# 金融贷款中企业多指标信用评估

**作者**：曾兆辉，冯廷玉，薛尚哲，黄旺辉，王彬，冯永

**单位**：重庆大学，计算机学院

**案例版权**：该案例归重庆大学计算机学院所有

**涉及的知识点**：数据预处理、数据可视化、随机森林分类、MapReduce

**案例来源及案例真实性情况**：该案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。

**摘要** 在金融行业中，对银行来说选择给哪家企业发放贷款是一个重要的问题，这一选择通常由银行负责相关方面的经理根据经验进行主观判断。在本案例中。我们使用大数据与机器学习的相关方法对企业的信用等级进行评估，以帮助银行进行合理的选择企业发放贷款。本案例使用了123家企业的多种数据，包括营运情况、发票数据、违约记录等。首先通过数据预处理，把这些数据处理成8项可以量化的指标，利用MapReduce计算出每家企业每项指标的数值，最后使用随机森林分类器将企业按照从优到差分为A、B、C、D四个等级。与来自银行的人工分级结果相比较，随机森林分类器的准确率达到了82.1%。

**关键词**：信用评估，数据挖掘，随机森林，决策树，MapReduce

## 1 引言

该教学案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。该案例的关键问题为金融贷款中企业多指标信用评估，需引导学生进行的主要内容有：（1）部署Hadoop+Spark大数据处理平台；（2）对数据进行预处理，计算企业多项数值指标；（3）使用随机森林算法，对企业信用进行评级，并使用准确率指标评估算法性能。

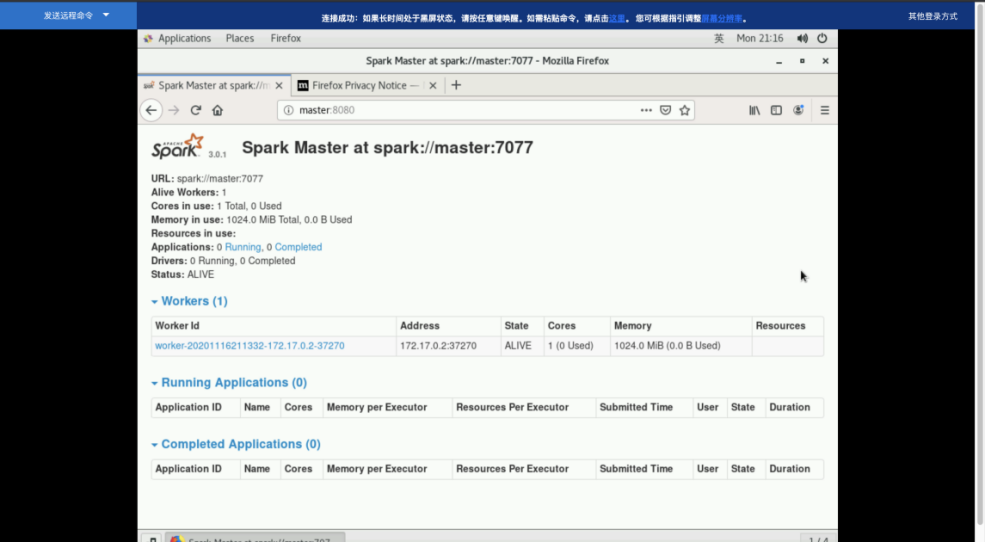
## 2 背景介绍

现如今，生活中大大小小的公司企业经常需要向银行申请贷款服务。银行资金也有限，如何选择优质企业发放贷款就成了需要解决的重要问题。在实际中，由于中小微企业规模相对较小，也缺少抵押资产，因此银行通常是依据信贷政策、企业的交易票据信息和上下游企业的影响力，向实力强、供求关系稳定的企业提供贷款，并可以对信誉高、信贷风险小的企业给予利率优惠。银行需要人工根据中小微企业的实力、信誉对其信贷风险做出评估，然后依据信贷风险等因素来确定是否放贷及贷款额度、利率和期限等信贷策略。但是由于人工评判，就必然会产生主观因素影响最终的评判结果。甚至会出现贿赂等不正当行为，带来坏账风险。因此，使用大数据相关技术与机器学习算法来完成企业信用评估任务，为银行放贷提供可靠的参考信息，具有极大的研究价值。

## 3 内容

**3.1 搭建大数据处理平台**

本案例采用Hadoop+Spark的大数据处理平台，具体的平台搭建过程可参考官方文档及相关教程，在此不再赘述。



**3.2 数据集及其预处理**

本案例的数据集来源于2020年数学建模国赛C题，是由国家银行给出的真实数据集。该数据集提供了123家有信贷记录企业的相关数据、贷款利率、客户流失率关系的2019年统计数据。数据说明如下：

(1) 进项发票：企业进货（购买产品）时销售方为其开具的发票。

(2) 销项发票：企业销售产品时为购货方开具的发票。

(3) 有效发票：为正常的交易活动开具的发票。

(4) 作废发票：在为交易活动开具发票后，因故取消了该项交易，使发票作废。

(5) 负数发票：在为交易活动开具发票后，企业已入账记税，之后购方因故发生退货并退款，此时，需开具的负数发票。

(6) 信誉评级：银行内部根据企业的实际情况人工评定的，银行对信誉评级为D的企业原则上不予放贷。

(7) 客户流失率：因为贷款利率等因素银行失去潜在客户的比率

在数据的预处理中，我们需要计算如下几个指标：

(1) 盈利能力：盈利能力体现的是企业获取利润的能力，使用销售净利润率指标表示利润能力，其比率越高，表明企业获取利润的能力越强，其计算方法如下：

销售净利润率=净利润/营业收入

对应到数据，使用当年发票进项总额与发票销项总额的差值表示净利润，使用当年发票进项总额表示营业收入，此为定量极大型指标。

(2) 营运能力：营运能力代表着企业的经营运行能力，是企业生存发展的核心指标。使用企业的总流水 (C21) 和进出单数 (C22) 表示企业的营运能力，其计算方法如下：

总流水 = 营业支出 + 营业收入

进出单数 = 进项发票数 + 出项发票数

(3) 履约能力：考虑到虽然作废发票不一定意味着被评价公司因故违约，但也能一定程度地体现出供求关系的不稳定性。同时，出现负数发票意味着因故退货即违约，而其责任在于购方，因此只统计进项的负数发票数，综上，违约比率可以定义如下：

违约比率 = （作废发票数 + 进项负数发票数）/ 总发票数

(4) 偿债能力：偿债能力直接关系到中小微企业能否按时偿还银行贷款，是银行最为关心的指标。使用累计资产对其偿债能力进行定量处理，其计算方法如下：

累计资产 = ∑净利润

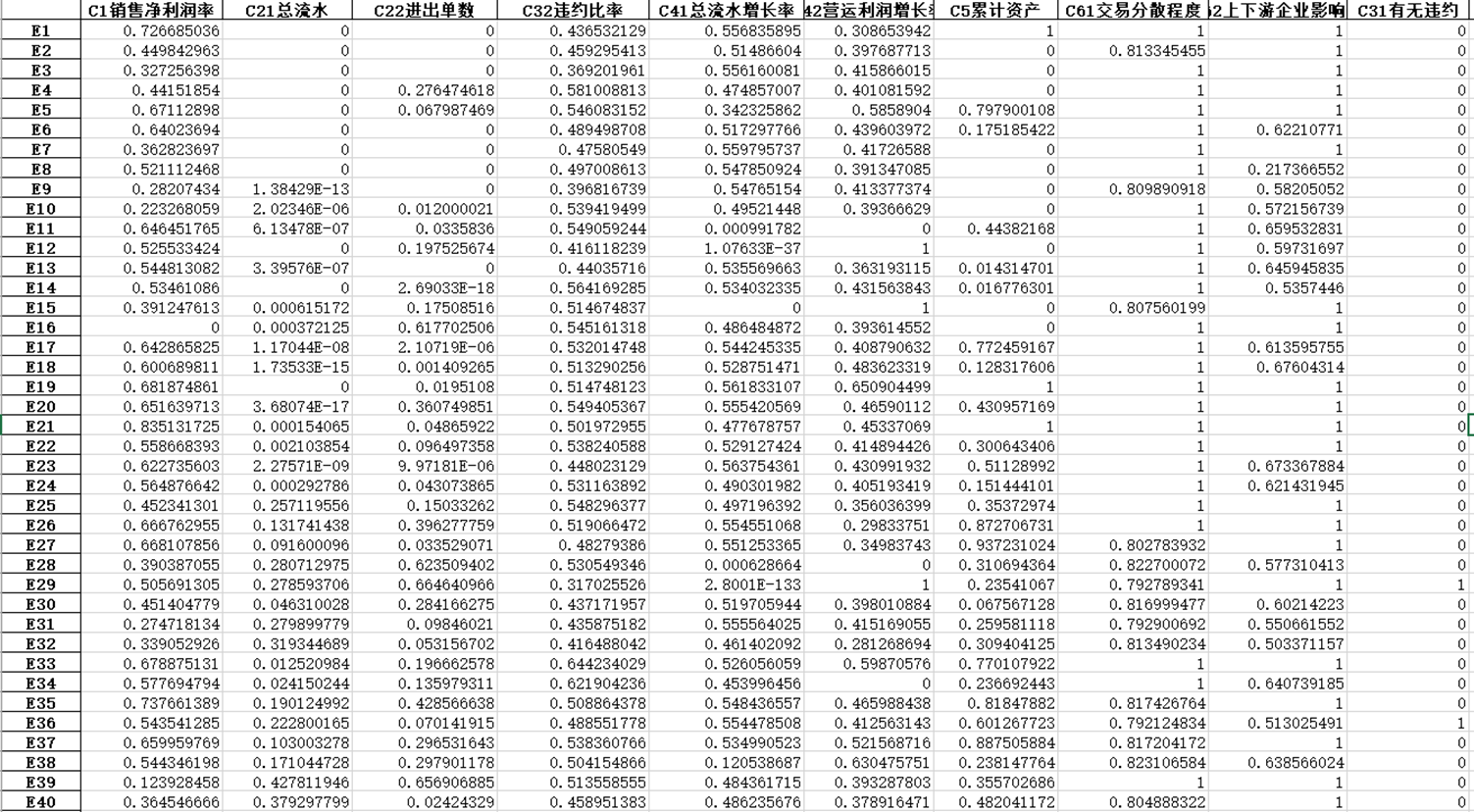
对每个企业，计算其创立至今进项发票总额与销项发票总额之差，即利润总和，可以得到累计资产。

5.上下游供需稳定度：上下游关系的稳定反映了企业供求关系的稳定，使得企业更高概率拥有偿还银行贷款的能力。对于上下游供需稳定度，重要的是企业是否能实现供需关系的多样化，采取交易分散程度表示这个指标，其计算方法如下：

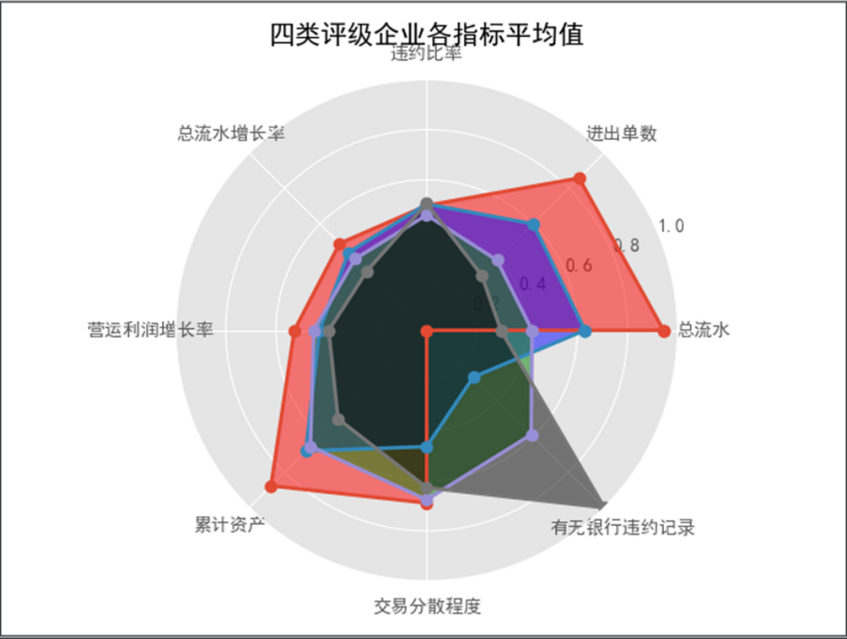
交易分散程度 =∑单笔交易额2 / 总交易额2

同时需要极大型和极小型指标进行标准化，以极大型为例，有：。另外，对于以“是否”判断的指标，进行离散化表示。

下面给出预处理后的部分数据实例，其中E1~E40表示公司代号。



另外，我们将不同类别的企业的各项指标进行可视化展示，如下图所示：



在上图中，不同颜色的色块表示不同等级的企业，红色为评级最好的A级企业，依次是蓝色、绿色、黑色分别代表B、C、D类企业。根据我们得到的可视化图形进行分析可以发现：

(1) 左上角增长率数据并没有显著差异，可见银行对其信誉评级基本不看增长率数据。

(2) 右下角有无银行违约记录是个很强的指标，表明银行A级必定无记录，D级必定有记录，B、C级则不一定。

(3) 累计资产/进出单数等存量指标是银行比较看重的指标。

可以看出，对于银行来说比较在意是否有违约纪律，其次流水量、进出单数、累计资产越大，对于申请贷款越有利。这也是和我们生活中的情景是相吻合的，银行希望企业具有良好的信誉情况，又有比较大的体量可以确保有能力归还贷款。

这也就形成了我们所说的“银行家悖论”，最需要钱的企业，信用风险很大，体量和流水量比较小，很难得到贷款。反而是不怎么需要贷款的企业，信用记录良好，体量大，流水量充足，很容易得到贷款。所以我们的借贷模型还有很大的提升空间，如果可以把评估维度更大一些，可以量化企业的创新指数可以很好地等，相信就可以很好的解决这一问题了。这一模型也就可以从银行贷款领域推广到风险投资领域。

**3.3 随机森林算法**

随机森林是通过集成学习的方式将多棵决策树进行集成的一种算法，其基分类器为决策树。每棵决策树都是一个分类器，对于一个输入样本，N棵树会产生N个分类结果。而随机森林集成了所有的分类投票结果，将投票次数最多的类别作为最终的输出。

本案例中的决策树采用C4.5算法。决策树算法本质上是一种分类算法，目标是将具有p维特征的n个样本分类至c个类别中。

Spark平台上提供了开箱即用的随机森林算法，封装于mllib包的RandomForest和RandomForestModel类中，可方便地进行调用。本案例中相关参数设置如下：

(1) data - 训练数据集：LabeledPoint中的RDD。

(2) numClasses - 用于分类的类数, 可以支持二分类，也可以支持多分类。

(3) categoricalFeaturesInfo - 存储类别特征的Map。

(4) numTrees - 随机森林中的树数目。

(5) featureSubsetStrategy - 分裂每个节点需要考虑的特征数。支持的值：“auto”, “all”, “sqrt”, “log2”, “onethird”。这里使用auto

(6) impurity - 用于信息增益计算的标准。支持的值：“gini”或“ entropy”。 这里使用gini。

(7) maxDepth - 树的最大深度，这里设置为4。

(8) maxBins - 用于分割特征的最大bin数量，这里设置为32。

**简单地采用法评估**

当简单的采用决策树算法进行分类时，结果中有92家企业与银行的人工评判相同，正确率达到了74.8%。

采用集成环境下的随机森林算法进行分类，有101家企业与人工评判相同，正确率提升到了82.1%。下表为部分公司在两种不同算法下的分类结果对比：



在分类结果中，随机森林算法中给到B类的企业，在决策树算法中给到了A级。对比人工评判结果可知，决策树错误而随机森林正确的部分，决策树似乎更加“宽容”，我们分析原因大致如下：

 决策树选用留出法，对选用的数据集求各级别企业指标平均值，发现的确比总的偏高，决策树习得了训练集的特点，出现了过拟合问题。

 其次是随机森林模型的泛化能力比较好，主要依靠了其中三个随机过程，即产生决策树的样本是随机生成，构建决策树的特征值是随机选取，树产生过程中裂变的时候是选择N个最佳方向中的随机一个裂变的。当随机森林产生的树的数目趋近无穷的时候，理论上根据大数定理可以证明训练误差与测试误差是收敛至一致的。

## 4 小结

该案例的选取金融贷款中企业多指标信用评估为主要问题，使用123家企业的多维度数据，结合数据挖掘、机器学习学习的相关方法，进行企业信用等级的评估。该案例主题新颖，结合了金融风控领域的现实需求与大数据分析与挖掘的多种理论与技术，可以充分增强学生的实践能力与理论基础。另外，本案例的内容仅为指导性的过程，在实际教学中，可保持基本研究内容不变，鼓励学生引入其它的数据预处理、数据挖掘、机器学习或深度学习方法完成任务，使用其它公开数据集进行实验研究，并考虑针对实际应用的进一步拓展。

## 附录

1. 本案例提供配套的PPT、视频、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例参考文献如下：

[1] Del Río S, López V, Benítez J M, et al. On the use of mapreduce for imbalanced big data using random forest[J]. Information Sciences, 2014, 285: 112-137.

[2] 王强.决策树过拟合问题研究[D].安徽:合肥工业大学,2008.

[3] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.