Emergence and Transfer of Wealth 财富涌现与流转, 2022, 12(1), 1-10
Published Online March 2022 in Hans. http://www.hanspub.org/journal/etw
https://doi.org/10.12677/etw.2022.121001

Hans汉斯

基于数据挖掘的公司债券风险预测分析

葛柯男，张有中*

厦门大学嘉庚学院管理学院，福建 漳州

收稿日期：2022年5月19日；录用日期：2022年6月20日；发布日期：2022年6月29日

摘　要

大数据时代的到来，网络与计算机技术的发展，给债券市场带来了风险预警的新工具。本文从公司内部经营状况的微观风险信息角度出发，利用数据挖掘技术找出影响公司债券到期偿还的关键因素，并建立预测债券违约的方法。研究以XGBoost极端梯度提升算法发现债券是否违约的主要影响因素是营业收入同比增长率和资产负债率，然后建立了债券是否违约的二元logistic回归模型，通过二元logistic回归模型可以进行债券违约的预测。

关键词

公司债券，数据挖掘，XGBoost极端梯度提升，逻辑回归

Prediction and Analysis of Corporate Bond Risk Based on Data Mining

Kenan Ge, Yu-Chung Chang*

School of Management, Xiamen University Tan Kah Kee College, Zhangzhou Fujian

Received: May 19th, 2022; accepted: Jun. 20th, 2022; published: Jun. 29th, 2022

Abstract

The advent of the era of big data and the development of the Internet and computer technology have brought new tools for risk early warning to the bond market. From the perspective of micro risk information of the company's internal operation, this paper uses data mining technology to find out the key factors affecting the maturity of corporate bonds, and establishes a method to predict bond default. Using the extreme gradient boosting algorithm, it is found that the main influencing factors of whether the bond defaults are the year-on-year growth rate of operating revenue and asset liability ratio in this paper. And then we establish a binary logistic regression model of the bond. The binary logistic regression model can predict whether the bond defaults.

*通讯作者。

文章引用: 葛柯男, 张有中. 基于数据挖掘的公司债券风险预测分析[J]. 财富涌现与流转, 2022, 11(1): 1-10.
DOI: 10.12677/etw.2022.121001

Keywords

Corporate Bond, Data Mining, Extreme Gradient Boosting, Logistic Regression

1. 引言

我国债券市场的发展，始于 1894 年清政府为支付甲午战争的军费发行的“息借商款”，发行总额达到白银 1 千多万两。1950 年我国发行人民胜利折实公债，规模为 2.6 亿元人民币，至 2020 年我国债券市场全年融资规模达到 12.78 万亿元人民币[1]。金融市场的不断发展与变化，使我国直接融资方式不断深化与改革，债券市场发展迅猛。债券市场的发行主体、融资的手段、渠道、条款也都更加的多元化。但是随着国内经济下行压力的加大、有效投资的增长乏力，许多公司的经济恶化，公司面临着财务、经营与信用风险增大等许多问题，债券违约事件开始出现。

2014 年 3 月 4 日 ST 超日于晚间发布声明称“11 超日债”债券将无法按时偿付全额本期债款，打破了我国债券市场刚性兑付的神话。此后公司债券违约事件频频发生。2016 年债券违约进入加速爆发期，2018 年债券违约浪潮加大，截至 2020 年我国债券违约金额高达 500 多亿人民币，这也标志着我国债券市场违约事件已经进入常态化的进程。因此在当前债券违约事件频发且呈现扩张之势的情况下，如何有效的预测并防范债券违约事件的发生，对于金融风险的防范具有重要的意义。

公司债券的风险预测对债券市场的健康发展有着积极的意义，因为一旦没有债券风险预测，或者是相关债券风险预测的模型准确度不高的话，投资者将无法对债券的情况做出正确的判断。若发行债券的公司到期违约，无力偿还本金及支付利息，将会挫败投资者对债券市场的信心。债券风险的预测可以使投资者对于债券有一定的认识和心理准备，有效的保证证券市场的稳定。

债券到期后能否还本付息的影响因素有很多，任何一个影响因素都有可能引发债券违约。无论是从外部环境的因素，例如宏观经济、行业因素、政府制度等方面，还是从企业内部经营因素，例如盈利能力、资产流动力，资产负债率等，都不是投资者能够完全掌握并全方位进行考虑的，并且市场上更存在一些信息披露不及时的情况，导致投资者无法及时掌握债券信息，从而造成债券违约事件频发。

企业债券的风险预测分析实质上是对企业的信贷风险进行评价。在信用市场，各式各样的风险随时存在。每一个信贷风险评估的决策，都是以一家公司偿还债款能力的 3-5 个关键指标为依据。本文希望通过数据挖掘的方法，找出反映公司偿还债款能力的关键因素，以及各项影响因素的重要性及占比，建立债券违约风险的预测机制，从而帮助投资者选择信誉良好、违约风险较小的公司。这样可以使得企业以更快、更低的利率获取资金，而不是通过提高利率去吸引投资者投资，导致偿还利息及本金的压力过大，以及债券违约风险的增加。

2. 文献综述

2.1. 公司债券

公司债券是公司依照法定的程序进行发行，约定在一定的期限内还本付息的有价证券[2]。对于持有者来说，公司债券只是向公司提供贷款的证书，所反映的也只是一种普通的债权与债务关系。简单来说

公司债券的持有者是企业的债权人，并非企业的所有者。债券持有者按照约定向企业收取相应的利息，并于约定的到期时间收取本金，这是企业债券人的基本权利，所取得的利息优先于股东的分红，就算企业破产清算时，也优先于股东拿回本金。只是债券持有者不能参与企业的经营、管理等各项活动。

2.2. 债券违约的因素

债券违约的外部成因主要包括宏观经济形式、行业环境、融资环境、信用评级[3]。例如：我国近年来的宏观经济下行，会加大企业出现财务困难的可能性，从而会导致债券违约现象的出现。如果行业大环境发展的不好，当新技术出现或一些新政策施行时，可能造成行业发展严重受挫，企业的收益更容易出现问题，从而导致债券违约。在信用评级方面，目前我国的信用评级机构对债券信息的披露相对缓慢，很多时候评级的下调公示是在违约曝光后才发布的，并没有起到一个很好的警示作用。

债券违约的内部成因主要是包括公司战略、公司治理、发行总额等[4]。企业的战略出现问题很容易造成企业出现资金紧张的情况。对于发行总额来说，如果企业盲目的进行集资扩张，容易造成企业的资金链出现问题。

2.3. 国内外研究现状

国外的债券市场发展比较早，在 1700 年便广泛流传。Ohlson 早在上世纪八十年代就提出考虑公司财务数据如资产负债率、流动比率等指标的逻辑回归(Logstic Regression)模型，来预测债券的信用风险[5]。Foxon(2007)认为，债券体系急需监管和完善相关的法律体系，因为投资者主要依靠债券的评级机构所提供的信息进行判断投资，而债券评级机构与债券发行人之间会存在利益关系，有的时候评级信息不公正，评级机构所公布出来的信息可能是进行过修饰的数据[6]。

Collin-Dufresne 等人通过对债券市场中的交易者报价和交易价格，对影响信用价差变动的因素进行了分析。发现主成份分析显示影响信用价差变动主要是由于一个共同的因子所致，但是他们不能发现和解释这个共同因子具体是什么，为什么会造成这种影响[7]。Bakshi 等利用可观察的经济因素对信用风险模型进行了实证研究，发现利率风险可能是定价和对冲的一级显着因素，考虑杠杆和账面市值的信用风险模型，可以减少样本外收益率拟合误差[8]。Azizpour 等人研究了美国企业违约聚集的原因，他们否认了公司债券违约与时间有关的假设，并且找到强有力的证据表明，传染是其中非常重要的一个聚集源。一个公司的违约可以直接影响到其它公司的运行健康[9]。

我国的资本市场结构在改革开放过程中不断得到改善和深化，而债券市场作为资本市场的一个重要组成部分，却始终处于相对落后的地位。我国在债券市场发展之初，政府和有关学者对债券的监管和认识还不够充分，制约了我国债券市场的发展。1984 年到 1992 年，我国的债券市场开始逐步发展，经过了 1993 年到 1998 年的整顿和规范，1999 年到现在已经进入了正规的发展时期。2007 年是我国债券市场的一个重大转折，其发债规模实现了质的飞跃。伴随债券市场发展壮大而来的是债券违约事件越来越多，国内学者对此也相应的展开了研究及探讨。

周宏、徐兆铭等通过对 89 家 2007~2009 年度的公司债券面板数据进行分析，发现宏观经济不确定性对中国公司债券信用风险具有显著影响。在这些因素中，金融危机的爆发、股市波动性、通货膨胀率、人民币兑美元汇率对公司债券的信用风险产生了重要的影响。在经济不乐观的情况下，企业债券的风险会增大[10]。周宏，林晚发等以 BS 模型为基础，构建包含信息不对称的企业债券风险评估模型，发现信息不对称程度与企业债券信用利差存在显著的正相关性，一个债券披露的信息越多，信用利差就越低，投资者所蒙受损失的风险就会越小[11]。

对于债券的信用风险预测，国内的部分学者建立了相关的模型进行风险预测。黄石、黄长宇通过建立 KMV 模型对于债券发行主体，进行了信用的风险评级，并且进行发行规模的推算[12]。曹萍则通过 KMV 模型建立对地方政府债券违约风险的评估体系[13]。

曾江洪、王庄志等通过基于统计学理论的 SVM 模型，对于中小型集合债券融资个体信用风险进行度量，经过数据检验，模型的预测准确率高达 90.77% [14]。刘慧芳通过建立 Logistic 模型，分析了各个主成分与违约概率之间的关系，进而得到了准确度较高的模型[15]。沙一诺基于数据挖掘模型构建企业债券违约风险的预测方法，发现 XGBoost 模型以及 Light GBM 模型的预测效果较好[16]。

2.4. 数据挖掘

数据挖掘是从大量、不完全、噪声、模糊、随机的真实数据中抽取出隐藏在其中的、人们不知道的、但具有潜在价值的信息和知识的过程[17]。数据挖掘与传统的数据分析有着本质的区别，数据挖掘是在没有明确假设的前提下挖掘有用的信息，这些信息及结果都是通过大量的搜索工作从数据中自动提取出来的，数据挖掘是要发现那些不能靠直觉发现的信息或知识、甚至是违背直觉的信息或知识，而所得到的信息也应该具有先前未知、有效和实用这三个特征。

按照学习目标对数据挖掘算法进行分类，一般常使用的算法有下列几种：

1) 概念学习：典型的有示例学习。
2) 规则学习：为了获得规则的一种学习，主要的规则学习有决策树学习。
3) 函数学习：典型的函数学习有神经网络学习。
4) 类别学习：主要的类别学习有聚类分析。
5) 贝叶斯网络学习。

3. 研究方法

3.1. 数据来源

本文的数据来源于 wind 数据库，通过数据库查找 2020 年发生债券违约的 15 家上市公司在违约当年度的财务报表，因为这 15 家公司其债券发行时间集中在 2015 年至 2016 年间，因此再收集 2015 年至 2016 年发行债券，但在 2020 年未发生违约事件的 870 家上市公司的 2020 年度财务报表，总计 885 家上市公司的 2020 年度财务报表。选定同一个年度的财务报表，主要是为了剔除各年度的宏观经济、国家政策、大环境(例如疫情原因)等其他因数的干扰。

3.2. 数据概览

上市公司的财务报表包括资产负债表、损益表、现金流量表或财务状况变动表、附表和附注，其中变量超过 45 项以上。本文经过文献的综合整理与比对后，删除其他与债券违约风险较无关的变量。数据一共 16 个变量，包括：X_1 速动比率、X_2 流动比率、X_3 应收账款周转率、X_4 现金流量利息保障倍数、X_5 存货周转率、X_6 净资产收益率 ROE、X_7 总资产净利率、X_8 销售毛利率、X_9 销售成本率、X_{10} 利润总额(同比增长率)、X_{11} 营业收入(同比增长率)、X_{12} 资产负债率、X_{13} 有息负债率、X_{14} 息税折旧摊销前利润 EBITDA、X_{15} 公司属性、X_{16} 资产总计。

3.3. 分析方法

本文拟通过构建 XGBoost 模型和逻辑回归模型对债券违约风险进行识别与预测，找出影响债券违约的关键因素，并建立相应的模型，希望对债券违约的预警有所帮助。

3.3.1. XGBoost 模型

XGBoost 极端梯度提升模型算法是基于 CART 回归树模型改进的，CART 回归树是二叉树模型，会根据样本特征来划分样本空间，不断分裂出左子树和右子树。XGBoost 算法是通过不断添加树，而新添加的树会根据样本特征再次分裂生长一棵树，添加树的过程就是学习新函数的过程。然后根据样本特征将每棵树落在对应的叶子节点上，每个叶子节点上的分数相加就能得到预测值。XGBoost 的优点是稳定性好、结果预测精确度高，并且对于数据中存在的问题如数据噪声、多重线性相关等问题的敏感度比较低，比较不容易受到影响。XGBoost 模型在缺失函数中加入了正则项，用来控制模型的复杂程度。从权衡方差和偏差的角度来看，XGBoost 模型的降低了模型的方差，使得学习所得的模型更加简单，可以防止过度拟合，减少计算量。

XGBoost 的函数迭加如下[18]：

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{1}$$

其中 $\hat{y}_i^{(0)} = 0$， $\hat{y}_i^{(t)}$ 是第 t 次迭代后样本 i 的预测结果， $f_t(x_i)$ 是第 t 棵树的函数。目标函数为

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega(f_i) \tag{2}$$

其中 $\sum_{i=1}^{n} l(y_i, \hat{y}_i)$ 代表损失函数，可由预测值与真实值表示，n 为样本数， $\sum_{i=1}^{t} \Omega(f_i)$ 为抑制模型复杂度的正则项。当 t 棵树生成后，目标函数变为

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_i) \\ &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \sum_{i=1}^{t} \Omega(f_i) \end{aligned} \tag{3}$$

由于前 $t-1$ 棵树的结构已确定，因此前 $t-1$ 棵树的复杂度之和，可以表示为一个常量，即

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + \text{常量} \tag{4}$$

对损失函数在 $t-1$ 棵树 $\hat{y}_i^{(t-1)}$ 处取泰勒公式的二阶展开近似值后，目标函数变为近似值

$$Obj^{(t)} \approx \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + \text{常量} \tag{5}$$

其中 g_i 为损失函数的一阶导数， h_i 为损失函数的二阶导数。由于在第 t 棵树时 $\hat{y}_i^{(t-1)}$ 是一个已知值，所以此时的 $l\left(y_i, \hat{y}_i^{(t-1)}\right)$ 为一个常量，去除所有常量后，目标函数近似值为

$$Obj^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{6}$$

复杂度由叶子结点个数 T 组成，为了抑制模型复杂度 $\Omega(f_t)$ 表示为

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{t} \omega_j^2 \tag{7}$$

其中 ω 表示叶子结点分数，以 γ 控制叶子结点个数，以 λ 控制叶子结点分数。第 j 个叶子结点的所有样本 x_i 以集合 $I_j = \{i \mid q(x_i) = j\}$ 表示，第 t 棵树的函数 $f_t(x_i) = \omega_q(x_i)$，并定义 $G_j = \sum_{i \in I_j} g_i$ 为第 j 个叶子结点所有样本的一阶导数累加之和， $H_j = \sum_{i \in I_j} h_i$ 为第 j 个叶子结点所有样本的二阶导数累加之和。因为 G_j 和 H_j 为常量，目标函数近似值变为

$$\begin{aligned} Obj^{(t)} &\approx \sum_{i=1}^{n}\left[g_i\omega_q\left(x_i\right)+\frac{1}{2}h_i\omega_q^2\left(x_i\right)\right]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{t}\omega_j^2 \\ &=\sum_{j=1}^{T}\left[\sum_{i\in I_j}g_i\omega_j+\frac{1}{2}\left(\sum_{i\in I_j}h_i+\lambda\right)\omega_j^2\right]+\gamma T \\ &=\sum_{j=1}^{T}\left[G_j\omega_j+\frac{1}{2}\left(H_j+\lambda\right)\omega_j^2\right]+\gamma T \end{aligned} \tag{8}$$

求目标函数最值，对 ω_j 一阶导数为 0，可得第 j 个叶子结点权重分数为

$$\omega_j^*=-\frac{G_j}{H_j+\lambda} \tag{9}$$

因此目标函数为

$$Obj=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \tag{10}$$

回归树的最佳划分点则为

$$\begin{aligned} Gain &= Obj_{L+R}-\left(Obj_L+Obj_R\right) \\ &=\left[-\frac{1}{2}\frac{\left(G_L+G_R\right)^2}{H_L+H_R+\lambda}+\gamma T\right]-\left[-\frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}\right)+\gamma\left(T+1\right)\right] \\ &=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{\left(G_L+G_R\right)^2}{H_L+H_R+\lambda}\right]-\gamma \end{aligned} \tag{11}$$

3.3.2. 二元 Logistic 回归模型

Logistic 模型虽然被称为逻辑回归模型，但实际上是一个分类模型。二元逻辑回归的分布函数和密度函数分别如下[19]：

$$F\left(x\right)=P\left(X\le x\right)=\frac{1}{1+\mathrm{e}^{-(x-\mu)/\gamma}} \tag{12}$$

$$f\left(x\right)=F'\left(X\le x\right)=\frac{\mathrm{e}^{-(x-\mu)/\gamma}}{\gamma\left(1+\mathrm{e}^{-(x-\mu)/\gamma}\right)^2} \tag{13}$$

如果因变量 Y 是二元分类，例如事件是否发生，发生编码为 1，未发生编码为 0。模型考虑 k 个自变量，事件发生 $Y=1$ 的概率是

$$P=\frac{\exp\left(\beta_0+\beta_1x_1+\cdots+\beta_kx_k\right)}{1+\exp\left(\beta_0+\beta_1x_1+\cdots+\beta_kx_k\right)} \tag{14}$$

则逻辑回归模型的公式为

$$\mathrm{logit}\left(P\right)=\ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k \tag{15}$$

4. 数据分析及模型构建

4.1. XGBoost 模型构建

本文首先针对缺失值进行处理，删除缺失值，再对结局变量即是否发生违约进行重新赋值，发生违

约的赋值为 1，没有发生违约的赋值为 0。885 条数据分为训练集和测试集，如表 1 所示。

Table 1. The allocation of the training set and test set
表 1. 训练集和测试集的分配

	未违约公司	违约公司
训练集	609	10
测试集	261	5

然后把数据转化成 XGBoost 算法需要的矩阵，构建目标如方程式(10)，计算 Gain 值如方程式(11)。XGBoost 算法输出结果如图 1 所示。图 1 输出的结果包含：Gain、Cover 等值，其中 Gain 是指相应的这个特征变量的重要性。这个变量特征是通过对模型中的每棵树采取每个特征的贡献，而计算出的相对贡献。与其他的特征相比，Gain 值较高意味着它对于生成预测更为重要。在所有变量中 X_{12} 资产负债率最重要,其次为 X_{11} 营业收入(同比增长率),然后是 X_6 净资产收益率 ROE 和 X_3 应收账款周转率。Cover 覆盖度量指的是与此变量相关的观测值相对数量。也就是说 X_{12} 资产负债率和 X_{11} 营业收入(同比增长率)两个变量可以覆盖 77.4%的观测结果。

Filter

	Feature	Gain	Cover	Frequency	Importance
1	x12	0.48319289	0.279218972	0.2	0.48319289
2	x11	0.34419405	0.498226368	0.4	0.34419405
3	x6	0.09954610	0.217094110	0.2	0.09954610
4	x3	0.07306696	0.005460549	0.2	0.07306696

Figure 1. XGboost algorithm output results
图 1. XGBoost 算法输出结果

综上所述，影响公司债券违约行为的主要因素是 X_{12} 资产负债率、X_{11} 营业收入、X_6 净资产收益率 ROE、X_3 应收账款周转率，其中影响力最大的是资产负债率。资产负债率本质上就是企业运用债权人所提供资金，进行经营活动的能力衡量指标，更是债权人提供贷款的安全保障指标，侧面验证 XGBoost 算法模型能在许多变量中有效找出影响公司债券违约的主要变量。

4.2. Logistic 回归模型构建

本文采用二元逻辑回归，针对债券是否违约进行分析预测，首先对多分类变量进行编码，外资企业为 1，民营企业为 2，公众企业为 3，国有企业为 4，其他企业为 5，然后使用 SAS9.4 中的 logistic 过程进行 logistic 回归模型的建立，由于 X_9 销售成本率存在多重共线性，删除此变量后，再次进行 logistic 回归，输出的结果如表 2。

由表 2 的输出结果我们可以发现，只有 X_{11} 营业收入(同比增长率)、X_{12} 资产负债率两项参数的估计具有意义，由此可以得到违约风险事件发生的概率为 P 时，逻辑回归模型方程为：

$$\text{logit}(P)=\ln\left(\frac{P}{1-P}\right)=-0.0413x_{11}+0.0614x_{12} \tag{16}$$

Table 2. The output result of binary logistic regression
表 2. 二元 logistic 回归的输出结果

最大似然估计			
参数	估计	Wald 卡方	显著性
Intercept	5.3120	1.5598	0.2117
X_1	−2.7024	2.1289	0.1445
X_2	−0.8977	0.3929	0.5308
X_3	−0.2722	2.4241	0.1195
X_4	−0.0725	1.0689	0.3012
X_5	−0.00736	0.1479	0.7005
X_6	−0.00437	0.3371	0.5615
X_7	0.0538	0.4740	0.4912
X_8	−0.0291	1.9211	0.1657
X_{10}	0.000356	1.6671	0.1967
X_{11}	−0.0413	3.9806	0.0460*
X_{12}	0.0614	1.4696	0.0380*
X_{13}	0.0607	1.2045	0.2724
X_{14}	−119E−12	1.3633	0.2430
X_{15}	−1.1685	4.3057	0.2254
X_{16}	8.29E−12	3.6755	0.0552

根据这些结果，为了排除其他变量影响，使模型更加精确，只输入营业收入(同比增长率) X_{11} 和资产负债率 X_{12}，进行二元 logistic 回归，可以得到修正后的二元 logistic 回归结果如表 3。由表 3，逻辑回归方程进一步修正为

$$\text{logit}(P)=\ln\left(\frac{P}{1-P}\right)=-8.233-0.44x_{11}+0.052x_{12} \tag{17}$$

Table 3. The output result of modified binary logistic regression
表 3. 修正后的二元 logistic 回归结果

最大似然估计			
参数	估计	Wald 卡方	显著性
Intercept	−8.233	46.017	0.000**
X_{11}	−0.044	18.744	0.000**
X_{12}	0.052	10.985	0.001**

由逻辑回归模型方程可以得到，一个企业当年度的营业收入(同比增长率)如果为 0 时，资产负债率必须高达 170%，才有很大概率陷入违约风险。

$$\text{logit}(\text{违约})=\ln\left(\frac{\text{违约}}{1-\text{违约}}\right)=-8.233-0.44\times0+0.052\times170=0.607$$

$$发生违约概率=\frac{e^{0.607}}{e^{0.607}+1}=0.647$$

$$未发生违约概率=\frac{1}{e^{0.607}+1}=0.353$$

通过逻辑回归模型的建立可以发现一个公司债券是否违约与营业收入(同比增长率) X_{11} 和资产负债率 X_{12} 有较大的关系，可以使用当年度的营业收入(同比增长率)和资产负债率来预测公司债券是否违约。一个公司其发行债券是否会发生违约事件与营业收入(同比增长率)呈负相关，即一个企业的营业收入(同比增长率)越小，发生债券违约事件的可能性也就越大；债券是否会发生违约与资产负债率呈正相关，即一个企业资产负债率越大，发生债券违约的可能性就越大。一个企业的收入增长越快，并且负债在企业资产的占比越小，企业的财务状况越好，债券发生违约的可能性越小。

5. 结论

本文收集我国债券市场 885 家上市公司的 2020 年度财务数据，分析造成公司债券违约的财务因素。通过 XGBoost 算法找出影响公司债券违约的主要因素，按照影响力大小排列依次为资产负债率、营业收入(同比增长率)、净资产收益率 ROE 与应收账款周转率，其中资产负债率和营业收入(同比增长率)两个变量，可以解释并覆盖所收集数据 77.4%的观测结果。再通过导入所有变量，建立二元 logistics 回归模型，发现资产负债率和营业收入(同比增长率)具有统计意义，因此修正二元 logistics 回归模型，建立以资产负债率和营业收入(同比增长率)为变因的二元 logistics 回归模型，为预测债券是否违约提供可量化的预测依据。根据逻辑回归方程可以得到，一个企业当年度的营业收入为 0，资产负债率高达 170%时，公司债券发生违约的概率高达 64.7%，因此有很大的违约风险。

通过 XGBoost 算法进行数据挖掘后，所计算出各变量的相对贡献，对比所建立的二元 logistics 回归模型，可以发现两个方法找出的公司债券违约的主要影响因素，都是资产负债率和营业收入(同比增长率)，结果相当一致。表示本文研究的结果是具有参考价值的，建立以资产负债率和营业收入(同比增长率)为变因的二元 logistics 回归模型，可以为公司债券违约的预测，提供量化依据。

在数据收集、分析与研究的过程中发现债券市场的一些问题，提出建议如下：

1) 完善债券信息的披露机制

目前我国的信用评级机构对于债券信息的披露是相对缓慢的，很多时候评级的下调公示是在违约曝光后才发布的。可以说并没有起到一个很好的警示作用。有时候债券的发行公司为了筹集资金，会刻意隐瞒相关信息，而债券人获取信息的渠道相对较少，只能依靠一些评级的机构进行了解，但是评级机构与发行人之间可能会有利益往来，这也使得投资者处于一个不利的地位。所以，我国监管部门应该不断完善相关法律信息，要求发行方在一定程度上披露信息。而监管部门对于公司债券评级、处罚等相关信息也应该及时披露在相应的网站上，对投资者起到一定的保护作用。

2) 完善评级制度

目现我国债券违约事件不断升级，可以看出我国债券评级制度不够完善，很多债券的评级与其实际情形并不相匹配。未来希望政府可以出台比较完善的债券评级制度。

基金项目

中国教育技术协会“十四五”规划一般课题项目(项目名称：新商科大数据应用实验实训平台与教学资源建设研究，项目编号：G002)；2021 年美林数据公司教育部产学合作协同育人项目(项目名称：新商科教改情境下经管类专业大数据应用实验实训平台建设，项目编号：202102344024)；厦门大学嘉庚学院

科研启动基金(项目名称：科研项目启动，JG2018SRF10)。

参考文献

[1] 东方财富网. 2020 年中国债券行业市场现状及发展趋势分析银行间拆借市场交易活跃[EB/OL]. https://baijiahao.baidu.com/s?id=1693106280415074653&wfr=spider&for=pc, 2021-03-02.

[2] 安义宽. 公司债券及其相关品种发展[J]. 经济管理, 2003(11): 6-14.

[3] 丁翠娥. 我国企业债券违约的影响因素及规避措施[J]. 财政监督, 2018(12): 98-102.

[4] 谭文杨. 我国民营企业债券违约原因及启示——“14 富贵鸟”案例分析[D]: [硕士学位论文]. 保定: 河北金融学院, 2019.

[5] Ohlson, J.A. (1980) Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, **18**, 109-131. https://doi.org/10.2307/2490395

[6] Foxon, T.J., Köhler, J. and Oughton, C. (2015) Innovation for a Low Carbon Economy. Edward Elgar, Cheltenham.

[7] Collin-Dufresne, P., Goldstein, R.S. and Martin, J.S. (2001) The Determinants of Credit Spread Changes. *The Journal of Finance*, **56**, 2177-2207. https://doi.org/10.1111/0022-1082.00402

[8] Bakshi, G., Madan, D.B. and Zhang, F.X. (2001) Investigating the Sources of Default Risk: Lessons from Empirically Evaluating Credit Risk Models. Social Science Electronic Publishing. https://doi.org/10.2139/ssrn.262673

[9] Azizpour, S., Giesecke, K. and Schwenkler, G. (2008) Exploring the Sources of Default Clustering. *Journal of Financial Economics*, **129**, 154-183. https://doi.org/10.1016/j.jfineco.2018.04.008

[10] 周宏, 徐兆铭, 彭丽华, 杨萌萌. 宏观经济不确定性对中国企业债券信用风险的影响——基于 2007-2009 年月度面板数据[J]. 会计研究, 2011(12): 41-45, 97.

[11] 周宏, 林晚发, 李国平, 王海妹. 信息不对称与企业债券信用风险估价——基于 2008-2011 年中国企业债券数据[J]. 会计研究, 2012(12): 36-42.

[12] 黄石, 黄长宇. 我国企业债券市场信用风险评级研究[J]. 当代经理人, 2006(21): 457-458.

[13] 曹萍. 基于 KMV 模型的地方政府债券违约风险分析[J]. 证券市场导报, 2015(8): 39-44.

[14] 曾江洪, 王庄志, 崔晓云. 基于 SVM 的中小企业集合债券融资个体信用风险度量研究[J]. 中南大学学报: 社会科学版, 2013(2): 5.

[15] 刘慧芳. 基于信用评级的企业债券信用风险预测研究[D]: [硕士学位论文]. 成都: 四川师范大学, 2017.

[16] 沙一诺. 基于数据挖掘的企业债券违约风险预测[D]: [硕士学位论文]. 上海: 上海师范学, 2021. https://doi.org/10.27312/d.cnki.gshsu.2021.002194

[17] 程照星. 数据挖掘在电信企业客户细分中的应用[D]: [硕士学位论文]. 重庆: 重庆大学, 2004.

[18] 王言, 周绍妮, 石凯. 国有企业并购风险预警及其影响因素研究——基于数据挖掘和 XGBoost 算法的分析[J]. 大连理工大学学报(社会科学版), 2021, 42(3): 46-57.

[19] 张孟迪. 基于 Logistic 回归和 XGBoost 的银行信用卡客户流失预测[D]: [硕士学位论文]. 济南: 山东大学, 2021.