高等学校人工智能理论与应用实践系列规划教材
教育部－百度－产学合作协同育人项目规划教材

百度云智学院人工智能方向校企合作指定教材

Python
机器学习实战案例

赵卫东　董亮 ◎ 著

《电子课件》　《程序源码》

清華大學出版社

Python
机器学习实战案例

赵卫东 董亮 ◎ 著

清華大學出版社
北京

内 容 简 介

机器学习是人工智能的重要技术基础，涉及的内容十分广泛。本书基于 Python 语言，实现了 10 个典型的实战案例，其内容涵盖了机器学习的基础算法，主要包括统计学习基础、分类、贝叶斯网络、文本分析、图像处理等机器学习理论。此外，还介绍了机器学习的推荐技术应用。

本书深入浅出，以实际应用的项目作为案例，实践性强，注重提升读者的动手操作能力，适合作为高等院校本科生、研究生机器学习、数据分析、数据挖掘等课程的实验教材，也可作为对机器学习感兴趣的研究人员和工程技术人员的参考资料。

图书在版编目(CIP)数据

Python 机器学习实战案例/赵卫东，董亮著. —北京：清华大学出版社，2019.12
ISBN 978-7-302-54189-9

Ⅰ. ①P…　Ⅱ. ①赵…　②董…　Ⅲ. ①软件工具－程序设计　②机器学习　Ⅳ. ①TP311.561　②TP181

中国版本图书馆 CIP 数据核字(2019)第 255965 号

责任编辑：闫红梅
封面设计：刘　键
责任校对：梁　毅
责任印制：丛怀宇

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-83470236
印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：12.75　　**字　　数**：312 千字
版　　次：2019 年 12 月第 1 版　　**印　　次**：2019 年 12 月第1次印刷
印　　数：1～2000
定　　价：39.00 元

产品编号：084483-01

FOREWORD 前言

当前，随着信息时代的快速发展，银行、投资、零售、互联网甚至传统的制造业都产生大量数据。各行各业开始逐步应用机器学习算法分析数据，以便在海量数据中总结出规律，辅助决策。这种发展趋势使得就业市场对数据科学、机器学习人才的需求不断增加，同时对人才的多元化、综合实践能力提出了要求。

随着数据分析相关行业的快速发展，数据分析在各个领域都得到了很多成功的应用，企业和政府部门都期望在各个业务方面的工作由数据分析能力强的人承担，更期望员工能够探索有效的数据分析方法，并根据实际数据场景分析结果做出决策，将分析和处理数据作为日常工作流程的一个环节，而不是将数据分析作为一项专业技能。同时，随着数据种类的繁多和数量的爆炸式增长，市场对毕业生的数据分析和处理能力提出了更高的要求，需要有数据分析技能的人才去预测行业前景，及时抓住发展机会，形成独有的竞争优势。高校的基本职能是培养人才，为了使学生更好地适应现代工作场所和终身发展，需要认真思考如何培养应用型人才，以适应当前的就业环境。机器学习相关专业以培养数据分析师、算法工程师、大数据工程师等数据分析、应用型人才为目标，这不仅要求学生理解算法本身，更需要学生具备跨学科的实践能力，将算法逻辑应用到实际生产、生活场景以解决现实问题。

企业对数据分析人才的数量和质量的高要求导致了大数据技术、人工智能人才的大缺口，而目前高校的机器学习教学偏向理论化，更多地注重算法本身，缺乏完善的实践教学体系和教学资源。学生的课堂学习只是面对多种专业理论知识的组合，缺少真实项目的实践过程，学生不能有效地将学习内容应用到实践过程中，这与应用型人才的培养目标存在一定的差距，毕业生不足以适应竞争激烈的就业市场。因此，高校需要更多地考虑就业环境与学生的真实需求，对传统的教学模式进行变革，掌握数据科学时代的新技术和新应用，在遵循教育规律的基础上，将实际项目实践与理论教学融为一体，逐步调整课程内容，培养学生自主思考与解决实际问题的能力，从而提高他们的竞争优势。

如何在教学过程中结合项目实践，已经成为各高校关注的话题。传统的机器学习教学在技能培养、数据与实际案例的选择上仍存在很大的提高空间，这与新时代机器学习人才发展的需求存在一定距离，有必要对人才培

养与项目实践相结合进行探索，尝试新的满足社会发展需要的教学模式，为培养具有专业素质和创新能力的机器学习人才奠定坚实的基础。

在学生理解算法原理的基础上，可采用灵活的模块化教学方法来培养学生对实际应用场景的认知。结合案例程序展示其应用，然后结合教学进度提出一些问题，学生通过模仿实现一个类似的验证型实验项目，该项目作为实验项目的原型，学生可访问、分析其功能、代码并测试其效果。随后，以此为基础做扩展实践，学生可以模仿教师提供的案例，通过自主设计并实现一个相对完整的项目，深化并巩固所学的知识，锻炼整体考虑问题的能力，提高灵活应用知识的能力和创新能力。

由于企业面对的很多问题并不能直接交由机器处理，数据的筛选、特征提取以及算法的整合与取舍是需要技巧的。同时，企业实践项目真实灵活并且与当前研究热点紧密相关，在项目解决方案的探讨中学生会面临很多瓶颈，例如样本的不平衡、算法存在的某些缺陷等，这些瓶颈不能直接地从课堂或其他途径上获取到有效的解决方案，更多地需要学生自身总结经验，在现有的思路上进行调优，从而帮助学生掌握算法缺陷，自主发现一些原有教学中被忽略的难点。

企业实践项目不同于常规教学实验，在大多数传统教学方法中，学生按照已有步骤进行规范化的实验，往往可以获得满意的结果。本书正是基于以上的现实需求，结合作者最近几年与企业合作的实战项目，通过一定的抽象和简化，精选了十个比较实用的实训案例，可以作为高校机器学习课程的实验教材，也可以作为学习 Python 课程的实训教材。

学习本书之前，读者需要掌握基本的机器学习理论，附录有测试题，可以在学习前检验。

在本书的写作过程中，研究生蒲实、于召鑫和本科生高名扬在资料收集方面做了很多工作，特此表示感谢。

赵卫东

2019 年 6 月

CONTENTS 目录

第 1 章　集装箱危险品瞒报预测 …… 1

1.1　业务背景分析 …… 1
1.2　数据提取 …… 2
1.3　数据预处理 …… 3
1.3.1　数据集成 …… 3
1.3.2　数据清洗 …… 3
1.3.3　数据变换 …… 6
1.3.4　数据离散化 …… 6
1.3.5　特征重要性筛选 …… 7
1.3.6　数据平衡 …… 8
1.4　危险品瞒报预测建模 …… 9
1.5　模型评估 …… 15

第 2 章　保险产品推荐 …… 19

2.1　业务背景分析 …… 19
2.2　数据探索 …… 20
2.3　数据预处理 …… 28
2.4　分类模型构建 …… 29
2.5　平衡数据集 …… 30
2.6　算法调参 …… 31
2.7　模型比较 …… 33

第 3 章　图书类目自动标引系统 …… 36

3.1　业务背景分析 …… 36
3.2　数据提取 …… 37
3.3　数据预处理 …… 38
3.4　基于贝叶斯分类的文献标引 …… 41
3.4.1　增量训练 …… 43
3.4.2　特征降维与消歧 …… 44
3.4.3　权重调节 …… 47

3.5 性能评估与结论 …… 49
3.6 基于 BERT 算法的文献标引 …… 49
3.6.1 数据预处理 …… 50
3.6.2 构建训练集 …… 52
3.6.3 模型实现 …… 54

第 4 章 基于分类算法的学习失败预警 …… 58

4.1 业务背景分析 …… 58
4.2 学习失败风险预测流程 …… 58
4.3 数据收集 …… 59
4.4 数据预处理 …… 60
4.4.1 数据探查及特征选择 …… 60
4.4.2 数据集划分及不平衡样本处理 …… 64
4.4.3 样本生成及标准化处理 …… 65
4.5 随机森林算法 …… 66
4.5.1 网格搜索及模型训练 …… 66
4.5.2 结果分析与可视化 …… 68
4.5.3 特征重要性分析 …… 72
4.5.4 与其他算法比较 …… 76

第 5 章 自然语言处理技术实例 …… 79

5.1 业务背景分析 …… 79
5.2 分析框架 …… 80
5.3 数据收集 …… 83
5.4 建立模型 …… 83
5.4.1 文本分词 …… 83
5.4.2 主题词提取 …… 86
5.4.3 情感分析 …… 89
5.4.4 语义角色标记 …… 90
5.4.5 语言模型 …… 91
5.4.6 词向量模型 Word2vec …… 92

第 6 章 基于标签的信息推荐系统 …… 96

6.1 业务背景分析 …… 96
6.2 数据预处理 …… 97
6.2.1 现有系统现状 …… 97
6.2.2 数据预处理 …… 98
6.3 内容分析 …… 99
6.4 基于协同过滤推荐 …… 103

6.4.1 用户偏好矩阵构建 …… 103
6.4.2 用户相似度度量 …… 104
6.5 基于用户兴趣推荐 …… 107
6.6 “冷启动”问题与混合策略 …… 110
6.6.1 冷启动问题分析 …… 110
6.6.2 混合策略 …… 110

第 7 章 快销行业客户行为分析与流失预警 …… 112

7.1 业务背景分析 …… 112
7.2 数据预处理 …… 112
7.2.1 数据整理 …… 113
7.2.2 数据统计与探查 …… 116
7.3 用户行为分析 …… 120
7.3.1 用户流失风险评估 …… 120
7.3.2 流失风险预警模型集成 …… 126

第 8 章 基于深度学习的图片识别系统 …… 129

8.1 业务背景分析 …… 129
8.2 图片识别技术方案 …… 130
8.3 图片预处理——表格旋转 …… 131
8.4 图片预处理——表格提取 …… 135
8.5 基于 PaddlePaddle 框架的文本识别 …… 141
8.5.1 环境安装 …… 141
8.5.2 模型设计 …… 142
8.5.3 模型训练 …… 143
8.5.4 模型使用 …… 146
8.6 基于密集卷积网络的文本识别模型 …… 146
8.6.1 训练数据生成 …… 149
8.6.2 DenseNet 模型训练 …… 152
8.6.3 文本识别模型调用 …… 154

第 9 章 超分辨率图像重建 …… 158

9.1 数据探索 …… 159
9.2 数据预处理 …… 159
9.2.1 图像尺寸调整 …… 160
9.2.2 载入数据 …… 160
9.2.3 图像预处理 …… 161
9.2.4 持久化测试数据 …… 162
9.3 模型设计 …… 162

9.3.1 残差块 …… 162
9.3.2 上采样 PixelShuffler …… 163
9.3.3 生成器 …… 164
9.3.4 判别器 …… 165
9.3.5 损失函数与优化器定义 …… 167
9.3.6 训练过程 …… 168
9.4 实验评估 …… 170

第 10 章 人类活动识别 …… 173

10.1 业务背景分析 …… 173
10.2 数据探索 …… 174
10.3 数据预处理 …… 175
10.4 模型构建 …… 176
10.5 模型评估 …… 180

附录 机器学习复习题 …… 184

参考文献 …… 195

第1章 集装箱危险品瞒报预测

随着经济全球化与国际贸易的快速增长，航运业发展之势非常迅猛。由于运输成本低以及货物运量大的优势，海运成为国际贸易货物运输的重要渠道，其中集装箱运输由于其便捷性、安全性和经济性的特点更是得到青睐。

1.1 业务背景分析

在所有集装箱海运货物之中，危险品的占比最高可达一成。危险货物的运输规范往往比普通货物的运输规范要更为严格。因为危险货物的运输经常会引发一些造成人员伤亡、货物损坏、船舶受损以及环境污染等严重后果的事故。而且经常有一些货主故意隐瞒危险品货物信息或因专业知识不够导致漏报、错报的情况出现，这就导致航运公司在不知情的情况下承载着运输危险货物的任务，这样就会导致航运公司在危险货物运输防范以及应急方面处于非常被动的局面。因此在集装箱危险品运输的问题上，航运公司如何能够化被动为主动是安全生产的一项重要任务。

为了提升工作效率，降低人工成本，提高客户体验，目前航运公司大多采用集装箱订舱系统来帮助航运公司规范订舱流程，然而目前的集装箱订舱系统在危险品风险控制方面暴露了很多不足，其中一些比较显著的问题是：

(1) 现有系统中，对于试图隐瞒危险货物的行为，主要通过销售以及客服人员审核识别，客服人员识别危险品瞒报订舱难度较大；

(2) 现有订舱系统积累了大量的历史数据，但是并没有充分挖掘数据中的价值，对历史数据的利用主要体现在报表统计方面，尚未使用数据挖掘；

(3) 识别危险品瞒报订舱以及纠正客户订舱错误信息的举措都会对订舱确认的响应速度造成较大影响，严重情况下甚至会导致客户流失。

在这样的背景下，如何高效、准确地找到危险品瞒报订舱，并通过进一步查验来缓解危险品瞒报风险就成为了航运公司的一个迫切的需求。

1.2 数据提取

由于航运业务是一个有着多年历史的业务领域，并且航运业务信息化建设也已经进行了很多年，所以航运公司积累了大量的订舱数据，而且订舱数据的维度也非常多。因此如何选择合理的数据采集范围以及关键的业务属性，对于后续数据挖掘工作至关重要。

首先确定采集订舱数据的范围。集装箱运输具有周期性、季节性的特点，以整年的数据为佳，而近年来危险品运输的规范性以及检测力度日益提高，危险品订舱数据也越来越准确，所以选取近一年的订舱数据。考虑到最近的订舱流程可能尚未完成，最终选取的数据范围，是以采集时间前一个月为结束点的一整年订舱数据。

确定好采集时间后，确定预测分析所需要使用的关键属性，经过实际可接触到的数据的筛选以及和业务专家的讨论，最终决定使用的关键属性如表 1.1 所示。表 1.1 包含关键属性名、字段含义以及选择该字段作为关键属性的理由。

表 1.1 关键属性表

关键属性名	字段含义	选择理由
CRE_MONTH	下单月份	集装箱运输具有周期性、季节性的特点，淡季与旺季的区分比较明显，所以将下单时间作为候选特征
AGMT_ID	合同协议号	航运公司与客户签订的协议编号，有协议的客户危险品瞒报、漏报及错报的概率较低
ISEBOOKING	电子订舱	是否使用的是电子订舱，订舱方式和瞒报与否是否有关系需要进一步分析与验证
OB_TRAFFIC_TERM	出口条款	集装箱运输起点类型，有到港、到门、到货运站等模式，危险品瞒报往往不会选取到门的方式
IB_TRAFFIC_TERM	进口条款	集装箱运输终点类型，有到港、到门、到货运站等模式，危险品瞒报往往不会选取到门的方式
TRADE_LANE	贸易区	订舱所归属的区域；不同的区域、港口对于危险品的要求与审查力度不同
OOCL_CMDTY_GRP_CDE	货物品名	与危险品信息紧密相关
BRIEF_DESC	货物简称	与危险品信息紧密相关
FULL_DESC	货物详细描述	与危险品信息紧密相关
SH_COOP_FREQ	发货人合作频率	发货人是货主，是负责危险品申报的主体，一般也是瞒报的责任方
FW_COOP_FREQ	货代公司合作频率	货代连接货主与航运公司，货代有确认货主申报信息准确的义务，专业的货代公司有助于控制危险品运输风险，同时也有部分货代公司存在违规操作，欺上瞒下赚取差价
CN_COOP_FREQ	收货人合作频率	收货人也是分析的重要对象，虽然瞒报责任一般不在收货方，但是有时危险品瞒报需要收货方的配合，所以也将其列为进一步分析的对象
SH_COOP_LVL	发货人合作等级	发货人是货主，是负责危险品申报的主体，一般也是瞒报的责任方

续表

关键属性名	字段含义	选择理由
FW_COOP_LVL	货代公司合作等级	专业、高级的货代公司有助于控制危险品运输风险，货代公司的合作等级越高，瞒报的可能性越低
CN_COOP_LVL	收货人合作等级	收货人的合作等级越高，其存在危险品瞒报的可能性越低

1.3　数据预处理

订舱样本数据采集完毕后不能直接使用，这是因为数据中存在大量的冗余属性、噪声、缺值记录和错误数据需要处理，不适合直接开展数据挖掘工作，因此需要进行数据预处理工作。

1.3.1　数据集成

危险品规则分析设计的数据来源于历史数据，而历史数据是存储在一系列结构复杂的关系表中的，为了进行后续的数据挖掘工作，首先需要根据订舱号这一订舱的唯一标识将这些表里的数据集成，将数据从业务分析友好视角转变成数据分析友好视角，便于后续的分析。

1.3.2　数据清洗

数据清洗主要是处理数据中的噪声、缺失值以及一些异常数据。这些问题数据可能是由程序漏洞、历史原因等导致的，这样的数据难以直接使用，会影响到训练的模型，所以要在数据分析工作开展前进行数据清洗。

1）处理缺失值

使用包含缺失值的数据进行数据挖掘会对训练得到的模型造成很坏的影响。所以要对缺失值做处理。利用 Python 中 Pandas 库的方法即可处理缺失值，同时可结合 MissingNo 库可视化展示数据中缺失值的密度，快速直观地了解数据的完整性，统计缺失值的代码如下：

```
def missingDataStat(data):
    total = data.isnull().sum().sort_values(ascending = False)
    percent = (data.isnull().sum()/data.isnull().count()).sort_values(...)
    missing_data = pd.concat([total,percent],axis = 1,keys = ['Total','Percent'])
    missing_data.head(20)
    #无效矩阵的数据密集显示
    msno.matrix(data)
```

本案例采集的订舱样本数据中存在缺失值的记录与比例以矩阵形式展示，如图 1.1 所示。从图中可以看出，样本数据中存在一定的缺失值记录。直接删除这些样本数据固然是最简单的处理方式，但是可能会丢失其中隐藏的知识点，对于预测结果造成较大偏差，甚至

失去预测的意义，需要对不同的特征分别进行分析。

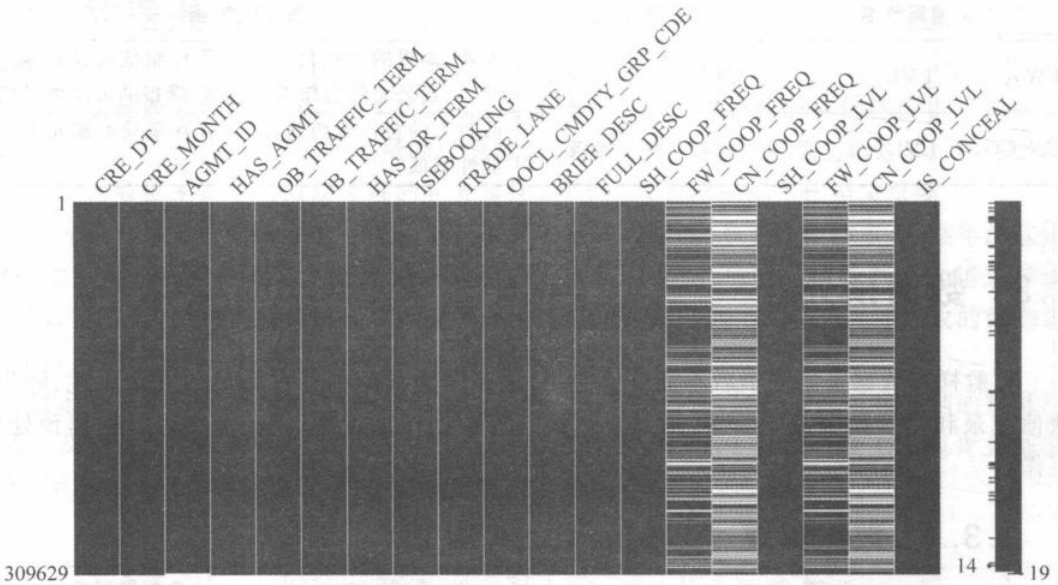

图 1.1 样本数据缺失记录分布

从图 1.1 中可以看出，缺失值主要集中在收货人以及货代公司的有关属性列，即 FW_COOP_FREQ(货代公司合作频率)、CN_COOP_FREQ(收货人合作频率)、FW_COOP_LVL(货代公司合作等级)、CN_COOP_LVL(收货人合作等级)这几列。订舱系统中，联系人信息是非常重要的特征，尤其是业务的主要参与者：发货人、收货人以及货代公司。发货人是必填项，理论上不应该存在缺失值，而在样本数据中存在个别缺失现象，经分析是程序漏洞导致，由于数据量比例极小，可直接删除。收货人与货代公司均存在一定比例的缺值，收货人空值在业务中代表订舱时尚未确定收货人，航运公司按发货人的指示交付货物；货代公司空值代表发货人直接订舱，未委托第三方。因此这里货代公司空值表示发货人自身充当货代公司角色，而收货人空值则代表暂时托运给发货人自己。

货物描述主要运用于分析货物智能分类的规则，存在少量的缺失值，在货物明细表中还有一项货物简称，可直接利用其填补缺失的货物描述。货物属性的缺失，可利用货物信息表中的冗余字段是否危险品、是否冷藏品以及是否大件货重新组合填充。另外，样本数据中还存在少量运输条款、下单方式、付费条款等特征缺失的记录，这些存在缺失的特征均为分类特征，一般可采用“默认值”“众值”等方式填补。经对比，本文采用 KNN 填充法，计算找到临近的 3 条记录，通过投票选出其中最多的类别用以填充缺失的值，此时填充效果较好，以 SH_COOP_FREQ、FW_COOP_FREQ、CN_COOP_FREQ 三列为例使用 KNN 填充的代码如下：

```
def missingData(data):
    #基于业务填补
    sub_data = data.loc[:,['SH_COOP_FREQ','FW_COOP_FREQ","CN_COOP_FREQ"]]
    #KNN 填补
    return KNN(k = 3).fit_transform(data)
```

2）处理错误数据

错误数据主要指数据集成前后有矛盾的数据，对于这些错误数据，第一选择是通过分析得出正确的记录信息并自动修复，若无法修复则删除该样本，避免影响模型的输出。

错误数据处理还包括对异常值的处理。样本数据中存在一些不合常规的数据，如货物的重量和体积等，业务中每个集装箱都有限重，货物体积也受限于集装箱的尺寸，在订舱过程中由于单位以及录入错误等问题会导致少量数据不合常理，对于这类数据要做检测和处理。本案例使用箱图检测异常值，并通过可视化的方式展示数据的离散状态。在箱图中异常值是超出上下限的数据，通过箱型图可以直观地辨别数据中包含的异常值。所以，在对订舱货物按品名分类后，分别计算每一种货物对应属性的上下限，通过箱型图的定义划分异常值后，添加是否存在异常作为新特征用于后续分析。以箱图检测异常值代码如下：

```
data.head()
data.plot.box()
```

通过箱型图检测样本数据中的联系人信息以及下单月份后，发现联系人信息存在异常数据，对于危险品瞒报订舱，联系人合作频率高的往往是一些值得信任的客户，合作频率比较低的客户存在更高的风险会有危险品瞒报订舱情况，因此异常数据不能直接删除。而且这些异常值还是接下来的重点分析对象。数据中下单月份以及联系人信息的箱型图如图 1.2 所示。其中，横坐标分别对应下单月份（CRE_MONTH）、发货人合作频率（SH_COOP_FREQ）、货代公司合作频率（FW_COOP_FREQ）和收货人合作频率（CN_COOP_FREQ）等联系人。纵坐标表示下单月份和联系人的合作频率值。

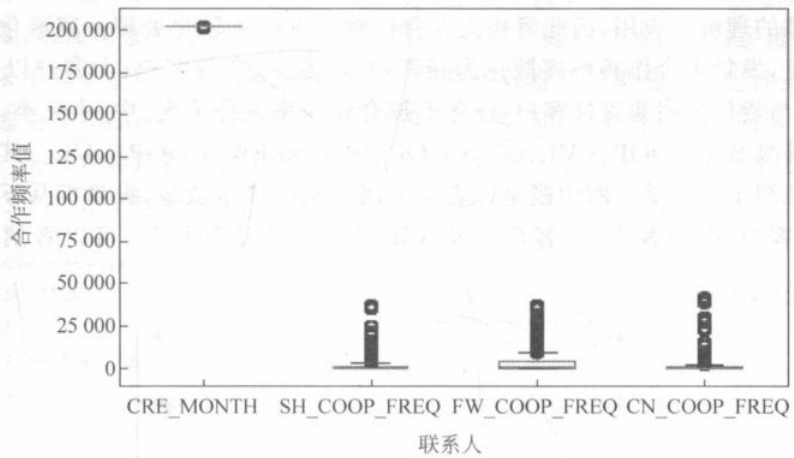

图 1.2　下单月份以及联系人信息的箱图

3）处理重复数据

重复数据主要是指同样的数据在数据库中被存储了多次，如货物信息在装箱后，若订舱含有多个箱子，则货物信息会被复制多份，装在对应的箱子中。此时直接使用会增加对应类别的样本数，使数据挖掘结果产生倾斜，所以预处理时删除多余记录，仅保留一条即可。

1.3.3 数据变换

对于危险品瞒报数据进行转换，得出新的瞒报标志 IS_CONCEAL。业务上认定瞒报的数据源主要有两部分：一部分为运输过程中被航运公司或海关检测出的订舱，此类订舱已被明确打上瞒报标志，另一部分为在审核订舱时发现是疑似危险品，客户又无法提供明确的非危证明而被拒绝的订舱，拒绝原因为疑似危险品，此类情况也认为是瞒报。

引入相关方合作频率，发货人、收货人以及货代公司的历史订舱量，代表着与航运公司的合作程度，一般来说，合作程度高的客户瞒报危险品的概率较低。为了避免合作时长对于历史订舱量的影响，改为使用客户合作频率作为衡量的标准，以(最近一年订舱总量/12)计算每月平均订舱量，分别得出 SH_COOP_FREQ、CN_COOP_FREQ 和 FW_COOP_FREQ。

进出口条款变换为是否含有到门条款 HAS_DR_TERM，条款分为到港、到门、到货运站等多种模式，但是对于瞒报预测，经业务专家分析，是否到门才有本质区别，所以将到港、到货运站等模式重新划分为非门条款。另外，下单方式变换为是否网上电子订舱 ISEBOOKING，并且根据是否存在协议号，引入新变量是否协议订舱 HAS_AGMT。

1.3.4 数据离散化

数据中部分属性为连续值，数据分布较为分散，对其进行离散化后，便于分析，并有助于模型的稳定，降低过拟合的风险。

发货人、收货人、货代公司合作频率，这些数据具有分布非常分散的特点，直接使用不利于后续对模型的理解和应用，因此将相关方合作频率进行离散化处理。离散化的方式是对数据进行分组，发货人合作频率离散化为新客户、小客户、中等客户、大客户以及 VIP 客户五类，收货人与货代公司非直接客户，业务上按合作频率细分为大、中、小三类，离散化后引入变量合作等级 SH_COOP_LVL、CN_COOP_LVL 和 FW_COOP_LVL。其中发货人的离散化结果如图 1.3 所示。图中横坐标表示不同类别的样本数量，纵坐标从下向上依次表示新客户、小客户、中等客户、大客户以及 VIP 客户五类不同标签，可以看到新客户数量最多。

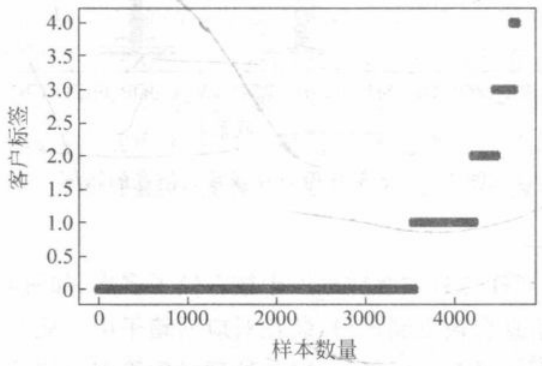

图 1.3 发货人合作程度聚类离散结果

联系人合作程度的离散化采用基于 K-means 算法的聚类离散法，获得聚类中心点的值后，通过 rolling_mean 函数平均移动，计算前后两个聚类中心的均值确定分类边界并切分数据，得到离散化后的新特征即联系人合作级别，代码如下：

```
def coopLvlDiscretization(data,fieldNme,newFieldNme,k):
  fieldData = data[fieldNme].copy()
  #K-Means 算法(k 为离散化后簇的数量)
  kmodel = KMeans(n_clusters = k)
  kmodel.fit(fieldData.values.reshape((len(fieldData),1)))
  kCenter = pd.DataFrame(kmodel.cluster_centers_,columns = list('a'))
  kCenter = kCenter.sort_values(by = 'a')
  #确定分类边界
  kBorder = kCenter.rolling(2).mean().iloc[1:]
  kBorder = [0] + list(kBorder.values[:,0]) + [fieldData.max()]
  #切分数据,实现离散化
  newFieldData = pd.cut(fieldData,kBorder,labels = range(k))
  #合并添加新列
  data = pd.concat([data,newFieldData.rename(newFieldNme)],axis = 1)
  return data
```

细粒度的订舱时间本身意义不大，由于航运业的淡旺季与月份有着紧密关系，所以将时间离散化为月份 CRE_MONTH 作为候选的特征之一。

1.3.5 特征重要性筛选

特征重要性筛选就是剔除与结果没有影响或影响不大的特征，有助于提高瞒报预测模型的构建速度，增强模型的泛化能力，减少过拟合问题，并提升对特征与特征值之间的理解。

本案例采用基于逻辑回归的稳定性选择方法实现对特征的筛选。稳定性选择方法能够有效帮助筛选重要特征，同时有助于增强对数据的理解。以六个特征：合同协议号、电子订舱、货物品名、发货人合作频率、货代公司合作频率、收货人合作频率为例(在图 1.4 中编号依次为 0,1,2,3,4,5)，代码如下：

```
def featureSelection(data):
A1 = data[['AGMT_ID', 'ISEBOOKING', 'OOCL_CMDTY_GRP_CDE', 'SH_COOP_FREQ', 'FW_COOP_FREQ', 'CN_
COOP_FREQ']]
B1 = data[['IS_CONCEAL']]
X1 = A1.values
y1 = B1.values
X1[:, 0] = leAgmt.transform(X1[:, 0])
X1[:, 1] = leEB.transform(X1[:, 1])
X1[:, 2] = leGrp.transform(X1[:, 2])
X1[:, 3] = leSH.transform(X1[:, 3])
X1[:, 4] = leFW.transform(X1[:, 4])
X1[:, 5] = leCN.transform(X1[:, 5])
y1 = LabelEncoder().fit_transform(y1.ravel())
x_train, x_test, y_train, y_test = train_test_split(X1, y1, test_size = 0.3, random_state = 0)
xgboost = XGBClassifier()
xgboost.fit(x_train, y_train)
```

```
print(xgboost.feature_importances_)
plt.bar(range(len(xgboost.feature_importances_)), xgboost.feature_importances_)
plt.show()
```

通过基于 XGBoost 模型的稳定性选择方法分析后，分析各个特征的重要性，得到特征选择中各特征评分如图 1.4 所示，可以看出发货人合作频率的特征重要性评分最高，其重要度最高。

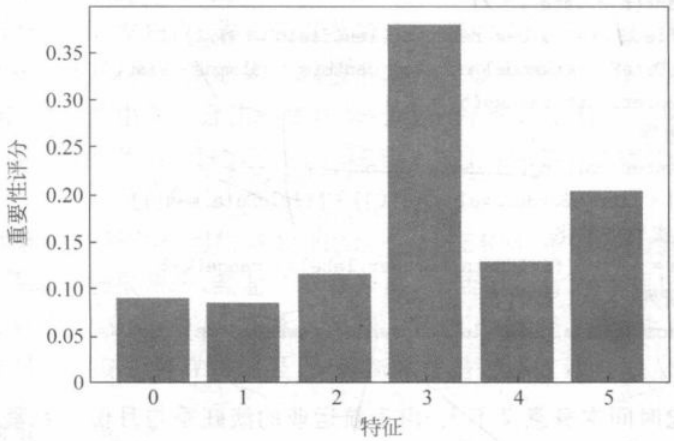

图 1.4 危险品瞒报特征重要性分析结果

1.3.6 数据平衡

危险品瞒报是小概率事件，本案例使用的原始数据中，瞒报记录有 2831 条，非瞒报记录有 341 982 条，占比约为 1∶121，属于严重的数据不平衡问题，这类数据会训练出准确率高但无实际意义的模型，因此需要处理数据不平衡问题。对于数据不平衡的问题，主要分为基于采样的方法和基于算法的方法。基于采样的方法，进一步可分为对于小类样本过采样以及对于大类样本欠采样。基于调整算法的方法有代价敏感方法和 SMOTE 算法等。本案例采用 SMOTE 算法，根据相邻样本数据合成新的样本，以补充小类数据的不足。

调用 Python 的 Imbalanced-learn 库来增加瞒报订舱量，设置瞒报订舱对比未瞒报订舱数据量为 1∶2，适当设置比例可以满足分类算法要求，避免过多地合成新样本数据，并设置随机种子使合成的新瞒报数据固定下来，避免变化的新样本对于模型结果产生干扰。为了避免合成的新样本影响瞒报订舱的预测效果，测试数据保留平衡前的分布，所以 SMOTE 算法仅针对训练集做平衡处理，实现代码如下：

```
def smoteData(data):
    #分为训练集和测试集 7∶3
A,A2,B,B2 = train_test_split(data[[{feature_columns}]],data[[{label_columns}]],test_size =
0.3)
    X = A.values,y = B.values,X2 = A2.values,y2 = B2.values
    #数据平衡(训练集)
    over_samples = SMOTEENN()
    X,y = over_samples.fit_sample(X,y)
    train = pd.DataFrame(np.hstack((X,y.reshape(-1,1))),columns = data.columns)
```