存活分析报告

一．数据

原始数据为融资租赁还款数据（上海+重庆），数据量12512x78，以合同号为主键，分别在上海与重庆的融租数据中提取出的。其大抵记录了在一个合同中的某个主体（公司）租借资产在每一期的还款情况。其中“0”代表还款，“1”代表未还款，逾期，“2”代表还款还未发生，还没到还款日期。

二．为什么需要做存活分析？

在这种情况下，公司需要知道，对于一个金融产品而言，我们在每一期的时间点，客户的违约概率（不还钱的概率）有多大。以此进一步根据EL=PD\*LGD\*EAD，PD为客户的违约概率，LGD为违约损失率，EAD为风险敞口。由此便可以推算出我们的在每一期的预期损失（EL）有多大。从而，提早进行坏账拨备。在国内，银监对商业银行的拨备计提率要求是150%覆盖，体现了对资本的审慎性监管要求。

三．方法

存活分析顾名思义，原本用作医学上统计患者在患某种病后每个时间周期点（可能是天，周，月）的存活的可能性。亦或者可以用作分析是否某种因素对于患者的存活可能性造成影响，例如：对于癌症患者来讲，患有高血糖，是否对于其存活率有影响。

如（二）中所述，当前我们需要进行的是得到每一期的时间点，客户的违约概率。在进行存活分析时，会使用到的方法有以下几个：

Kaplan-Meier plots  --------------------------------用来画存活曲线的图

Nelson-Aalen plots ----------------------------------用来画死亡曲线的图

Log-Rank test ----------------------------------------用来对不同主体的存活曲线进行比较

Cox-proportional hazards regression ------------用来探索不同特征对于存活率的影响

在进行存活分析时，会涉及到两个function，一个叫survival function，其定义了在某一个时间点，一个人（或者一个事件）存活的概率，例如：S(10) = 0.3 意味着在第10天的时候，该subject存活的概率为30%；另一个叫hazard function, 与survival function相反，其定义了在某一个时间点一个人（或者一个事件）死亡的概率。

在实际操作中，用来预测/模拟 survival function 表现的方法叫做Kaplan-Meier Estimator，

用作计算存活概率的公式如下



其中 S(ti) 代表在ti时间，该subject存活的概率；ni代表在到达ti时间点前subjects存活的数量；di代表在时间点ti死亡的数量。

更笼统地可以表达为：



利用python实现时，需要调用lifelines中的Kaplan-MeierFitter



之后按照kmf 的argument，让该function适配数据。共两个argument，第一个为时间的序列，第二个为观察事件。



在适配好数据之后，可以利用kmf该包自带的一些method来进行分析，如event\_table，



生成的event\_table 如下，在输出的table中，以event\_at为index，和5个columns. removed 代表了在当前时间点，不再算作为该实验一部分的subjects的数目，即死亡的subjects加上censored subjects.

observed 代表了在当前时间点，观察到的兴趣事件的次数，在这次project中，其代表了发生违约的次数。

censored 是一个医学概念，一般指患者长时间未出现兴趣事件(死亡，病变等)，另一方面可能说明该患者已经痊愈等，这样的情况下，患者会被计在该列中。

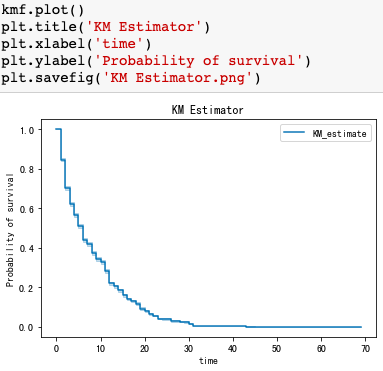
entrance 代表了在当前时间点，新进入的subjects，可以理解为新的确诊的患者。

at\_risk 代表了在当前时间点，正在接受观察的subjects的数目。

Table

Description automatically generated

在得到了table后，我们就可以得到大致的关于subjects随着时间的变化情况。但为了更具体且直观的看出存活率随时间的变化，我们可以使用如下Plot出其中的关系。



另外，非常重要的是，我们可以使用以下method预测非连续时间的subjects的存活率。



对于hazard function其使用方法完全一样，唯一不同的地方在于，调用的包的名称不同。



值得注意的是，与医学上的概念不同的是，融租项目中客户违约不同于死亡。客户可能出现例如，第3期违约但第4期不违约的情况，但利用以上的survival function 和 hazard function 计算出的概率都是默认只要发生兴趣事件（死亡），那么就直接将该subject移出实验。所以，两个function预测的概率并不能直接实用用于计算融租中客户的违约概率。需要对得到的概率进行处理。

为了计算出融租项目每一期真实的当期的违约概率，在计算时，我们可以将第n期发生违约与前n-1期均不发生违约看作是两个独立事件。通过以下公式计算出每一期的真实的当期PD，PD代表probability of default, PS 代表 probability of survival

A picture containing text

Description automatically generated

四．结果

理论上来讲，其真实的PD违约概率，应该随着期数的增加逐渐变大，但由于融租项目目前大部分都还未结束，很多项目都处于开始阶段，导致每期的PD随着时间增长越来越小。后续随着时间越来越久，可以再对此进行更新。其结果在下表中体现出来。

Code 我上传到 我的github：

https://github.com/JackieSun66/Glp\_Job/blob/main/%E8%9E%8D%E7%A7%9F%E7%94%9F%E5%AD%98%E5%88%86%E6%9E%90.ipynb

