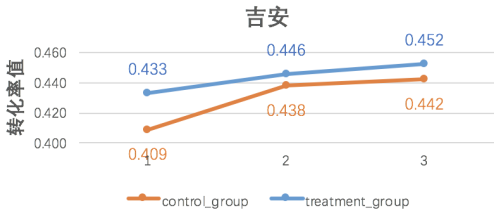
目前NMF乘客留存分层实验模型已经完成了一期实验，整体来看，GMV增幅为3.2%，△ROI为1.55。

模型预测的用户留存率与实验中对照组的留存率趋势一致，实验组完单率有明显提升。

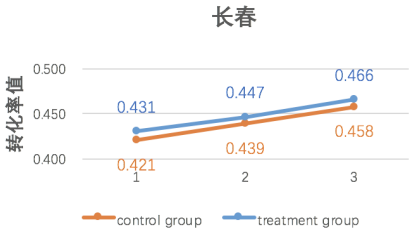
其中，石家庄、长春、吉安的实验结果如下：



△GMV	△ROI
3.2%	1.24



△GMV	△ROI
5.0%	2.56



△GMV	△ROI
4.6%	2.52

备注

NMF乘客留存率分层可视化结果可见tableau：NMF乘客分层结果tableau链接

NMF乘客留存率结果已上线特征平台的标签系统，按周维度产出，标签名称为passenger_remain_rate_nmf

NMF乘客一期实验wiki见：快车NMF乘客留存率分层实验wiki

如有需要查看tableau，可按照如下步骤申请权限：

- 点击链接申请权限：http://tableau.intra.xiaojukeji.com/#/views/_3/sheet0?:iid=1
- 选择全部-大数据策略数据可视化-乘客增长-NMF乘客次月留存分层，点击申请，申请人选yangkaidi
- 申请通过后，可点击上述链接查看模型可视化结果