# Python 技术在商业银行信贷风险数据方面的运用

文/陈婕

摘要

本文在借鉴以往经验及典型案例的同时，从总体结构、数据重组、风险监控、外部信息、非结构化数据分析等方面，对数据分析技术、外部数据与银行自身业务数据的结合进行了探讨与实现。为各银行在设计 python 信贷风险数据控制时提供一些思路。

【关键词】Python 商业银行 信贷风险 数据

如何完善信贷风险数据作为商业银行的主要利润来源，如何准确有效地防范与管理商业银行的信贷风险数据与信贷系统质量，是未来银行企业稳定的关键因素。商业银行低效率持续上升的原因，究其原因为，个人信贷额度多、信贷周期短、机构凝聚力不强、收益率不高。因此，在银行企业客户信贷风险管理中，基于银行企业信贷风险的控制是商业银行信贷风险的关键。本文在定性与定量分析的基础上，首先对 Python 技术在商业银行信贷风险数据方面，作为数据挖掘工具的应用分析介绍，然后对 Python 在商业银行信贷风险数据管理中的应用进行分析，最后构建 Python 对客户信贷风险评估分析模型。为信贷风险管理与预警方面的科学完成提供借鉴。

## 1 Python在商业银行信贷风险数据管理中的应用分析

Python 在商业银行信贷风险数据中，主要作为数据挖掘工具来使用，因此想要分析 Python 技术在商业银行信贷风险数据方面的应用，需要结合数据挖掘在银行信贷风险控制中的应用，特别是在客户信贷分析与贷款调度中的应用。银行的稳健经营有赖于全面的信贷风险管理与控制。许多因素可以影响客户信贷，包括大量数据、干扰信息与数据噪声。收集有利于银行的客户信贷风险分析信息是非常重要的，然而，数据挖掘技术是显而易见的。

（1）通过相关分析、特征选择等技术手段进一步分析客户信用状况帮助银行有效识别重要客户信用水平，客户信用水平与各因素之间存在一定的关系，如何考察确定申请人信用额度的不同因素之间的关系？利用数据挖掘技术来解决这些问题，可以更好的评价信用因素，在数据开发后，预测未来客户信用质量的变化，在此基础上，实施相应的预警措施。降低风险成本和信用风险管理。

（2）在整个信用风险管理周期中，可以利用数据挖掘技术，从风险预测、识别、预警、控制、传导等方面建立科学有效的信用风险管控机制；建立完善的体系，制定有针对性的管控体系，提高银行核心竞争力。

（3）数据挖掘可以用来分析客户的信用风险。它可以分为五类：正常、关注、次别、可疑和损失，即使用信用和偿还客户信用评级是根据客户的信用状况、基本信息，以及客户当前的信用状况，实现客户信用风险控制。

要对此类贷款进行后期持续监管，全面及时掌握担保人和信贷客户的银行融资情况，如公司生产经营情况、对外担保情况、还本付息情况、融资总额等信息。风险等级为“次级”及以下的问题贷款，需要银行有关部门进

```
In : safe loans raw = loans[loans[target] == +1]
     risky loans - raw = loans[loans[target]一一I]
     ratio = len(risky loans raw)lfloat(len(safeto loans raw))
     risky loans二risky loans raw
     safe loans = safe loans raw.sample(frac = ratio, random state = 1)
     loanswe data = risky loans.append(safe loans)
In [12] : Nl二len(safe_loans)
     N2 = len(risky loans)
      N == N1 + N2
     print(?%o    f safe loans : %.2f%%"%(N I lN * 100.0))
     print(?%o    f risky loans : %.2f%%"%困2!N * 100.0))
     print( " Total number of loans in our new dataset :?,   N)
输出: %of safe loans : 50.00%
     %of risky loans : 50.00%
      Total number of loans in our new dataset : 14368
```

图 1

```
In : decisionto tree一 model = tree.DecisionTreeClassifier(max_depth = 5)
     decisiones tree model = decision一eees model.fit(X, Y)
     print(decision一ee^ model)
输出: DecisionTreeClassifier(class weight = None, criterion ='gini', max - depth = 5,
                二搗,  splitter ='best')
In : small_ tree一 model二tree.DecisionTreeClassifier(max_ depth = 2)
     small一ee一 model = small一ee一 model.fit(X, Y)
     print(smallee tree model)
输出: DecisionTreeClassifier(class weight = None, criterion ='gini', maxes depth = 2,
                ……,  splitter ='best')
```

图 2

```
In : big tree model = tree.DecisionTreeClassifier(max depth = 9)
     big tree - model = big tree model.fit(X, Y)
     print(big tree model)
输出: DecisionTreeClassifier(class_weight = Nonea criterion ='gini', max - depth = 9,
              sputter ='best')
In : print( biges tree model.score(X, Y))
     print( big tree一 ~ model.score(Xv, Yv))
输出: 0.901861840961
      0.894572025052
```

图 3

表 1：关键性特征因素提取后实例表

| 样本序号 | Layered | mode | term | type | safe_loans |
|---|---|---|---|---|---|
| 26 | B | xinyong | 1Y to 3Y | A | 1 |
| 27 | B | xinyong | 6Y to 1Y | B | -1 |

表 2：数据平衡化后的样本表

| 样本序号 | Layered | mode | tern | type | safe_loans |
|---|---|---|---|---|---|
| 27 | B | xinyong | 6M to 1Y | B | -1 |
| 28 | B | xinyong | 6M to 1Y | B | -1 |
| 29 | B | xinyong | 6M to 1Y | B | -1 |
| 196 | A | xinyong | 6M | A | 1 |
| 82877 | A | zhiya | 6M | A | 1 |
| 9496 | A | zhiya | 6M | A | 1 |

行管理和推广。不良贷款专项归集。商业银行需要对担保人和信贷客户财务指标的变化进行监测和分析，掌握变化的根本原因和相关风险因素。根据实时发现的风险预警信息或紧急风险事件，对客户的生产经营或财务状况、偿付能力有重大影响的，要评估其对信用风险的影响，动态调整客户的风险分类，进一步落实有效措施。为降低风险，实时跟踪监控风险控制效果，最终减轻或终止风险事件的不利影响提供保障。综上所述，利用 Python 技术，我们可以对信用客户的信用风险进行建模和分析，预测信用情况是否良好。为科学支持信用认证工作，推动信用风险管理向质量管理量化管理转变，提供借鉴。

## 2 Python构建对客户信贷风险评估分析模型

本节根据层次分析法确定了四个关键特征因素：贷款分类 (type)、担保方式 (mode)、贷款期限 (term)、客户特征分层 (Layered) 构建决策树模型。四个关键特征因素的定义以及决策目标代码如下所示 :

```
In: features=['Layered', 'mode'
'type']
target='safe loans'
loans=loans[features+[target]]
```

如表 1 所示。

### 2.1 数据平衡化

```
In : Xs = sample - validation data[features]
     Ys = sample validation_ data[target]
In [30]: prediction1 = decision tree model.predict(Xs)
     prediction2 = decisiones tree - model.predict_proba(Xs)
     print( prediction1)
     print( Ys.values)
     print( prediction2[:,1])
 输出:[1  1  -1  -1]
      [1  1  -1  -1]
      [0.96913854  0.96734621 0.0051458   0.00136147]
```

图 4

经过数据处理后，首先要平衡严重失衡的信贷数据。数据集意味着一类记录中的样本数远低于其他类，其中少数阶级被称为少数阶级，而多数阶级是多数阶级。企业信贷风险研究

少数未偿还贷款的分类准确性往往更为重要。为了改进少数类别的分类，有必要对样本进行平衡。如上图所示，良性贷款与非良性贷款的百分比分别为 95.57% 与 4.43%o 为了解决这一问题，比例由良好贷款与不良贷款的数量决定，在这种方法中，从好的贷款中随机抽取一定比例与一笔不可偿还的贷款，由函数 append（）形成。新的信贷数据通过删除训练集中大多数类的样本数来平衡数据集。虽然它可以提高分类性能，降低计算复杂度，但缺点是它会影响大多数类的分类。导致样本信息丢失。代码如图 1 所示。

平衡精简后贷款比例为一比一，总样例数从 162289 减少到 14368。表 2 为数据平衡化后的样本表。

### 2.2 决策树模型的建立

处理后的样本集按随机程序分为训练样本和测试样本。利用训练数据集进行数据分析，得到了记录分类的决策模型。验证的目的是获得训练模型在集合上的预测误差，以便选择模型。对测试数据集中的数据进行分类，对分类结果和实际结果进行评价，确定真实的测试误差和评价指标，模型预测的精度为模型评价，三组数据的划分没有严格的标准，基于不同样本量的决策也不尽相同 。然而，一个原则是测试集必须保持未知，并且独立于训练集和验证集。首先将前一节处理的数据按四比一

**<< 下转 141 页**

# 数据挖掘技术在软件工程中的应用

文/唐海燕[1]　兰兵[2]

摘要

本文结合数据额挖掘的内涵，探析软件工程应用数据挖掘的意义，提出相应的应用对策。在现代信息技术快速发展的过程中，数据挖掘技术被逐渐应用在社会各领域，不仅为我国社会市场经济的快速发展提供助力，更推动了我国现代化建设的步伐。

【关键词】数据挖掘 软件工程

在信息化背景下，我国传统的数据信息技术已经难以满足现代企业发展的需求。而数据挖掘的发展与普及，能够有效实现数据信息的即时保存与精准传送，推动我国现代企业信息化建设的步伐。现阶段，作为我国产业发展最快的软件功能，往往与数字信息技术存在紧密的联系，而将数据挖掘应用到软件工程产业体系中，不仅能够激发数据挖掘的全部潜能，更能提高软件工程的发展质量，降低软件工程的成本投放，规避相应的风险出现。因此软件工程企业应用数据挖掘，不仅具有重要的现实意义，更有显著的时代意义。

## 1 数据挖掘的基本内涵及内容

数据挖掘具体指现代信息技术，与传统数据技术相比，具有强大的信息处理、传送、存储等功能。在我国社会各领域中得到广泛的应用。然而现阶段，我国部分企业对该技术的应用价值了解甚少，依旧采用传统的数据信息处理手段，导致信息处理质量与效率相对低下。而数据挖掘具体包括了数据分析、数据转换以及数据处理等功能，不同功能间具有紧密的联系，可以有效实现对数据信息的综合评估。对于软件工程产业来讲，应用数据挖掘技术，不仅能够有效增强数据信息的处理质量，防止失误出现，更对企业的全面发展具有重要的推进作用。在具体的数据挖掘环节层面，软件工程公司首先应“界定”商业问题，进行相应的数据准备，进而以数学建模的方式理解数据，并对比既定的商业问题，对数学模型进行评估。最后，将模型应用与商业问题中，明确数学模型的时效性，以此彻底解决原有的商业问题。其中，在模型评估的过程中，需要围绕数据源对商业问题的具体内容进行二次界定，以此保障模型评估的准确性与科学性。

## 2 软件工程应用数据挖掘的意义

### 2.1 深化对信息的理解

基于数据挖掘含有传统数据信息技术的各项功能，在将大量数据信息进行集中采集后，企业能够根据数据信息的基本类型进行自动化分类管理。而在此种数据采集模式下，数据采集所涉及的范畴广泛，内容丰富，可以根据不同数据信息的基本体系展开针对性分析，建立出体系完善的管理平台，方便企业快速查询及获取所需的信息资源。而在软件研发阶段，数据挖掘可以将大量零散的数据资源进行集约化的整合处理，使企业能够多角度、全方位的了解并掌握不同数据信息的内涵。简而言之，数据挖掘能够以类似大数据技术的手段对大量信

**<< 上接 140 页**

的比例分为训练数据与验证数据，然后确定训练数据的输入值 x 与初始值 y，得到训练数据（11494，20）的验证数据（2874.20），然后利用决策树辅助函数建立两个树模型。函数变量可以给出不同的值来控制树的复杂度。因此，每五层与两层的最大深度使得第一树（最优树）比第二树（简单树）更复杂。从结果的树信息输出可以看出，模型的 gini 索引是用来共享属性的。代码如图 2 所示。

使用 score（）函数评估树的分类精度。然后生成具有 9 层最大树深度的复杂树。代码如图 3 所示。

复杂树的训练数据和验证数据的正确率分别为 90%、19% 和 89.46%，训练数据和验证数据的正确率分别为 89.99%。简单树的训练数据分别为 89.24%。88.90%，如果树由简单到复杂，则训练精度逐渐提高，验证精度先升后降，树深过大导致过度适应。综上所述，选择树深为 5 的中树模型作为企业客户信用风险评估模型。

### 2.3 决策树模型的验证

评价模型的预测由验证集中的两个正、两个负检验样本进行检验。因此我们必须得出以下结论：首先，使用决策树模型来预测贷款是否合理。其次，它提供了良好贷款的可能性，预测结果如图 4 列代码所示。

从结果可以看出，通过关键因素指标识别出的决策树模型可以很好的为贷款提供决策支持。决定并提高企业信贷风险管理的效率。

## 3 结束语

随着全球经济金融一体化的不断发展，信贷风险不可避免，一旦银行进入经营过程，难免会产生不成功的贷款业务，贷款利率上调不成功，不仅会给银行造成巨大损失，还会对整个银行业乃至经济的长期稳定发展产生负面影响。同时，信贷风险数据管理是一项长期而艰巨的系统技术。本文通过对 Python 技术在商业银行信贷风险数据方面的运用，假设构建了基于Python的商业银行信贷风险管理系统。基于一个模型对我国商业银行信贷风险数据管理进行实证研究，旨在提高我国商业银行信贷风险管理水平。

## 参考文献

[1] 高喆，禹朝帅，刘钊宾等．基于 Python-Matlab 的 Abaqus 后处理技术在柴油机有限元分析中的应用 [J]．拖拉机与农用运输车，2017(4):46-49.

[2] 孙哲，韶丹，郭建兴．基于 Python 的地震影响场自动生成与发布技术的研究与实现——以陕西省为例 [J]．华北地震科学，2018,36(3).

## 作者简介

陈婕（1982-），女，江苏省无锡市人。大学本科学历，毕业于上海同济大学，工程师，信贷审批高级专家。主要研究方向为大数据风控研究。

## 作者单位

中国工商银行　上海市　200120