分券模型的构建，训练及对比

I. 模型目的：

* 在出现同一日内多个客户对同一只标的提出预约，但可分配的标的量无法满足所有客户需求时（一般情况下，可分配标的量 < 单一客户的需求量），基于【地域】，【时间】，【费率】，【规模】等影响因子，确定最终获得标的的客户。

II. 数据：【历史锁券合约信息】

* 请将全部数据按照80-20 split为train data和test data

III. 模型

1. 连续决策树模型：按照【地域 - 时间 - 费率 - 规模】的顺序对同一日针对同一只标的的申请进行连续筛选

* 对同一日针对同一只标的的申请按照4个因子进行rank，每层决策树淘汰最底部一定比例的预约，直至整个决策树流程结束后保留一条预约
* Issue1：每层决策树淘汰的预约比例如何确定？
* Solution: 按照经验，很少出现同一日针对同一只标的的申请超过4条的情况，所以可以每一层决策树淘汰最底部1条预约。如在决策树行进到最末层之前出现只剩1条预约的情况，则连续决策树提前终止。
* Issue2: 如何基于4个因子对预约进行rank？
* Solution: 4个因子中，对时间/费率进行简单rank即可。规模参考【weekly summary】。客户的地域划分和地域优先等级请自行打上华北，华南，华中tag，并按照华南>华北>华中划分优先级。

1. Regression model

* Dependent variable: 发起预约的客户是否最终获得了标的
* 构建一个binary variable来标示每一个提出预约的客户最终是否得到了券。0代表未获得，1代表获得。
* 数据获取：【历史锁券合约信息】-【订单状态】为【已锁券】的合约中的客户，即为最终获得标的的客户
* Independent variables:

A. 预约时间：

* Discrete numeric variable
* 需要对于在同一天内，对于同一只标的的预约时间进行标准化处理
* 考虑到申请时间对于此模型的影响仅为时间优先原则，可以将申请时间rank后打上分类标签即可。但由于每日对于每只券的预约总数量不同，把时间看成categorical variable没有意义，所以不如将此看成discrete numeric variable来惩罚该客户在竞争客户中的迟来程度：即数值因子的值代表【在此因子上优于该客户的竞争者数+1】如时间因子数值越小，则表示该客户预约在竞争者中越早，反之数值越大，则该客户预约在竞争者中越晚。
* 获取方法：通过【证券代码】和【有效期起始日】作为ticker筛选出针对同一只标的的一批预约。将【申请时间】column转化为yyyy/mm/dd/hh/mm/ss格式后进行倒序rank

B 费率：

* Discrete numeric variable
* 基于人工分券时只考虑费率高低排名而非具体数值，该因子构建逻辑同【预约时间】因子。 由于我们此时用排名数值奖励高费率预约，故对于费率进行【rank】，即预约费率在竞争者预约中排名越高，费率因子数值越小。
* 获取方法：参考【历史锁券合约信息】-【最高费率】

C 地域：

* Multi-categorical variable
* 不同的客户对应不同的地域。
* 简单打标签即可

D 规模：

* Discrete numeric variable
* 参考weekly summary，取得数值后，逻辑同因子B费率
* 在费率相同的情况下，偏向给规模大的客户

对于regression models的备注：

* 训练完baseline regression model后，再分别拟合Lasso & Ridge regression对loadings和因子数量进行regularization
* 已知4个因子重要程度为：地域 > 时间 > 费率 > 规模，其中规模在人工审批过程中仅作为辅助规则（在费率相同的情况下，分券给规模较大的客户），所以为了实现最优模型表现，尝试将规模因子刨除后再分别拟合baseline logistic regression，lasso和ridge regression
* 训练完上述6个regression model后，对比out of sample performance，选择表现最佳的模型

3. Classification model

* 在regression analysis中构建完毕所有因子和dependent variable后，分别基于构造好的数据训练KNN, Decision Tree, SVM和Naive Bayes 4 个classification model
* 其中knn如果要应用于classification model，需要构建比较合理的distance function，可以参考以下链接：<https://stackoverflow.com/questions/50335203/how-to-apply-knn-on-a-mixed-datasetnumerical-categorical-after-doing-one-hot> 里面有sample code
* （实在难的话就不要做knn了）
* 刨除规模因子，再次分别拟合模型
* 训练完上述8个classification model后，对比out of sample performance，选择表现最佳的模型

IV 模型选择和fine tune

* 训练完上述所有模型后，将每个模型的Accuracy, MSE和adjusted r-squared汇总在一张表单中
* 选择3个表现最佳的模型进行更细致的调参（如果算法里有parameter可以fine tune的话）
* fine tune后再次测试Accuracy, MSE和adjusted r-squared，选择表现最佳的模型

备注：

1. 整体代码用Python写，模型训练算法尽量用sklearn
2. 遇到模型/数据/逻辑上的问题参考<https://stackexchange.com/> ，coding问题参考<https://stackoverflow.com/> ，或者直接来问我