OSID

机器学习在我国高收益债投资中的应用研究

鲍俊颖，施成湘
（重庆第二师范学院 数学与信息工程学院，重庆 400065）

摘要：近年来我国债券违约频发，违约数量和违约金额不断创新高。高收益债具有高收益、高信用风险等特点。首次违约债券具备高收益债的特征，兑付概率相对较高，以首次违约高收益债券为研究对象，运用高斯朴素贝叶斯模型和逻辑回归模型筛选高兑付概率违约债券，对投资高收益债具有一定的参考价值。

关键词：高收益债；机器学习；高收益；高风险
中图分类号：F832.48　　文献标识码：A　　文章编号：1008－6390(2021)03－0015－06

我国债券市场快速发展，信用违约问题日益凸显，违约事件频发甚至发生连环违约[1]。自 2014 年我国债券市场出现第一只违约信用债以来，违约规模和违约数量不断创新高。投资违约债券存在一定的信用违约风险，同时也是一种全新的投资方法，即投资高收益债承担高违约风险的同时，能够获取高额收益。截至 2019 年 9 月 7 日，我国债券市场累计违约债券 480 只，其中仅 62 只兑付，实际兑付率为 12.92%，投资违约债券风险相对较大。债务违约是市场经济的必然表现，应当客观理性地看待当前债券市场存在的违约风险及影响[2]，通过研究违约债券的特征，提高违约债券投资风险管理能力，并以市场化为导向，鼓励机构投资者入市，开辟高收益债专门通道，逐步建立中国版垃圾债市场[3]。

机器学习属于人工智能范畴，是数据分析的一种方法，可分为监督学习和无监督学习。监督学习是从输入和输出中学习的算法，高度依赖确定的输入和输出信息。目前比较流行神经网络和决策树技术，主要应用于统计分类和回归分析，常用的经典算法包括朴素贝叶斯、逻辑回归、支持向量机和邻近算法等。无监督学习是从输入中学习的算法，只有输入数据且无标记，通常用于处理无法预知样本标签的问题。无监督学习是让算法自己学习如何处理问题，主成分分析和深度学习算法属于无监督学习。机器学习应用到高收益债投资领域能够帮助投资人做出投资决策。

一、高收益债投资市场

（一）高收益债投资价值

我国债券市场已经发展成为以银行间市场为主、交易所市场为辅，涵盖各类机构投资者和债券投资品种的多层次债券市场体系，债券市场已成为我国金融市场的重要组成部分[4]。债券投资包括投资银行债券承揽、承做、承销等债券买方（自营、资管）投资交易，交易类型可以分为现券买卖、债券回购交易、债券借贷业务、债券远期以及利率期权投资等。债券投资交易主要是杠杆交易，获取债券长短期利差以及行业和品种之间信用利差，一般机构风险偏好相对较低。

自 2014 年首只债券违约后，我国债券市场规模增加、债券市场做市商制度发展、债券内外部评级水平提高，我国高收益债投资市场发展前景广阔。交易所竞价交易系统、上证固定收益平台和银行间交易市场部分高收益债见表 1，其中 15 宜华 01 债券到期收益率达到 287.24%，若不考虑信用风险，持有到期完成兑付将获得相对较高的收益。以 19 方正 MTN002 为例，2019 年 12 月 2 日，北大方正集团有

收稿日期：2020－11－16
基金项目：重庆第二师范学院课程思政特色课程建设项目“概率论与数理统计”（KCSZ202016）；重庆第二师范学院校级科研项目“基于大数据分析的巨灾债券定价模型的实证研究”（KY201728C）；重庆市科委项目“多源异构教育大数据的汇聚与融合技术研究与平台建设”（cstc2020jscx－msxmX0152）
作者简介：鲍俊颖，讲师，研究方向：金融统计；施成湘，教授，研究方向：应用数学。

限公司(以下简称“北大方正”)发行的2019年度第二期超短期融资券未能完成债券的本息兑付,作为知名高校的下属企业,北大方正主体为AAA级大型国有控股企业集团,此消息引起市场各方瞩目。2019年12月2日,19方正MTN002债券估值全价下跌至30.80元,此后受各方面预期因素影响,2019年12月23日,19方正MTN002由30.80元上升至54.77元。19方正MTN002作为高收益债的一种,价格短期内波动相对较大,通过深入研究其债权主体后续偿还能力和高收益债信用风险特点,选择恰当的时机参与投资以获取高额回报具有较大的可能性。

表1　部分高收益债到期收益率

债券名称	剩余期限/天	净价/元	收益率/%	交易场所
17鹏博债	120	47.90	47.34	上证固收平台
14瀚华01	114	65.00	173.34	上证固收平台
15九鼎债	196	92.75	21.52	上证固收平台
15宜华01	150	45.00	287.24	交易所竞价系统
16太安债	351	79.60	32.87	交易所竞价系统
19方正MTN002	606	52.72	27.24	银行间市场

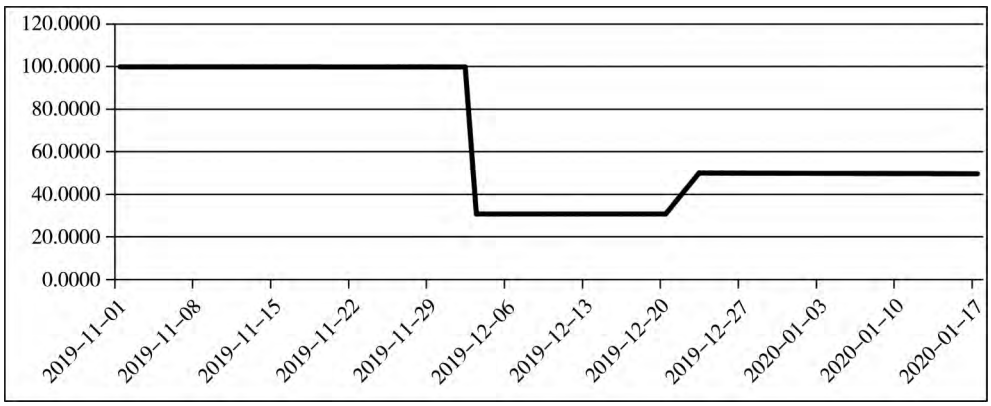

图1　18方正01债券中债估值

投资高收益债需要承担债券违约风险,如果无法按期兑付,在我国现有法律体系和机制下,投资人可通过以下几种方式维护合法权益:一是投资人和债务发行主体协商,通过债务重组偿付债券;二是通过求偿诉讼、破产重整和破产清算等司法途径;三是对于有担保增信措施的债券,通过向担保人追偿违约债务或通过处置抵押物获取补偿收益。目前,我国债券违约后债券持有人主要通过前两种方式受偿,其中债券发行人自筹款项和不良资产收购是周期最短的方式,但要求债券发行人要有较强的偿还意识和外部资金支持。考虑到以上特点,高收益债参与主体的资金应当以中长期资金为主,来源不稳定的短期负债资金要谨慎参与,否则,投资违约债券有可能触发流动性风险,对投资收益产生较大影响。

(二)高收益债投资风险

在成熟资本市场,高收益债已经成为企业融资工具和投资者参与投资的重要渠道,高收益债市场发展助推了实体企业,尤其是高科技企业的发展。高收益债市场发展主要由金融市场本身决定,影响因素包括参与债券的投资者结构、债券整体违约水平、债券市场流动性和债券违约处置效率等宏观因素,以及债券投资评级市场、利率衍生品和信用衍生品发展水平等微观因素。

与股票投资相比,债券投资需要承担市场利率风险、买卖流动性风险、债券信用违约风险。其中,债券投资信用风险是指参与债券投资无法按时兑付利息或本金风险。高收益债券对应较高信用风险债券,部分高收益债已经发生信用风险,通过研究和判断高收益债信用发展趋势,择机参与债券投资是目前债券市场比较主流的投资形式。高收益债投资收益主要由三部分构成,持有到期票息收入,信用资质改善后债券净价上升获取资本利得收益,以及债券违约后通过诉讼等手段最终能够获取的合法收益。与之对应的是无法按照约定获得票息收入的风险,

债券主体信用资质继续恶化导致债券净价大幅下跌的风险，违约诉讼无法获取预期收益的风险。高收益债发生违约，会同时触发以上三类风险，往往获取收益难以覆盖投资成本，高收益债风险远高于一般债券投资，适合较高风险识别和承受能力的投资者参与。

（三）高收益债投资方式

高收益债投资一般有三种方式，一是通过分析高收益债市场，投资低估价值债券获取资本利得收益，一般投资期限相对较短，同时借助利率衍生品对冲市场利率风险；二是通过分散投资高收益债，获取高收益债投资收益，收益覆盖违约损失；三是主动参与高风险违约债券，通过研究债务主体偿还能力，低价收购即将违约或者已经违约债券，获取延期支付或破产清算收益。该类投资期限较长，适合负债长期稳定的专业机构。

本文使用机器学习模型，主要探讨第二种和第三种高收益债投资方式，投资信用风险预期较好且已发生违约风险的高收益债券，最大限度降低信用风险水平，获取最终兑付收益。对已经发生实质性违约的高收益债，可通过合法途径获取部分或全部权益。

二、高收益债数据特征处理

（一）数据特征处理方法

运用机器学习解决实际问题首先要对数据进行预处理，构建良好的数据表征，这一过程一般被称作特征工程或者特征提取，即通过提取数据特征并将特征向量化作为学习输入值；其次是根据解决问题的实际需要和样本数据量选取合适的应用模型；最后是对各类模型结果进行学习和检验，并选用最有效的学习模型。测试模型数据称为训练数据，其又可分为测试数据和留出数据，测试数据用于训练模型，留出数据用于评估模型，机器学习需严格按照学习流程进行学习和检验，确保模型的实用性。

（二）数据特征选择

使用机器学习模型分析高收益债，首先要根据事先的数据分析并借鉴实际操作经验选取有预测效果的数据特征，数据特征选取直接决定了模型结果的准确性。下面以高收益债历史数据为基础，分析高收益债企业性质、募集方式、发行主体评级和违约日期等特征是否具有预测区分效果。

1. 企业性质

债券发行主体企业分为央企、地方国企、民营企业、公众企业、中外合资和外资企业。央企通常是中央直属企业，其企业主体信用风险低于一般企业，央企债券发行规模大、协调难度高，违约后能够全额兑付，但最终兑付周期比较长。地方国有企业回收周期长，回收率较低，例如广西有色进入破产清算，预计回收率不到10%，东北特钢违约后破产重整两年多，最终兑付率为22.09%。民营企业抗风险能力较弱，与外部融资环境密切，在融资高杠杆模式下，违约后再融资困难极大，违约后实际回收率低于央企和地方国有企业。目前，已违约企业主要以民营企业为主。由以上特点可知企业性质具有明显的区分性，发行主体性质能够作为数据分析的一个特征使用。

2. 债券类型

我国信用债分为公司债、企业债、短融、中票和非公开定向债务融资工具（PPN），不同债券类型对应不同发行审批制度。其中，公司债发行采取核准制，由证监会审核；企业债由发改委审核，每半年确定一次发行额度；短融、中票和PPN为注册制，由交易商协会审批。当前企业债违约率低于公司债、公司债违约率低于短融、中票和PPN，实际违约率具有明显区分性，因此可以作为数据分析的一个特征使用。

3. 外部评级

已违约债券发行主体评级为AAA的占比3.72%，评级为AA+的占比21.28%，评级为AA及以下的占比51.24%，债券外部评级数据具有明显的区分性，历史数据表明其可以作为一个特征进行分析，详细数据见表2。

表2　违约债券评级分布

违约债券发行时主体评级	数量	占比/%
AAA	18	3.72
AA+	103	21.28
AA及以下	363	75.01
–	484	100.00

4. 其他特征

结合历史经验和数据分析，债券资金募集方式、债券违约时间、主体是否A股上市、担保措施、主体企业所属区域等指标都有一定的区分性，可以作为数据分析特征。

（三）数据特征标准化

违约债券首次违约样本数量为135个，违约债券特征分为违约日期、行业、民营企业债或国有企业债、所属品种、募集方式、上市与否、发行评级、违约日评级等，预测目标为违约本金兑付率。样本数据实例如表3所示。

将样本数据特征标准化，特征数据见表4。

表3　违约债券数据实例

简称	余额	发行评级	违约评级	类别	上市	区域	行业	发行方式	兑付
11 超日债	10.00	C	AA	民营企业	是	上海	半导体	公募	兑付
12 东飞01	1.03	A	A	中外合资	否	江苏	机械	私募	未兑
…	…	…	…	…	…	…	…	…	…
15 沈机床股	5.00	C	AA	地方国有	是	辽宁	机械	公募	未兑
17 三鼎01	3.44	C	AA	民营企业	否	浙江	综合	公募	兑付

表4　违约债券特征

序列	特征	特征值	取值
1	违约日期	按年份分为2014至2019	距首次违约事件时间间隔
2	债券性质	一般公司债、企业债、中期票据	5
		超短融、短融	4
		定向工具、私募债	3
		可交换债，证监会ABS	2
		股权交易中心债	1
3	企业性质	地方国有企业	2
		公众企业	－1
		民营企业	1
		外资企业	－2
		中外合资企业	0
4	是否上市	上市	1
		未上市	0
5	发行方式	公募	1
		私募	0
6	担保措施	有担保	1
		无担保	0
7	发行时主体评级	按照评级由低到高依次取值	－1～7
8	违约后一周主体评级	按照评级由低到高依次取值	0～17
9	首次违约时债券余额	首次违约时债券余额(亿元)	债券余额(亿元)
10	行业评级调低	所属行业评级调低数量	1～10
11	区域	按照GDP排名划前50%	1
		按照GDP排名划前50%	0
12	兑付	兑付	1
		未兑付	0

观察样本距离首次违约间隔时间、发行时主体评级、企业性质、所属行业，样本能够区分为两个类别，这在一定程度上说明机器学习模型可以学会它们的区分方法，如图2所示。

三、机器学习模型测算

以首次违约债券为例进行测算，将首次违约债券投资简化为分类问题，即根据历史违约特征分为能够兑付和不能兑付两类，投资能够兑付的债券，获取最大收益。以下分别应用高斯朴素贝叶斯模型和逻辑回归模型进行测试。

(一) 朴素贝叶斯模型

1. 模型说明

分类问题是数据挖掘领域的重要研究课题，而朴素贝叶斯分类是最常见的分类算法之一[5]，其优点是简单高效、运算速度快、参数相对较少，经常

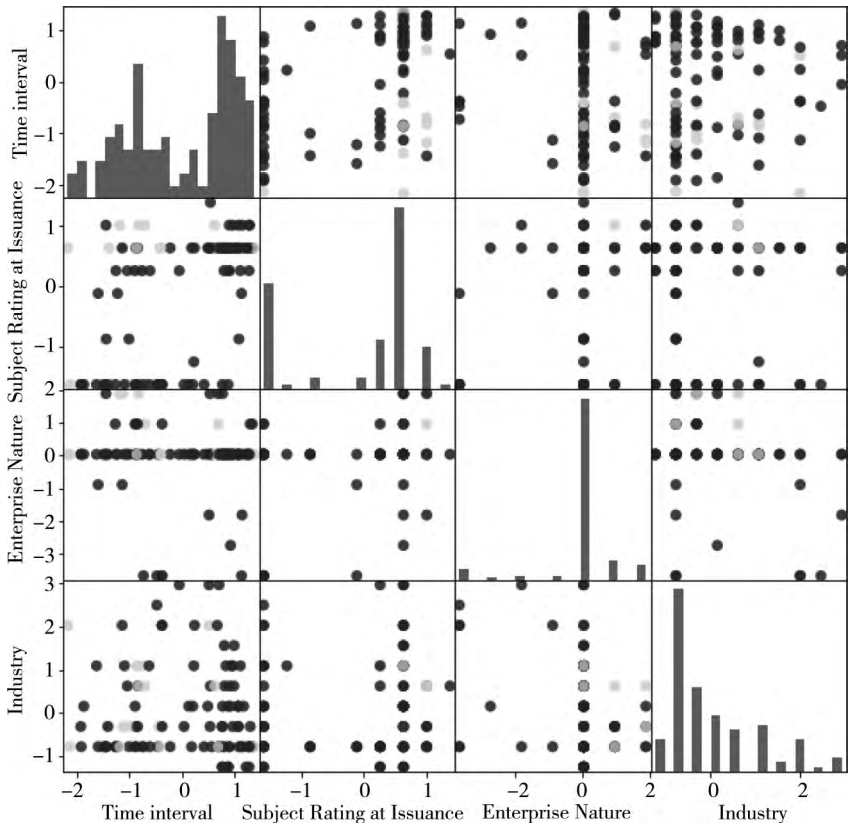

图2　数据集的散点图矩阵

用于高维数据分类问题。其理论基础是贝叶斯定理，主要特点是假设各属性之间是相互独立的，具体应用到违约债券投资算法如下：

(1) 假设违约债券投资样本为 x，类别标记为违约债券兑付和未兑付两种可能，并记为 $y=\{0,1\}$。

(2) 对联合概率分布 $p(x,c)$ 建模，获取 $p(x\mid c)$，由条件概率公式及贝叶斯定理可知：

$$p(c\mid x)=\frac{p(c)\,p(x,c)}{p(x)}$$

其中，属于某类先验概率记为 $p(c)$，样本 x 对应类别为 c 的类条件概率为 $p(x\mid c)$。

(3) 假设违约债券各属性相互独立，即违约债券各属性相互独立地对分类结果产生影响，则 $p(x\mid c)$ 变为：

$$p(c\mid x)=\frac{p(c)\,p(x,c)}{p(x)}=\frac{p(c)}{p(x)}\prod_{i=1}^{d}p(x_i\mid c)$$

其中，d 表示违约债券属性个数，x_i 表示样本在第 i 个属性的值，样本集可确定 $p(x)$ 且各个类都相同。

(4) 由于以上模型为二分类问题，分类问题简化为求解以下公式结果的最大值，并将其作为分类结果：

$$p(c=1\mid x)=p(c=1)\,p_{x_1\mid c=1},\cdots,p_{x_i\mid c=1}$$

$$p(c=0\mid x)=p(c=0)\,p_{x_1\mid c=0},\cdots,p_{x_i\mid c=0}$$

(5) 假定样本特征的概率分布为正态分布，以下采用高斯朴素贝叶斯模型对数据进行测算，测算结果如图3所示。

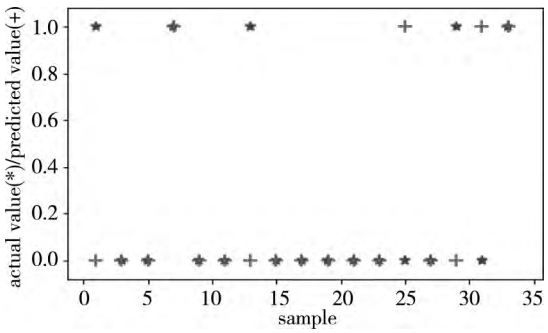

图3　高斯朴素贝叶斯模型测算结果

2. 测算结果分析

样本数据75%作为训练集，25%作为验证集，根据高斯朴素贝叶斯模型测算，验证集预测正确率为73.53%，其中预测投资能够兑付成功率为40.00%，预测违约成功率为87.50%。若以兑付数量为口径计算，样本兑付率为20.59%，预测投资违约债券成功率高于平均兑付率。详细数据见表5。

表5　高斯贝叶斯模型测算结果

预测样本	验证集数量	正确数量	成功率/%
兑付样本	10	4	40.00
违约样本	24	21	87.50
合计	34	25	73.53

表6　逻辑回归模型测算结果

预测样本	验证集数量	正确数量	成功率/%
兑付样本	3	1	33.33
违约样本	31	25	80.65
全部样本	34	26	76.47

(二)逻辑回归模型预测

1. 模型说明

如何高效地对数据进行分类成为迫切的需求[4]，其中逻辑回归是一种广义线性回归模型，主要应用于解决二分类问题。逻辑回归模型如下:

$$P(y=1\mid x)=\frac{e^{w_0+w_1x_1+w_2x_2+\cdots+w_dx_d}}{1+e^{w_0+w_1x_1+w_2x_2+\cdots+w_dx_d}}$$

假设特征的数量为 d，债券特征为 x 时能够兑付的概率表示为 $P(y=1\mid x)$，债券特征为 x 时不能兑付的概率表示为 $P(y=0\mid x)=1-P(y=1\mid x)$。当预测概率值大于0.5时，预测可以兑付，反之不可以兑付。

逻辑回归模型的似然函数为:

$$L(\vec{w})=\Pi_{i=1}^{N}p(y_i=1\mid x_i)^{y_i}(1-p(y_i=1\mid x_i))^{1-y_i}$$

逻辑函数的损失函数为似然函数的负对数:

$$C(\vec{w})=-\log(L(\vec{w}))=-\sum_{i=1}^{N}(y_i\log p(y_i=1\mid x_i)+(1-y_i)\log(1-p(y_i=1\mid x_i)))$$

使用逻辑回归模型进行测算，测算结果如图4所示。

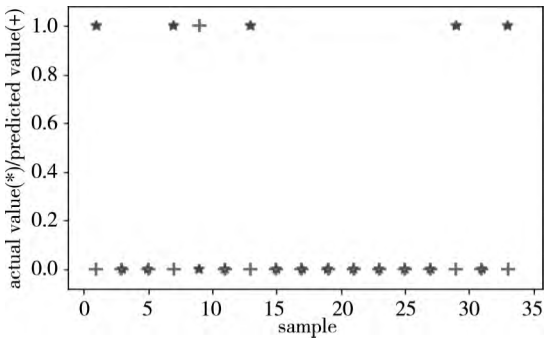

图4　逻辑回归模型测算结果

2. 测算结果分析

根据逻辑回归模型测算，样本数据75%作为训练集，25%作为验证集，验证集预测正确率为76.47%，其中预测投资能够兑付成功率为33.33%，预测违约成功率为80.65%。以兑付数量为口径计算，样本兑付率为20.58%，按照模型预测投资违约债券成功率高于平均兑付率。详细数据见表6。

(三)模型测算对比

样本兑付率为20.58%，使用高斯贝叶斯模型和逻辑回归模型预测成功率分别为40.00%和33.33%，明显高于样本平均值，高斯贝叶斯模型表现优于逻辑回归模型。若考虑以上两个模型的预测结果，预测兑付成功率为50%，预测违约成功率为86.96%，整体预测成功率高于单一模型。综合两个模型能够提高预测成功率，但实际投资机会大幅减少。

表7　综合测算结果

预测样本	验证集数量	正确数量	成功率/%
兑付样本	2	1	50.00
违约样本	23	20	86.96
全部样本	25	21	84.00

四、总结

以债券发行主体首次违约数据为基础，选择高斯贝叶斯模型和逻辑回归模型对投资违约债券兑付成功率进行预测，结果显示:预测违约债券投资兑付概率明显高于样本平均值，在一定程度上能够指导债券投资决策。目前，研究样本数量相对较少，随着违约债券数量的增加，还可探讨使用其他机器学习模型进一步提高预测的稳定性。

参考文献:

[1] 袁彦娟，曹晓黎，马莉娜. 我国债券市场信用违约问题研究[J]. 区域金融研究，2018(3):42-45.

[2] 马保明，李文，邓惠中. 证券公司自营业务预期违约债券处置研究[J]. 北方金融，2019(7):55-59.

[3] 刘雨桐. 浅谈债券市场的信用违约及风险管理[J]. 现代经济信息，2019(9):317.

[4] 孙司宇. 我国债券市场现状、问题和发展方向初探[J]. 白城师范学院学报，2017，31(1):93-96.

[5] 李文超，王彦焱. 朴素贝叶斯模型及朴素贝叶斯假设改进[J]. 数码世界，2017(7):112-13.

[6] 朱嘉诚. 基于逻辑回归的分类算法研究[J]. 数字化用户，2017，23(46):152.

责任编辑　乡　下

Abstracts of Major Papers in Issue

On the Moral Practice Implication of Carrying Forward the Great Anti-epidemic Spirit: from the Perspective of Moral Narrative Theory by GUO Xiaoyu P. 5

As a kind of specific moral inquiry methodology originated from virtue ethics, moral narrative theory pays deep theoretical attention to the external context and internal path of moral practice and character cultivation, and has becomes a theoretical perspective to explore and promote the moral practice implication of the great anti-epidemic spirit. To vigorously promote the spirit is to strive to integrate it into the spiritual character of the Chinese nation, so that the anti-epidemic spirit can lead the moral cultivation of individual citizens. From the perspective of ethics, it is helpful for the great anti epidemic spirit to give full play to the value leading role in the process of creating a new historical cause and speed up the pace of Chinese people to achieve a better life.

Key words: anti-epidemic spirit; moral narrative; spiritual character; moral cultivation

Research on the Application of Machine Learning in Default Bond Investment in China by BAO Junying, SHI Chengxiang P. 15

In recent years, China´s bond defaults occur frequently, the number and the amount of defaults continue to reach new highs. High yield bonds have the characteristics of high yield and high credit risk. Referring to the international mature capital market, China´s high-yield bond market has broad development space. First time default bonds have the characteristics of high-yield bonds, and the probability of cashing is relatively high. This paper takes the first default high-yield bonds as the research object, and uses Gaussian naive Bayesian model and logistic regression model to select Default bonds with high cashing probability, which has certain practical reference significance for participating in high-yield bond investment.

Key words: default bond; machine learning; high yield; high risk

Dramaturgical Analysis of Early Etiquette Activities in China by LI Xiangxiang P. 26

The etiquette activities before Confucius paid too much attention to the external form rather than the inner true feelings, thus with the characteristics of performance. With the help of Goffman´s dramaturgical theory, it is more clearly to see the similarities between Chinese early ritual activities and stage performances: the roles on the stage were assigned according to ritual system, ritual utensils were used as props for performances, and ceremonies were stage performances one after another. Although the etiquette activities aimed to educate all social strata to be content with their roles, it was difficult to integrate the performers with their roles because of the lack of real feelings in this kind of "performance", which eventually led to the collapse of etiquette and music. Therefore, Confucius emphasized that etiquette should be supported by benevolence, and the moral significance of etiquette is to realize the integration of roles and performers, so as to enhance the binding and persuasive power of etiquette.

Key words: character; etiquette; dramaturgical theory; original Confucianism

Study of English Translation Style of Chapter Titles in *Hong Lou Meng* by H. B. Joly: Comparing with Three Complete English Translation Versions by JI Shufeng P. 40

H. B. Jolly´s translation of *Hong Lou Meng* (56 chapters) serves as a connecting link between the preceding and the following in the history of translation of *Hong Lou Meng*. It is a pity that it has not attracted much attention