统计学习大作业

一.研究内容

根据与Repayment有关的相关特征，试图构建每一间学校的学生贷款的违约指数，利用该指数来衡量每一间学校的学生违约严重性。同时，在不使用earning有关字段的前提下，尝试从统计学习模型来实现对构建指数的预测，达到识别存在高潜在违约风险的学校。为方便后文叙述，本文所有指高潜在违约风险的学校即为该学校办理贷款的学生具有较高的违约可能性，潜在违约风险即为多数办理贷款的学生违约的风险。

二.特征的选择

2.1违约指数相关特征

2.1.1CDR

CDR全称Cohort default rates，指的是队列违约率。其中The three-year cohort default rate (CDR3)和The two-year cohort default rate (CDR2)是已有的两个指标，分别表示学校学生在近三年内违约的比率和学校学生在近两年内的违约比率。

2.1.2BBRR

BBRR全称Borrower-Based Repayment Rate，指的是基于借款人的还款比率，在特征中以BBRR[YR]_[LOAN]_[GROUP]_[STATUS]的形式出现，这里YR是指进入还款状态后的年数，有一年和两年两个情况；LOAN是指贷款类型，有与个人相关的 Parent PLUS 贷款 [LOAN]=PP和联邦直接贷款 [LOAN]=FED；GROUP是指学生学习的状态，分别有 UG 本科毕业、UGCOMP 本科即将毕业、UGNOCOMP 本科肄业、UGUNK 本科状态位置、GR 研究生毕业、GRCOMP 研究生即将毕业、GRNOCOMP 研究生肄业；STATUS是指借款人的状态，分别有：

1. 违约状态DFLT 借款人有一笔贷款超过 360 天未能按照期票中的规定付款
2. 拖欠状态DLNQ 借款人有一笔贷款在 31 到 360 天之间未能按照期票中的规定付款且没有一笔违约
3. 延期状态FBR 借款人暂时停止或减少每月还款，且没有违约或者拖欠
4. 宽容延期状态DFR 借款人暂时延期支付贷款，在此期间，补贴贷款通常不产生利息
5. 无进展状态NOPROG 借款人按期还款，但是所有贷款余额总和超过原贷款余额总和，且前几类均不适用
6. 取得进展状态MAKEPROG 借款人按期还款，所有贷款余额总和小于原贷款余额总和，且不适用先前类别
7. 足额支付状态PAIDINFULL 借款人按期还款，所有贷款均已足额支付。
8. 取消状态DISCHARGE 借款人因死亡、残疾、破产、欺诈或身份盗窃等原因偿还义务已被取消。

2.1.3DBRR

DBRR全称Dollar-Based Repayment Rate，是关于学生还款速度状况的指标群，在特征中以DBRR[YR]_[LOAN]_[GROUP]_[METRIC]的形式出现，这里的YR指的是统计年份，有1，4，5，10，20

五个统计年份；LOAN是指贷款类型和GROUP的定义与BBRR中的相一致；METRIC是具体的指标，分别有：

1. 未偿还余额NUM
2. 最初贷款金额DEN
3. 未偿还余额与最初贷款金额的比例RT
4. 在两年内任何时候进入还款的借款人数量N

2.1.4DEBT

DEBT即债务相关指标，特征中主要有[TYPE]_DEBT_[METRIC]和CUML_DEBT_[METRIC]两种形式，其中CUML_DEBT表债务累积分位数的数值，有债务人统计人数，和10，25，75，90四个分位数共五个特征；而[TYPE]_DEBT_[METRIC]中[METRIC]有N人数和MAD债务中位数两个指标，TYPE表示学生的类别，分别为：

1. 家族第一代大学生FIRSTGEN
2. 非家族第一代大学生NONFIRSTGEN
3. 独立于家庭的学生IND
4. 依托于家庭的学生DEP
5. 女性学生FEMALE
6. 男性学生MALE
7. 家庭高收入（家庭年收入7.5w美元以上）学生HI_INC
8. 家庭中等收入（家庭年收入3w-7.5w美元）学生MD_INC
9. 家庭低收入（家庭年收入3w美元以下）学生LO_INC
10. 已毕业学生GRAD
11. 退学的学生WDRAW
12. 获得PELL奖学金的学生PELL
13. 未获得PELL奖学金的学生NOPELL
14. 全体学生

2.1.5RPY

RPY即repayment，是还款相关指标，在特征中以[TYPE]_RPY_[YR]_[METRIC]形式出现，其中YR为统计年份，共有1，3，5，7四个统计年份；METRIC则为N和RT，分别代表还款人数和还款人比例；TYPE表示学生类型，和DEBT的TYPE类型基本一致（GRAD用COMPL表示，WDRAW用NOCOM表示），这里便不多赘述。

2.2违约指数特征构建

2.2.1缺失值的处理

由于数据本身具有不同程度上的缺失，为不影响数据本身的无偏性，对本处的缺失值均采取均值代替的方式，对于全部缺少的特征，则采取全为0的方式代替。

2.2.2CDR_index

正如前文所述，CDR指标非常单一，考虑CDR的实际含义，以如下方式定义：

$$CDR\_index = \frac{CDR2 + CDR3}{2}$$

2.2.3BBRR_index

BBRR的四个变量中，因为是构建学校的违约指数，因此本文只考虑联邦直接贷款FED类型，而不考虑个人贷款PP类型。同时只考虑本科/研究生中已毕业和即将毕业的学生的数据，减少由学生是否毕业带来的还款状态偏差。由于重点关注违约，因此只考虑八个状态中的前六个状态，并赋予不同的权重如下表所示：

STATUS	DFLT	DLNQ	FBR	DFR	NOPROG	MAKEPROG
weight	1	08	0.6	0.3	0.2	0.1

权重的数值代表着该STATUS的违约恶劣程度，同时考虑到不同统计年份的可比性不同，年份越长情况下同等违约状态理应更加严重，因此给出了一个基于年份、数量、权重的加权指标BBRR_index，其定义如下：

$$BBRR\_index = \sum_{group} \frac{\sum_{year} \sum_{w_i} w_i \cdot {}^{year}_{w_i} BBRR^{group}_{rate} \cdot {}^{year} BBRR^{group}_{N} \cdot year}{12 \cdot \sum_{year} {}^{year} BBRR^{group}_{N} \cdot year}$$

其中，${}^{year}_{w_i} BBRR^{group}_{rate}$表示第i个STATUS下，指定year和group的还款人比例，${}^{year} BBRR^{group}_{N}$表示指定year和group的还款人总数。通过这样构建，性质恶劣的还款比率数值会被放大。至于[group]，则是采用等权重平均的方式。

2.2.4DBRR_index

DBRR的四个变量中，和BBRR一样，[LOAN]只考虑联邦直接贷款FED类型，[GROUP]只考虑本科/研究生中已毕业和即将毕业的学生的数据。首先可以明确的是，违约恶劣程度与RT数值成正比，且相同RT数值下，年份越久远性质越恶劣，还款人相对越少，性质越恶劣。但是因为不知道借款人总数，因此这个进入还款的借款人数量N没有特别大的参考价值，但是同所学校的不同[GROUP]的进入还款的借款人数量N是已知的，因此考虑用进入还款的借款人数量N和年份来对未偿还余额与最初贷款金额的比例RT进行加权，构建DBRR_index指标，其公式如下所示：

$$DBRR\_index = \sum_{group} \frac{\sum_{year} {}^{year} DBRR^{group}_{rate} \cdot {}^{year} DBRR^{group}_{N} \cdot year}{4 \sum_{year} {}^{year} DBRR^{group}_{N} \cdot year}$$

其中，${}^{year} DBRR^{group}_{N}$表示指定year和group下的两年内进入还款的借款人数量，${}^{year} DBRR^{group}_{rate}$表示指定year和group下的借款人未偿还余额与最初贷款金额的比例。通过这样构建指标，可以放大年份和人数在最终指标计算上的影响，人数大的相对权重更大，年份久的，说明性质越恶劣，比重越大。至于[group]，则是采用等权重平均的方式。

2.2.5DEBT_index

DEBT的变量中，为了更纯粹的表示学校学生的债务情况，这里用[TYPE]_DEBT_[METRIC]来构建DEBT_index指标，其中[TYPE]只采用全体学生、已毕业、高收入家庭、中等收入家庭、低收入家庭五种类型的学生，并赋予不同的权重，具体如下所示：

type	全体	GRAD	HI_INC	MD_INC	LO_INC
weight	1	2	1.5	1	0.5

通过赋予毕业生高权重，可以更好的反映该学校毕业生的债务情况，同时给不同家庭收入的学生不同的权重，以达到减少不同家庭收入对债务数额带来的影响，最终通过人数加权的形式得到DEBT_index如下：

$$DEBT\_index = \frac{\sum_{w_i} w_i \cdot DEBT_{MDN}^{w_i} \cdot DEBT_N^{w_i}}{\sum_{w_i} w_i \cdot DEBT_N^{w_i}}$$

其中，$DEBT_N^{w_i}$表示w_i权重的[type]类型学生的数量，$DEBT_{MAD}^{w_i}$表示w_i权重的[type]类型学生的债务中位数。通过这样构建指标，可以放大毕业生的真实债务水平，缩小不同家庭收入的学生对债务指标的影响，同时根据人数加权的办法切实反应大体的债务情况。

2.2.6RPY_index

RPY_index的构建思想和前面类似，也是通过学生类型权重，学生人数，年份加权的形式得到。不同的是，相同的还款人比例下，年份越长，学校学生还款速度越慢，因此构建指数时采用和之前不同的处理方式，年份加权的方式由乘法改为除法；同时其他条件相同的情况下，学生家庭收入越多，说明债务越重。其中[TYPE]只采用全体学生、已毕业、高收入家庭、中等收入家庭、低收入家庭五种类型的学生，并赋予不同的权重，具体如下所示：

type	全体	COMPL	HI_INC	MD_INC	LO_INC
weight	1	2	0.5	1	1.5

这样在放大完成学业学生的还款率的基础上，减少了家庭收入给还款率带来的影响。RPY_index定义如下：

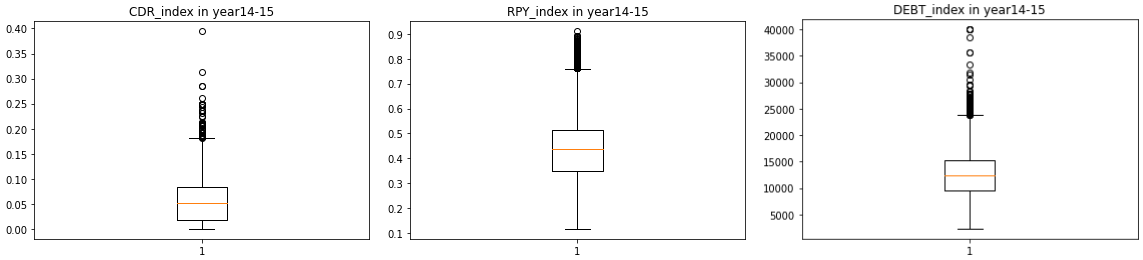
$$RPY\_index = \frac{\sum_{w_i} \sum_{year} w_i \quad {}^{year}RPY_N^{w_i} \cdot {}^{year}RPY_{rate}^{w_i} \div year}{6 \sum_{year} \quad {}^{year}RPY_N^{w_i} \div year}$$

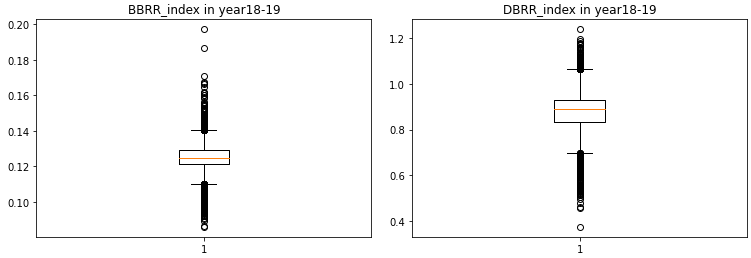
其中${}^{year}RPY_N^{w_i}$表示指定year内w_i权重的[type]类型进入还款的学生的数量，${}^{year}RPY_{rate}^{w_i}$表示指定year内$w_i$权重的[type]类型进入还款的学生占比。通过这样构建指标，能缩小久远年份的占比，放大毕业生的占比，缩小家庭收入因素带来的影响。

2.3违约指数特征选择

2.3.1各指数特征的可视化表示

本文尝试用14年–19年的数据对2.2中得到各个指数绘制箱线图，得到的部分结果如下：





根据箱线图的结果得出以下结论：

- CDR_index，DEBT_index，DBRR_index，BBRR_index，RPY_index五个指标反映良好，且能捕捉到许多过高异常值，意味着部分学校学生违约严重程度极大。
- DBRR_index和BBRR_index指标能捕捉到许多过低的异常值，意味着通过该指标能找到一些潜在的学校学生违约程度低的学校。
- 不同年份指标均存在不同程度上的缺失，具体表现为缺失1~2个指标，例如18-19年没有RPY_index的数据，17-18年没有DBRR_index和RPY_index的数据，14-15年没有DBRR_index和BBRR_index的数据

2.3.2违约指数的构建

尽管指标的缺失很让人感觉无奈，但是通过同年份的数据研究发现三个现象：一是同年份的DBRR_index和BBRR_index的相关系数接近于1；二是RPY_index和DEBT_index的相关系数接近于1，本文的理解是，高额的债务值导致了较高的还款人比例；三是所有的指标均呈现正相关。为尽可能用到已有的数据，考虑到违约程度是一个绝对指标，本文定义*_index其中之一大于指定阈值则为潜在高违约风险的学校，反之则为违约风险不显著的学校。各个指标的阈值由箱线图主观给出，并如下表所示：

index_name	RPY_index	BBRR_index	DEBT_index	DBRR_index	CDR_index
threshold	0.6	0.14	25000	1	0.1

由此法构建指标正负样本比约为0.9负样本：0.1正样本，最终指标由以下公式给出：

$$Final\_index = \begin{cases} 1 & ,*\_index > threshold \\ 0 & ,else \end{cases}$$

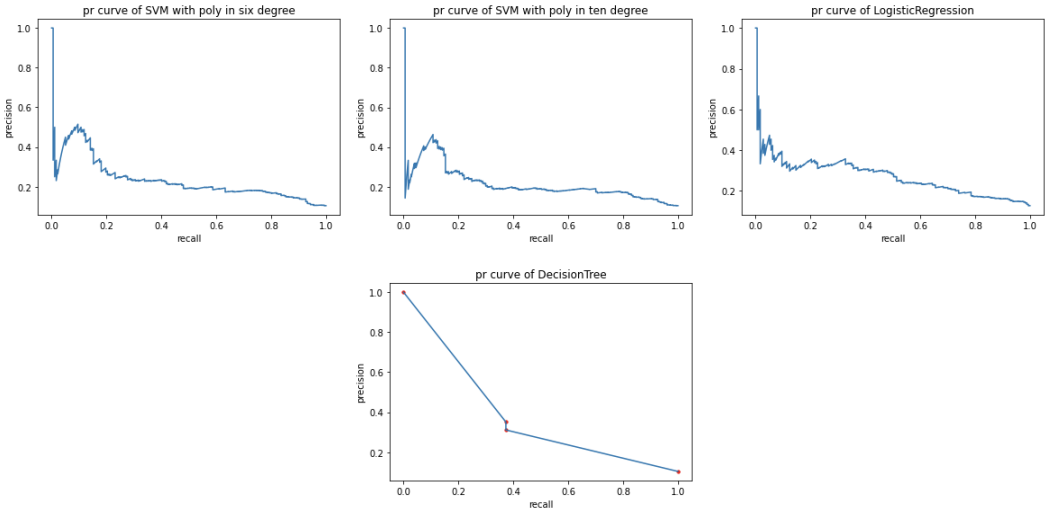
2.3.3因变量的选择

根据特征选择了一些因变量，这里不一一赘述。因变量主要有类别型和数值型，在预处理的过程中，数值型的特征将进行标准化处理，而类别型数据则不额外进行处理。最终有194个数值型特征和6个类别型特征恰好共200个特征构成模型因变量。

三.模型建立

3.1逻辑回归，支持向量机与决策树

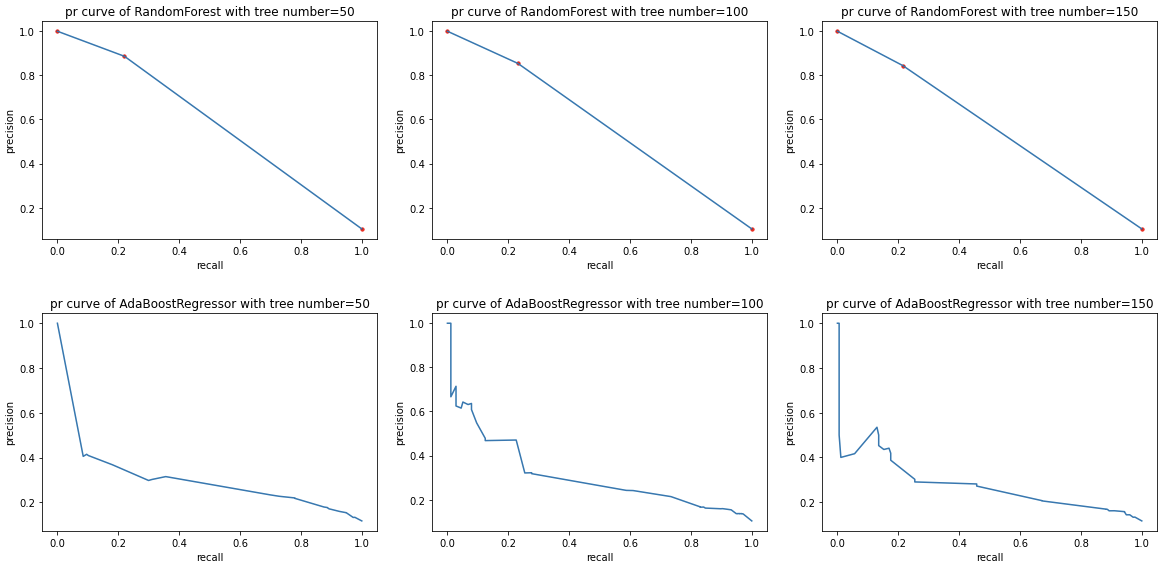
统计学习，传统的方法必然得用一用，得到如下PR curve：



可以看出，模型的效果非常不好：svm算法出现了一个驼峰，意味着有很多负样本被预测为正例，这有可能是样本标签不合理，不过峰值处的数值与逻辑回归和决策树一样不理想，这意味着模型的表达能力不行，这也是意料之中的事情。

3.2集成学习

尝试一下Adaboost和randomforest（基学习器为CART决策树），得到如下PR curve：

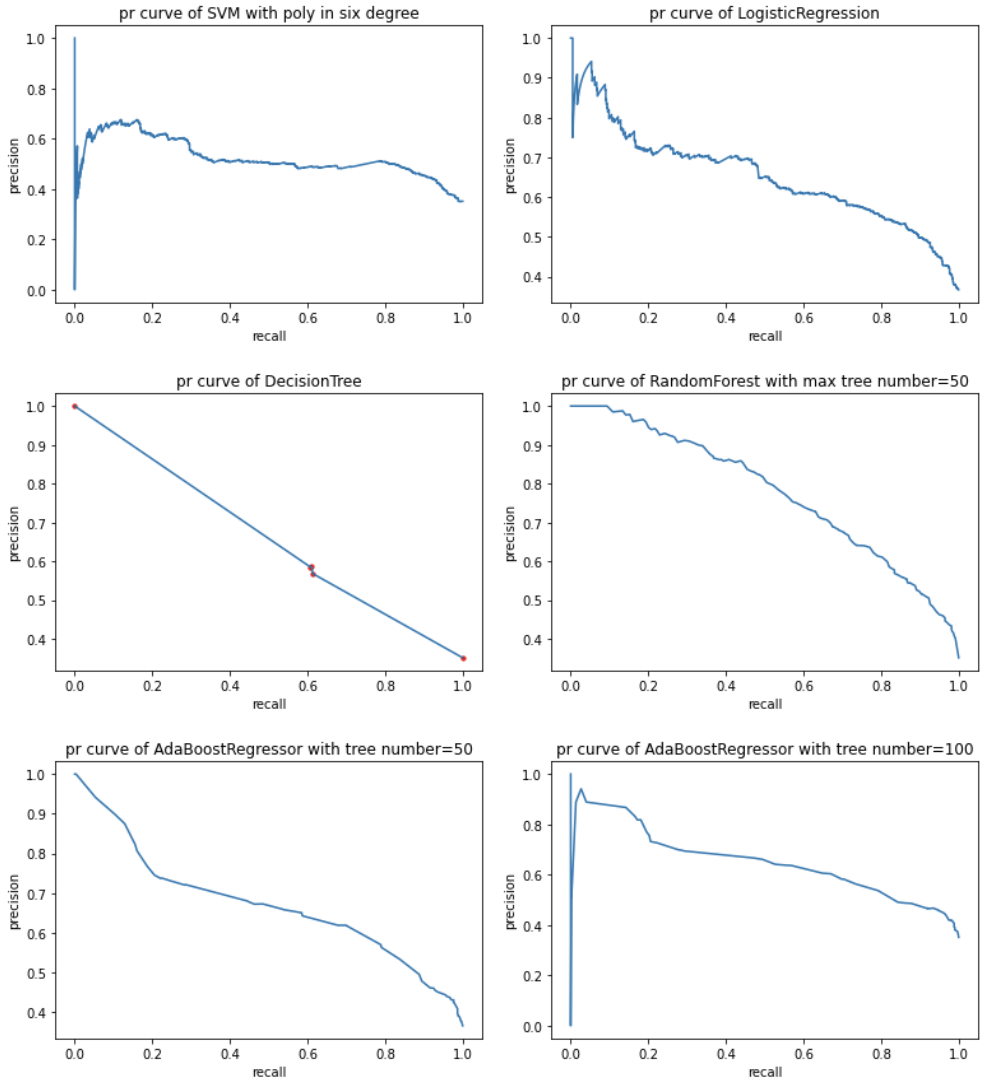


可以看出，模型的效果依旧不好，randomforest尽管看起来不错，但是由于研究的目的是识别有潜在违约风险的学校，0.2出头的召回率不能说是一个好的结果，基于模型PR曲线出现驼峰的现象，本文下面将放宽违约指数的构建，新的各个指标的阈值如下表所示：

index_name	RPY_index	BBRR_index	DEBT_index	DBRR_index	CDR_index
threshold	0.6	0.14	23000	1	0.1

3.3模型重建

重新设定阈值后，正负样本比变为0.66负样本:0.34正样本。本文再尝试一下之前的模型，得到如下的PR curve：



可以看出，相比于之前的阈值，现在的模型结果有了极大的提升，特别是随机森林模型的结果。下面给出各个模型（根据正负样本比确定分类阈值）的PR数值：

	LR	SVM	DecisionTree	Adaboost	Randomforest
Precision	0.630	0.671	0.587	0.650	0.737
Recall	0.532	0.157	0.610	0.518	0.629

3.4结果讨论

本文先从一个严格的标准来区分具有潜在严重违约倾向的学校，尝试进行建模并取得了不好的效果。从模型结构的PR曲线图中观察发现PR曲线左处有驼峰，这意味着有部分负样本在预测时具有很高的预测值，因而怀疑是不是标准制定的太严格，以至于很难将部分负样本区别开，同时也没有很好的学习到正样本的特征。于是决定严格标准，即判定为高风险的指标阈值下降一点，再重新建立模型。从重建的模型的结果不难发现，适当放宽构建指数的阈值，相当于更加严格界定潜在违约的标准，能提高识别具有潜在违约倾向的学校的准确率。在有限的特定指标上能达到这个效果，本文认为此次构建的随机森林模型对联邦发放贷款具有一定的借鉴和指导作用。此外，本文的模型在指标上兼容了多个Repayment相关指标，这使

得在不同年份有不同程度指标缺失的情况下，也具有鲁棒性，可以使用；同时，由于不同年份间的指标之间有差别，借此模型也可以借用某年的数据作为训练数据训练模型，对目标年份数据作测试，可以测得在特定年份的违约标准下，目标年份中具有潜在违约倾向的学校，作为一个参考。