– **Unlimited Distribution** –

How to choose bond

junlin.cai
Datayes, Ltd

2014 年 3 月 19 日

1 简介

在 MBS 资产池的研究中，关于债券怎么选择无疑是一个很大的学问，值得我们去深入探讨。`Datayes` 固定收益团队通过对资产池中各债券各方面属性的研究，试图寻找出债券评级或者其违约可能与其其余属性和相关经济变量之间的关系。

2 研究方法

我们选择作为研究的变量如下：总期数、已还期数、截至基准日剩余期数、基准利率、贷款目前利率、贷款初始利率、贷款现有余额、本息偿还方式、已偿还本金总额、性别、年龄、婚姻状况、职业、年收入、贷款剩余期限与年龄之和、逾期利率等，所研究的对象为五级分类。

为了研究的方便，对定性性质的变量我们均做了生成 dummy 变量的处理，dummy 的数量比该变量的值的总数少一，例如性别有男女两种，则我们只需生成一个 dummy, 是男则取 0，不是则为 1，以此类推。五级分类中，分为正常、次级等等，我们取正常为 1，其余为 0，数据缺少的则删除该条变量，累计删除数量为 127，剩余有效变量数 153035。

2.1 主要研究工具

模型的研究中,我们主要采用的是 `Python` 的一个机器学习的库,叫做 `scikit-learn`。其中核心的是一个 `RFE` 函数，（Recursive feature elimination）。这个函数的主要作用

是通过循环递归的逐渐在原有变量的更小的子集里寻找对所研究对象更有解释能力的变量，直到达到所需求的水平。通过这个函数，我们可以在债券的众多属性里，挑选出和债券评级有很强关联性的一些变量。以后，我们则可以通过对这些变量的研究，挑选出合适的债券来构成我们所需要设定的资产池，如风险收益高的一组和风险收益低的一组。

2.2 研究步骤

在模型的选择中，考虑到五级分类作为被研究对象取值只有 0 和 1 两种，我们选取 `LogisticRegression` 来研究。结合该模型，我们通过 `RFE` 挑选出 14 个适合的变量（这里包括 dummy），然后再考核所选出的参数的系数和 *score*，来判定该参数对研究对象的影响和其显著性。考虑到截至基准日剩余期数、总期数和已还期数之间存在线性相关，我们需要删除截至基准日剩余期数。

- 挑选出能够定量分析或者能够生成 dummy 变量的，并且有可能对所研究对象产生影响的变量。然后对适当的变量生成 dummy，记得模型中加入的 dummy 要比变量分类始终少一个以防止产生共线性；
- 读入数据，转变数据格式，并对缺失的数据进行一定处理。在这里的研究中，我们把所有缺失的数据都填充了 0。在 dummy 变量中，这对我们结论的影响比较小。其他可能产生影响的变量的缺失数据则相对比较少。
- 是用 `RFE` 方法进行变量的筛选，选出对结果有显著影响的一些变量，并考察其参数和显著性。然后根据结果进行一定的判断，并核实这样的结果是不是符合现实中的逻辑。

3 模型结论

最终的研究结果显示，贷款目前利率、7 个本息偿还方式的 dummy、性别（女 dummy）、婚姻状态（已婚 dummy）、职业（中小企业 dummy）、职业（金融企业 dummy）、历史信用记录等因素对债券的五级分类有较大相关性。也就是一共有 8 个变量对我们关心的五级分类有较为显著地影响。以下除了带星号的是偿还方式的 dummy 的系数外，其余依次是贷款目前利率、性别（女 dummy）、婚姻状态（已婚 dummy）、职业（中小企业 dummy）、职业（金融企业 dummy）、历史信用记录。[[-2.62679051 -0.3020649* 0.44834436* -0.69189383* -6.35614492* 0.62225526* -1.02892788* 2.61163847* 3.17296065* 0.6532721 0.45535409 -0.53939601 0.37965081 -0.37759654]] 我们得到结论如下：

- 贷款目前利率越高，信用评价越差，系数为 −2.627;
- 女性的贷款，信用普遍比男士好，系数为 0.653;
- 婚姻状况对一个人贷款的信用评级由较大影响，系数为 0.455，已婚的人贷款的信用比较好;
- 职业对一个人贷款的信用评级有显著影响，中小企业员工信用偏差，系数为 −0.539，金融企业员工信用较好，系数为 0.380。

4 模型缺陷与拓展

目前的模型只是在众多变量中挑选出一定数量的因变量，这些因变量能够较好的解释所研究对象五级分类的值。这有助于我们在接下来的研究中知道哪些变量对我们关注的结果有一定影响。但是由于解释性的变量含有很多定性变量所生成的虚拟变量，对于我们定量的研究有一定的影响。再者，线性模型本身就存在局限性，它缺乏足够的经济金融理论作为基础。譬如性别对债券评级分类的影响，具体是多少，怎么衡量等等。